Designing Games to Collect Human-Subject Data

David Gundry

Doctor of Philosophy

University of York

Computer Science

March 2022

Abstract

Applied games align the 'fun' of gameplay with real-world outcomes to achieve social good. For data collection outcomes (e.g. where games are used for experiments or citizen science) various templates and taxonomies for design have been proposed to achieve this alignment. However, existing approaches assume it is always possible to evaluate (and validate) collected data against either 'ground truth' or intersubjective consensus. On the contrary, a significant proportion of human-subjects research is concerned with datums that cannot be validated in this way, such as latent traits and beliefs (e.g. ice cream preference, which cannot be 'validated' against a correct value). Despite extensive knowledge from the social science methodological literature, we do not have comparable templates or taxonomies that can help to design and analyse data collection games for these kinds of data: we cannot yet turn experiments into 'elicitation games'.

This thesis develops a theoretical model for such 'elicitation games', using language elicitation as a case study. Elicitation games must satisfy requirements of validity and motivation. First, I survey validity threats characteristic of the use of games in experiments. Second, I construct a grounded theory of speech motivation to understand what motivates data-providing behaviours in applied games. Integrating these, I theoretically justify a generalised model of data elicitation in games: *Intrinsic Elicitation*. Finally, to identify which validity threats are of primary importance within this model, I run a series of controlled experiments comparing accuracy rates using a novel elicitation game for eliciting adjective order.

This thesis contributes a framework for integrating game design and social science experimental concerns and how they may influence each other for the design and analysis of elicitation games. I find that games incentive rational players to misalign data to experimental outcomes. This can be solved by novel game designs that follow Intrinsic Elicitation.

Contents

List of Tables

List of Figures

Acknowledgements

My great thanks to Sebastian Deterding, whose patient guidance, sharpness of insight, and breadth of knowledge have supported me from wild idea to finished thesis. His support kept me going through difficult times and helped me to emerge the other side a happier and, I hope, wiser person.

My thanks to Seth Cooper and Paul Cairns for examining this thesis and providing helpful feedback. All errors and omissions are my own.

Thank you to Sonia Eisenbeiss for your encouragement and for your valuable perspective from linguistics as my secondary supervisor. My thanks also to Dimitar Kazakov, my first supervisor; though our interests diverged he set me upon upon a path that would become this thesis. My thanks to Jo Iacovides for stepping in as my third and final primary supervisor when needed after Sebastian's move to Imperial. Finally, thanks to Udo Kruschwitz who took over Sonia's role of secondary supervisor.

Thank you to my playtesters and co-designers for the many enjoyable hours playing my games (enjoyable for me at least). And though *Pastry Chefs* will not be in this thesis (much), it remains one of the games I am most proud of. Thank you to the support and friendship of my fellow IGGI Ph.D. students and members of YCCSA. Thanks to Jo Maltby for helping to arrange, among many other things, my parental leave and change to part time. And thanks again to Paul Cairns who helped me resolve a number of administrative (and a few statistical) issues.

Thank you to the reviewers who provided feedback on the papers that were adapted into the chapters of this thesis and to everyone else who shared their knowledge and time in contributing to my research. Finally, thanks to my participants, whoever you are. I hope you enjoyed the game.

To Stacey Gundry, for everything.

And to Jacob Gundry, for all your help.

Author's Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as references.

Chapter 2 has been adapted from the following journal paper:

> David Gundry and Sebastian Deterding. 2018. Validity Threats in Quantitative Data Collection with Games: A Narrative Survey. Simulation & Gaming. SAGE Publications, 50, 3 302–328. https://doi.org/10.1177/1046878118805515

Chapter 4 has been adapted from the following conference paper:

> David Gundry and Sebastian Deterding. 2018. Intrinsic Elicitation : A Model and Design Approach for Games Collecting Human Subject Data. In Foundations of Digital Games 2018 (FDG '18). ACM, New York. 10 pages. https://doi.org/10.1145/3235765.3235803

Chapter 5 has been adapted from the following conference paper:

> David Gundry and Sebastian Deterding. 2022. Trading Accuracy for Enjoyment? Data Quality and Player Experience in Data Collection Games. In CHI Conference on Human Factors in Computing Systems (CHI '22). ACM, New York. 14 pages. https://doi.org/10.1145/3491102.3502025

Sections of these papers have also been included elsewhere within the thesis.

Chapter 1

Introduction

The motivations are familiar. Applied games motivate participants to engage in an otherwise boring task, generating hours of productive play and large amounts of data. This data can then be used for the good of humanity, for instance to solve scientific problems, thereby unlocking new ways of doing science, harnessing human leisure time for social good, and engaging ordinary people in scientific research. Such ambitions first appeared with Luis von Ahn's human computation games (von Ahn, 2005), also known as games with a purpose (von Ahn & Dabbish, 2008), which introduced novel 'agreement' game mechanics for gamifying and validating the collection of intersubjective data, such as image labels. A flurry of research since has extended these principles to the collection of more kinds of data (Quinn & Bederson, 2011), but data always of a recognisable *type*: data about our shared experience, about the intersubjective consensus of individuals. They appeared again with Seth Cooper's scientific discovery games (Cooper, 2014), which presented a paradigm for gamifying scientific tasks by comparing data against a 'ground truth' scientific model. Modellable problems such as protein folding (Cooper, Khatib, et al., 2010) have been targeted: data of another distinct *type*. Driven by the same motivations, each of these paradigms extended the *types* of data that we knew how to collect with a game, each focusing on a different desiderata.

Now in this thesis, I intend to once again extend the franchise to a third, and neglected, *type* of data. The story is familiar: I will identify a new type of data to collect: human-subject data. I will present a model for how to collect it: Intrinsic Elicitation. I will then evaluate this model with the aid of a novel game developed for a specific case study. Just like my academic forbears, I will propose that this example can be generalised for the development of future games and the collection of more kinds of data.

1.1 Applied Games for Human-Subject Data

Applied games are whole games that are designed to achieve some purpose in addition to entertainment (Schmidt et al., 2015). They are a increasingly popular way of eliciting data from people, particularly large online populations (Cooper, 2015; Deterding et al., 2015; Ross & Tomlinson, 2010; Slegers et al., 2016). For this thesis, I will be considering applied games for collecting data about *individual human participants* that is useful for scientific research. The reasons for using applied games over traditional data collection methodologies are as follows:

Participant Motivation Applied games are primarily used to achieve the motivational benefits that game playing supposedly provides. This is common with scientific discovery games (Cooper, 2014) and human computation games (von Ahn, 2005). Standard methods like surveys or experiments are often boring, which can lead to low retention and study completion (Bell et al., 2013; Kelders et al., 2012), poor data quality (DeRight & Jorgensen, 2015; Kirkwood et al., 2010; Ratcliff, 1993; Sawin & Scerbo, 1995), and in general raises ethical concerns (D'Angiulli & LeBeau, 2002; Thackray, 1981). Motivation may correspond to better data (Molina et al., 2013), and fewer missed attention checks (Oppenheimer et al., 2009). Furthermore, presenting or framing a survey or experiment as a game can potentially improve motivation (Friehs et al., 2020; Hawkins et al., 2013; Levy et al., 2016) and data quality (Van Berkel et al., 2017).

Scalability In quantitative experimental research, detecting statistically significant effects with small effect sizes or high variance requires a lot of participants (Lenth, 2001), as does sampling widely from different population groups (Anand et al., 2011). The cost of running large in-person studies can be prohibitive. While standard online surveys and experiments can allow access to more participants, more quickly, recruitment is a challenge. While commercial platforms such as MTurk (Aguinis et al., 2021) and Prolific (Palan & Schitter, 2018) provide easy access to participants, the cost of large samples can be prohibitive. Furthermore, the work-for-pay model introduces its own threats to validity such as bots (Kennedy et al., 2020), satisficing (Litman et al., 2015) and participant non-naivety (Chandler et al., 2012). Applied games, on the other hand, may be able to collect data from a large number of voluntary players at low cost. Successful examples include *The ESP Game*, with 13,000+ players voluntarily producing 1.2 million labels in the course

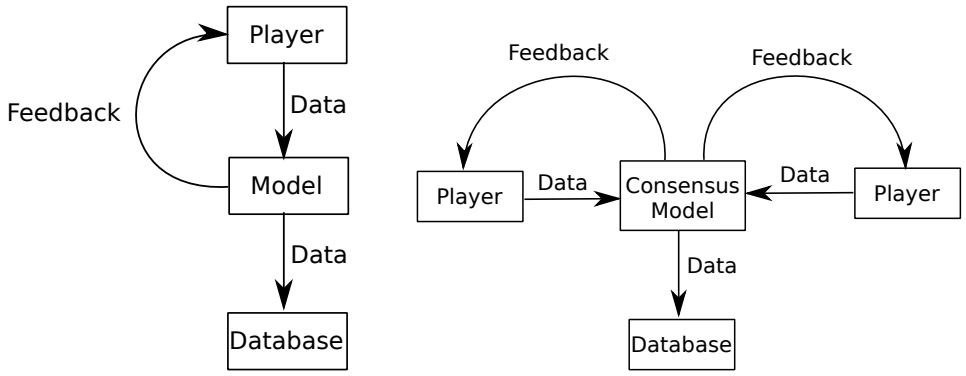
of four months (von Ahn & Dabbish, 2004). And *Phylo* collecting 350,000 solutions from over 12,000 registered users in 16 months (Kawrykow et al., 2012). Games have also been used as a participant recruitment strategy for collecting large experimental samples (e.g. Zendle, Kudenko, et al., 2018).

Methodological Opportunities Experimental tasks, with their apparently arbitrary demands, are often made intelligible by being presented as a 'game'. This is particularly common when working with children. Reaction time tasks (Berger et al., 2000; Dunbar et al., 2001) and reinforcement schedule experiments (Case et al., 1990; Ploog et al., 2009) might appear arbitrary and unnatural: what easier way to introduce such experiments than to say "it's a game"? The use of games provides methodological opportunities. Games in economics allow the study of strategic action in contrived situations (Eckel, 2014). Word games in linguistics use made up words and rules (S. R. Baum, 2002). Situational norms of experiments which can give rise to demand effects (Orne & Whitehouse, 2000) might be avoided, either by obscuring the true purpose of the study within a game, or by replacing the norm of being a 'good participant' with norms of gameplay. Finally, (game)play may serve as a more ecologically valid interaction than task unfamiliar to a participant's everyday life (Eisenbeiss, 2009).

1.1.1 The Challenge of Validity

The challenge that any applied game for data collection must overcome is that of validity. The data that an applied game collects will be used for a purpose and will affect real-world decisions and outcomes. These might be, for example, drawing justifiable inferences (standard scientific validity (Messick, 1995)), making good hiring decisions based on in-game performance, or directing crisis responders to likely survivors spotted on satellite imagery of a flooded region. Therefore it matters that we can both trust and understand what inferences and what outcomes game-collected data support. Moreover, we must be able to reliably design such games that will be able to collect data of a kind and quality that is fit for its intended purpose.

Threats to validity, on the other hand, are properties of data or in its theoretical justification that render it less adequate or appropriate for its intended use. Creating games to collect data is fraught with threats to validity, real or assumed. Various authors in experimental research have suggested potential methodological problems characteristic



(a) Solution-based games (b) Intersubjective consensus games

Figure 1.1: The core components of the Solution-based and Intersubjective Consensus templates

to using games which threaten validity (e.g. Washburn, 2003). One fundamental issue is that motivations within a game will likely be different to those within an experiment. Furthermore, when game events are used to determine real-world outcomes, the issue of gaming the system arises (Baker, 2005). Unhelpfully, the threats to validity that arise from the use of games has not previously been approached systematically.

1.1.2 Current Solutions

The response to the issue of validity within the applied games for data collection literature has been *validation*. The idea is that, if there is a method to validate the collected data post-hoc, the issue of validity is vastly simplified: validated data can be considered valid however it was collected (to the extent the validation mechanism itself is convincing). There are two ways this validation has been achieved in games. These are illustrated in Figure 1.1.

Solution-based games such as *FoldIt* (Cooper, Khatib, et al., 2010) and *EteRNA* (Lee et al., 2014) validate user inputs against a computational model of a target domain. Correctness is therefore defined by the model and the question of validity is reduced to the accuracy of the model. A solution-based applied game for data collection is appropriate for the type of data where it is not practical to computationally *generate* good solutions, but is able to efficiently *evaluate* solutions. The problem might be defined within the game rules in order to constrain solutions to valid ones (Das et al., 2019). Player performance scores can be calculated from the model. These serve to motivate players and direct their efforts towards high-quality solutions (Cooper, Treuille, et al., 2010). However, this template is

only applicable when such a computational model can be defined.

Efficiency is the key concern in such games: designers want the player base of such games to produce high-quality solutions as quickly as possible. As solutions are not known in advance, a key challenge is ensuring the game supports their discovery, for example by iteratively developing the game based on analysis of player solutions (Cooper, 2014). This involves the refinement of in-game tools, such as visualisations and interactions that players may use to locate good solutions, as well as effectively teaching players how to use the system (Cooper, Treuille, et al., 2010). Importantly, none of these changes risk making individual validated solutions that are collected any less valid.

Intersubjective consensus games such as *The ESP game* (von Ahn & Dabbish, 2004) and *Eyewire* (Tinati et al., 2017) typically use agreement mechanics whereby multiple users collectively converge on a consensus solution. They are frequently used to classify and describe datasets (Das et al., 2019). For instance, in *The ESP game*, two geographically separated players agree (without any communication) on the same label for the same image. A statistical argument can then be made that the collected label data is likely to be correct. The more converging data that is collected, the more certain we can be that the data we have collected reflects the intersubjective consensus of our players. Correspondence with such a consensus (whether with a single other player or the whole population of players) can be used to derive a score to direct player effort toward facilitating consensus. So long as there is no obvious way to 'cheat' the consensus (such as by conspiring to agree on a sub-optimal label), the validity argument of such data – as a measure of consensus opinion – is compelling. However, this is only appropriate where the type of data to collect is in fact the intersubjective consensus of a population, not, for instance, the unbiased opinion of an individual.

The essential component of either of the above templates is a mechanism for validation – whether to a model or a consensus. To this mechanism can be added any amount of gamification (game elements), so long as this does not disrupt that validation. This I will summarise as the *gamification+validation* approach. Significantly, with gamification+validation, we do not have or need an integrated understanding of *how* this gamification affects the accuracy or validity of the data. Game elements can be added relatively freely with little risk that the data will become unusable. However, each template applies to only a certain type of data and neither encompasses the kind of data I will be concerned

with here: human-subject data.

1.1.3 Human-Subject Data

Humans are one of the most interesting problems in science. To understand them we need data about them, individually. In particular, much of the data we are interested in is not directly observable (e.g. heart rate) or verifiable (e.g. age), such as preferences, beliefs, latent traits, abilities, judgements, and competences. This is what I will call human-subject data.

Typically, to get such data we run experiments: contriving a particular set of circumstances where participants, motivated by the social norms of the situation provide data that is then recorded, analysed, and distributed as the product of 'science'. Payment essentially does not change this picture (unless payments are performance-contingent). Even where participants are paid to provide *some* data, the moment-by-moment motivation to provide *'good'* data emerges from a willingness to obey with the social norms of the experimental situation (honest, diligent answering, focused attention, etc.) (Orne & Whitehouse, 2000), interacting with the structure that the experimental task provides. This is the familiar way of doing science with human subjects. Yet, as von Ahn (2005) and Cooper (2014) have shown, this is not the only way of doing science. Replace the structure of the experimental situation with the structure of a game. Replace the motivation to behave appropriately in an experiment with the desire for a high score. What would you get? A whole lot of threats to validity.

Unfortunately, the existing templates for *validated* data collection do not apply. Human-subject data can neither be validated against a computational model, as with solution-based data collection games, nor against an intersubjective consensus. Attempting either would undermine the validity argument that the use of human-subject data relies on. In the former case, it would reduce our 'empirical' data to the a priori expectations encoded in our model. In the latter case, it would change the nature of our data through sampling groups instead of individuals; these are different: the preference of an individual need not relate to the preference of a group that individual is part of. Thus validation does not give us an easy way out of dealing with the potentially various but as yet unsystematised threats to validity characteristic of the use of games.

While previous research from applied games does not give us models and templates for collecting this kind of data, this hasn't stopped games(-like) approaches being used in

disciplines such as economics (Eckel, 2014), game theory (Crawford, 2002) and psychology (Washburn, 2003). These disciplines have tested the waters of creating games to replace experiments. However, they have not done so in a way to incorporate the applied games and game design literatures to make enjoyable 'whole' games designed to stand on their own. Yet as I suggested above, 'whole' applied games hold potentially significant benefits for participant motivation, scalability, and experiment methodology.

An alternative to *creating* a whole game for data collection is to use an existing game to collect human-subject data: so-called game intelligence (Devlin et al., 2014). Player actions in a game can express skill or knowledge, expose preferences, or reflect beliefs and unconscious biases. Game analytics can identify these through statistical correlations with real-world measures (e.g. Kokkinakis et al., 2017). Even better, real-world gameplay data avoids the 'experiment' frame, potentially removing various threats to validity that are characteristic of the experimental situation such as demand effects (Orne & Whitehouse, 2000). However, relatively few studies are amenable to the 'natural experiments' that existing games constitute. In such instances – if we wish the benefits of using games – we have no alternative but to create them ourselves.

Currently we are held back from creating and using applied games for collecting human-subject data by a lack of models and templates comparable to what the applied games field has developed for other types of data (Quinn & Bederson, 2011). This makes designing such games a risky business as the resulting data might lack the required validity. Further, it prevents people trusting applied-game studies if they do not have a clear idea how such game data should be interpreted. What is needed is a framework for the design and analysis of such games. This will help us better evaluate existing games in the literature. Furthermore, by articulating design principles, it may make it possible to identify novel game designs that open up new methodological opportunities for collecting human-subject data.

1.1.4 Structure of this Chapter

In order to make this argument more fully and thus set up the research questions of this thesis, I need to introduce two areas more fully: 1) the kind of human-subject data collection applied games under discussion, and 2) the issue of validity. First, to clarify the particular kinds of games I will be discussing, I begin by situate the research in relation to the wider field in section 1.2. Then, in section 1.3, I identify within this wider field

a category of games I will call *elicitation games*, of which I survey a number of existing examples. Second, I sketch the concepts of validity and threats to validity in section 1.4.

Having done this, I will identify the research gap in section 1.5 and proceed to state the research questions of this thesis in section 1.6. I will then describe how the thesis will be structured (section 1.7) and close this chapter with a brief mention of ethics (section 1.8).

1.2 Situating the Research

This research is about using applied games for data collection for human-subject research, games which I will label *elicitation games*. Such games are designed for both enjoyment and to collect useful data about human participants. This label corresponds to the intersection of two research areas relating to games: Applied Games and Human-Subject Research with Games.

Applied Games is the design and development of novel games to serve a purpose other than or in addition to entertainment (Schmidt et al., 2015). While related to gamification, "the use of game design elements in non-game contexts" (Deterding et al., 2011, p. 10), a distinction is usually drawn between gamification as involving game elements, while applied games is the development of whole games.

Human-Subject Research with Games groups together a range of fields that use games in order to study human subjects, including games research, HCI, media effects, economics, psychology, and linguistics.

Further, within this intersection I am concerned with the use of games for data collection, particularly for valid data. Data collection is a methodological concern within human-subject research and a desired outcome for a class of applied games. The research presented here is specifically about collecting data, and not about other uses of applied games for human-subjects research, such as structuring participatory design and design ideation (e.g. Kultima et al., 2008; Muller and Druin, 2012; Slegers et al., 2016).

At the intersection of applied games and human-subjects research with games, I will introduce *elicitation games*. This is shown diagrammatically in Figure 1.2. Elicitation games are 1) applied games, thus they are designed for a particular purpose which is to

collect data, and this data is 2) about human subjects.

1.2.1 Applied Games for Data Collection

'Applied games' is the design of games to serve a particular purpose other than entertainment (Schmidt et al., 2015), such as teaching mathematics. It is primarily distinguished from 'applied gaming', which is the use of games *per se* for an ulterior purpose (whether or not designed for the purpose). Thus *developing* a novel game to teach mathematics is an applied game, whereas the *use* of an existing game for the same purpose is applied gaming. As such, the applied games field is primarily concerned with the design of a stimulus or artefact as well as evaluation of the performance of this artefact with regards to the goals that it was designed for. 'Applied games' is a relatively recent term, coined after many more specific terms had been developed for applications of similar principles emerging from particular fields. To add to the confusion, many of these originally specific terms have widened in usage, to become co-extensive with the 'applied games' label itself. One such popular term is 'games with a purpose', originally used for human computation games. 'Serious games' were originally games with an educational intent. 'Gamification', popularly defined as "the use of game design elements in non-game contexts" (Deterding et al., 2011, p. 10), is usually divided from applied games by being the use of game design elements rather than whole games, however the 'gamification' label is also frequently used for interventions that could be called applied games. For clarity, I will use the 'applied games' label throughout until I have introduced the term 'elicitation games', which I explicitly consider a subset of applied games.

The assumption that underwrites the field of applied games is that because games are *intrinsically* motivating they can be used to motivate behaviour that would otherwise require *extrinsic* motivation such as paying participants. There are multiple accounts of this intrinsic motivational quality of games (Boyle et al., 2012; Mekler et al., 2014). Self-Determination Theory (Ryan et al., 2006) models all behaviours as situated on a spectrum from wholly extrinsically motivated – motivated for example by payment or social obligation – to wholly intrinsically motivated – performed without reference to an external goal. Here the locus of autonomy of an action is described as either situated externally to the individual, giving rise to external regulation of the behaviour and loss of intrinsic motivation, or internally, corresponding to internal regulation of the behaviour and feelings of autonomy. In addition to autonomy, two other factors contribute to intrinsic

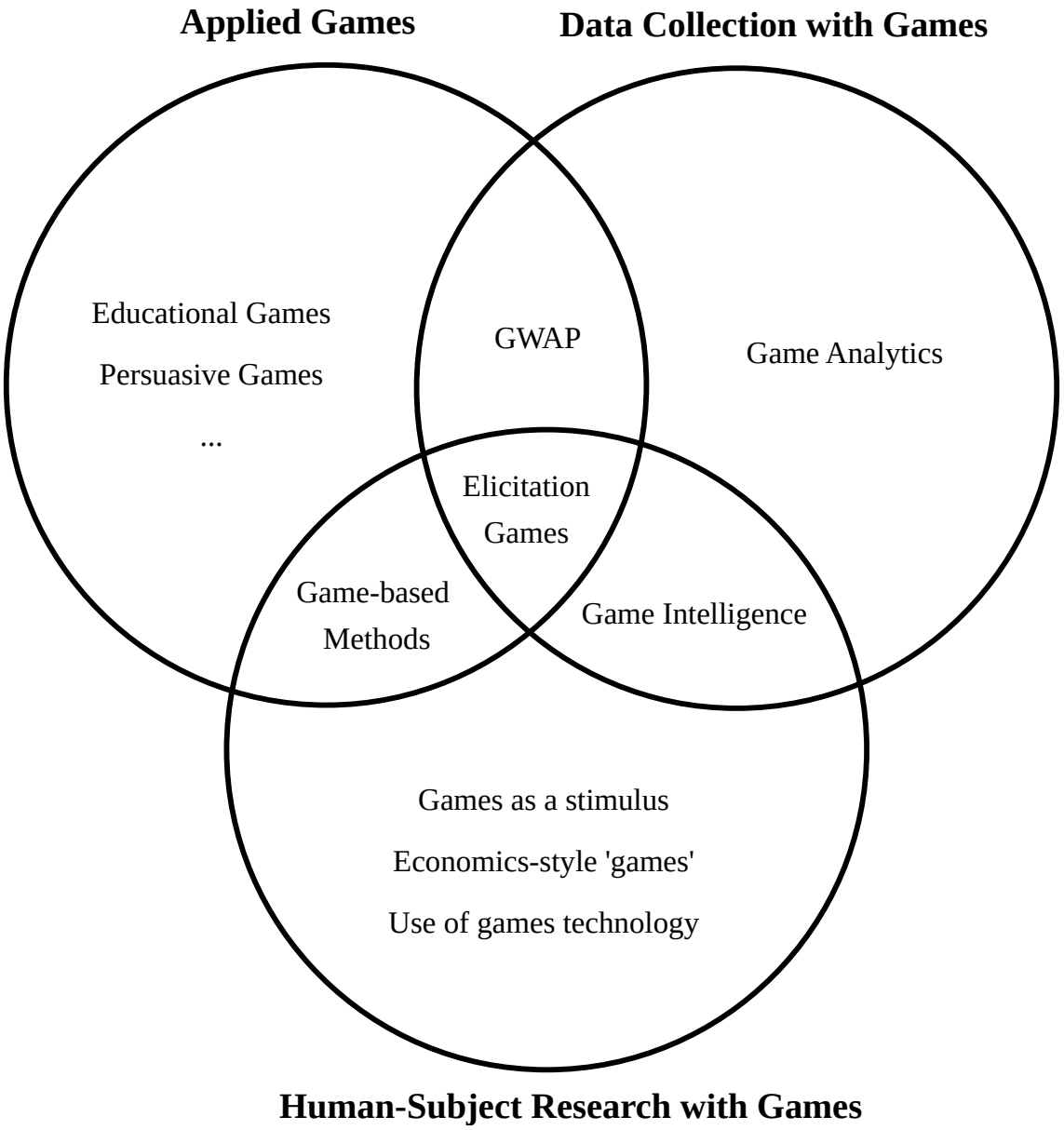


Figure 1.2: Venn diagram of related fields: Applied Games, Data Collection from Games, and Human-Subject Research with Games

motivation: Competence and Relatedness. The theory of Flow (Chen, 2007; Nakamura & Csikszentmihalyi, 2014) describes such actions as autotelic – being performed for their own sake. Autotelicity, combined with the right level of challenge, leads to a 'flow' state, characterised by a positive experience of intense absorption into an activity, loss of sense of time, and so on. When understood through the lens of Frame Analysis, a sociological theory of situational frames introduced by Goffman (1986), the normative social expectations of 'gameplay' are characteristically autotelic (Deterding, 2013). All of these accounts point towards games being enjoyable primarily because of the intrinsic experience of playing them and this enjoyment potentially being disrupted by external demands placed upon the play.

Applied games *for data collection* is the design of games to elicit a particular kind of data through play for a particular purpose. It is often situated within the field of *human computation*, which is a general term for the use of humans to perform computational tasks that would be difficult or impossible for a machine (Quinn & Bederson, 2011); human computation games are such tasks presented in an applied games format. The original examples of human computation games, including the well-known image labelling game *The ESP Game* were presented by Lois von Ahn in his thesis entitled "Human Computation" (von Ahn, 2005). Such tasks might be *crowdsourced*, which refers particularly to the distribution of labour, but there is overlap in that similar concerns of data validation arise. Solutions to this shared problem, such as intersubjective agreement and reputation systems have been proposed (Quinn & Bederson, 2011). *Citizen science* refers to the extension of the scientific research process to non-scientists, often through games or gamified tools which make use of the player's interest in science or desire to contribute to scientific research (Raddick et al., 2013).

Gaming enjoyment is just one of several motives identified for citizen science game participation (Curtis, 2015). These motivations can be classified following self-determination theory into intrinsic and extrinsic motivations (Tinati et al., 2017): intrinsic motivations include competence from solving puzzles, altruism, and community relatedness; extrinsic motivations include achievement, social status, and recognition (e.g. authorship on scientific papers). Community features may be just as important as game design features (Curtis, 2015). While game features help to *sustain* volunteer involvement, participants may be initially attracted to the game because of their interest in science (Iacovides et al., 2013) – to the benefit of charismatic sciences such as astronomy (N. Prestopnik et al.,

2017). However, these studies have been on relatively large and complex games (with complex user interfaces) often requiring significant practice in order to make a meaningful contribution (e.g. Lee et al., 2014). In contrast, most citizen science players 'dabble' (Eveleigh et al., 2014), suggesting that small games with manageable contribution sizes might provide a rich 'long tail' (Dix, 2010).

1.2.2 Human-Subject Research with Games

Human-subject research is a category that groups together diverse fields interested in studying individual human subjects, such as HCI, market research, linguistics, economics, and psychology. When experimental data is collected in such fields, it is characterised by being *about an individual* (and their consumer preferences, linguistic judgements, deviations from rationality, etc.) While we typically generalise such claims as relating to a wider population, e.g. movie-goers, English speakers, or market participants, such generalised claims are generally derived from aggregated data collected about individuals.

Scholars have identified trends in using games, or game elements, as part of human-subject research and HCI (Deterding et al., 2015; Slegers et al., 2016). Games may serve as ideation tools, discussion prompts, and design research, as well as experiments and data collection tools. It is important to note that usage of the term 'game' in human-subjects research has typically been broad and non-prescriptive, and for present purposes will require narrowing. It commonly refers to:

1. The phenomenon of **entertainment games and gameplay** as an object of study, studied experimentally for e.g. player experience (Cutting & Cairns, 2020), media effects (Zendle, Kudenko, et al., 2018), and psychophysiological effects (Kivikangas et al., 2011) of gameplay.

2. An experimental paradigm that involves **arbitrary rules for behaviour**, such as word games (S. R. Baum, 2002; Treiman, 1983) – using artificial morphological rules for studying linguistic structures – or formalisable incentive structures in economics games (e.g. Fehr and Gächter, 2002).

3. Entertainment games as a **useful operationalisation of a motivating task** in the study of e.g. motivation (Isen et al., 1978; Sharek & Wiebe, 2011).

4. The identification of an experimental task as a game used a **metacommunicative framing device** through language or an adoption of surface features, commonly

used as a way of engaging participants (particularly children) (Washburn, 2003)(e.g. (Soboczenski et al., 2016)

5. The use of **games technologies** or interfaces characteristic of games as a methodological tool, regardless of the participant's experience of using those technologies (Frey et al., 2007; Slater et al., 2006; Washburn, 2003).

As can be seen, 'games' can be used as a methodological tool to structure (1, 2, 3, 5) and motivate (3, 4) experimental tasks. As such, the primary concern for the use games or game technologies used is not to be enjoyable, well-designed games, but to contribute to the *sine qua non* of experimental research: validity. Satisfying necessary validity claims need not require the use of a whole game. Indeed, given the variance, complexity, and social norms associated with games (as will be reviewed in chapter 2), it seems likely that the use of a whole game might indeed be directly counter to validity goals.

The applied games field, in contrast, sets its goal at creating games comparable to entertainment games that are not only be perceived to be, but actually are enjoyable. Of course, as in this thesis, applied games are also studied and such studies and experiments must also be concerned with validity (we want to ensure, for example, that it is *our game* responsible for an effect observed and not an experimental confound). However, studies on applied games are particularly concerned with *ecological validity*: if applied games are supposed to be intrinsically motivating, then for us to be able to generalise our results from an experiment to applied games in general, our game must also be intrinsically motivating.

Most categorisations of games in experiments look at how games are used, such as using games to manipulate variables or gaming as a performance metric (Washburn, 2003). Here, in contrast, we will be particularly interested in how the game relates to the data that is ultimately collected. Within the experimental situation, this gives only two ways in which games can be used. First, and most commonly, games are used as a stimulus. Second, they may be used as a measurement instrument.

Games used as stimulus or manipulation present a task as an independent variable or a stimulus, perhaps with the intention of contrasting this with an alternative stimulus. This may be to study the effect of (some aspect of) gameplay on, e.g. affect (Isen et al., 1978), priming of aggressive concepts (Zendle, Cairns, et al., 2018), or attention (Cutting & Cairns, 2020). Or it might be to study the experience of gameplay itself (e.g. Cairns et

al., 2006). While this is commonly achieved with self-report scales, it can also be achieved in real-time with psychometric instruments (Kivikangas et al., 2011).

Here the primary concern is that the game suitably operationalises the desired stimulus. If this is achieved, the games might relatively freely be existing (McMahan et al., 2011), modded (Elson & Quandt, 2016) or novel (e.g. Zendle, Kudenko, et al., 2018), depending primarily on needs for ecological validity, ease of development, and need for fine-grained control. If some aspect of the gameplay experience, such as contriving win/loss (Isen et al., 1978), or wider context, such as distractors (Cutting & Cairns, 2020), is to be manipulated, this may constrain the suitable games[1].

Importantly, here the experimental data is not encoded within the gameplay. It is the player who mediates the relationship between the gameplay (as an experience) and a separate measurement instrument (e.g. survey, neurophysiological measure, etc.). As such the kind of game may be relatively interchangeable (if violent games *per se* cause an effect, this should be observable with any violent game), and certainly it is unusual that fine-grained subtleties of the game mechanics are likely to invalidate the data collected.

Games used as measurement instrument are responsible for operationalising a measurement of a dependant variable in the study design, while necessarily also acting as stimuli. Performance at the game as a whole may be the dependant variable (e.g. Arnold, 1976; Jones et al., 1981; Kennedy et al., 1981), measured via some abstraction of gameplay, such as game score. For another class of games, data arises from specific game actions. In gamified reaction time experiments (e.g. Berger et al., 2000; Dunbar et al., 2001) the data derives from the timing of actions within the game. Another example is a study on number-entry error (Oladimeji et al., 2012), where the data of interest is the accuracy of interactions with various number-input devices in a virtual hospital ward. The process of measurement may be automatic (e.g. through game telemetry) or manual (e.g. noting down the score of each trial).

Here what is being observed is not the effect of the game on the player, but the effect of the player on the game, analogous to the way temperature affects a thermometer. However, unlike a thermometer, the game also significantly affects the player in a dynamic loop. The actions a player takes are likely to affect the state of the game, affecting their future actions, perhaps making certain actions either necessary or impossible. Player performance

[1] Järvelä et al. (2012) and Järvelä et al. (2014) provide guidance for using such games.

is liable to lead to vicious cycles (failing once makes the game harder), or virtuous cycles (succeeding once makes the game easier), or be affected by dynamic difficulty balancing. Finally, nuances of the game might easily invalidate the data collected and the games are not likely to be interchangeable. It is not surprising that most games used in this way are rather simple (e.g. Oladimeji et al., 2012; Soboczenski et al., 2016).

Finally, while not experimental in nature gamified surveys and assessments are another way human-subject data is collected with game(element)s. Gamified surveys use techniques of gamification (typically points, leaderboards, etc.) to motivate survey participation. Such approaches tend to find that gamification increases motivation but does not necessarily impact behaviours such as satisficing, omitting items, or abandoning surveys, and with those, data quantity and quality (Keusch & Zhang, 2017; Lumsden et al., 2016).

1.3 Elicitation Games

At the intersection of the two fields of applied games for data collection, and human-subjects research with games sits a set of games that have a number of useful properties for developing scalable human-subject data collection tools. By games, I refer explicitly to *whole* games and not simply the use of game design elements as in gamification. Elicitation games, as I define them, have the following three necessary conditions:

1. As applied games they should be **designed for enjoyment** as a primary goal (and ideally *be* enjoyable) and thus in principle likely to encourage voluntary play

2. The gameplay itself should be **designed as a measurement instrument** for some meaningful data **about their players as individuals**[2]

3. This data should be in a form that is (or can in principle easily be) **recorded and transmitted to researchers**[3]

Games that possess all of these properties are *elicitation games*. For example, an elicitation game might present a psychometric task (2) that is designed to be enjoyable to

[2] A game does not need to be successful to be an elicitation game, though obviously that is to be preferred. An elicitation game need not be *in fact* enjoyable, though it should be designed for enjoyment. Similarly, the data it collects might not be valid, so long as it was designed to be a measurement instrument. However, while such games game might be elicitation games, they would not be *good* elicitation games

[3] An elicitation game, might for example, require an experimenter to set up an audio recorder. It would be strange to say that the game commences being an elicitation game only when the record button is pressed.

complete (1). Relevant data about the performance of this task is uploaded to a database (3). An elicitation game might be an enjoyable (1) real-world game whose performance is an expression of individual players' orienteering skills (2), which is designed so relevant information about the player's orienteering skills is captured in a recordable form through play (on paper, GPS tracking, etc.)

Such a game need not in principle be digital, although a digital game would likely be easier to scale. If the above psychometric game was an analogue card game, the final measurement might be encoded within, for example, the particular order of cards in the player's deck at the end of the game. It might be expensive to pay for each participant to post their deck back to the researchers, but were the game to be deployed in a physical space with an experimenter attendant, this would not be a concern.

Diverse examples of possible elicitation games under this definition include a carnival high-striker as a measure of upper-body strength, a play-by-post combat-strategy game as a measure of risk-taking behaviour, and a cooperative party game as a measure of particular personal identify claims and beliefs (though the effectiveness of each would of course need to be justified). Elicitation games could be used to survey or compare populations, or automatically assign players to one or more experimental manipulations. For example, an online game involving reading words on screen could automatically assign players to conditions containing a variety of different typefaces and colours in order to e.g. compare performance between languages.

Educational games often incorporate assessments of an individual's performance, which is data that cannot be validated against an external model or intersubjective consensus. If such assessments are themselves delivered as a game or portion of gameplay, they can also be considered to be elicitation games, and they have the same concerns: design of an appropriate stimulus, ensuring validity of the data.

1.3.1 Survey of Elicitation Games

While the term 'elicitation game' is new, several existing games meet the criteria above. These are surveyed below. The games discussed are summarised in Table 1.1. For the purposes of the discussion below, I have categorised elicitation games based what they reward their players for. In-game rewards guide player actions, and are thus both are a key feature of game design and a key potential threat to validity. The games identified either reward players for their level of performance at a challenging task, reward players

Game	Mechanic	Reward
Sea Hero Quest	(Wayfinding task) Navigate through memorised map	Task Performance
	(Path integration) Shoot flare back to buoy	Task Performance
The Great Brain Experiment		
— Working memory	Click memorised positions in grid	Task Performance
— Attentional blink	Identify second target image presented in rapid series	Task Performance
— Selective stop signal	Tap only ripe fruit falling past marker	Task Performance
— Decision making	Choose between lotteries with different payouts	Payout
	Mark happiness on line	None
'Number entry'	Enter numbers quickly	Task Performance
'Spaceships'	Move to memorised location	Task Performance
Text Text Revolution	Type words on smartphone	Task Performance
Dragon Master	Pick word matching definition	Knowledge
Literate (Ekapeli)	Identify orthographic form of phoneme	Knowledge
BeFaced	Make facial expressions that match tiles to clear	Comparison to Model

Table 1.1: Elicitation games surveyed with their primary data-eliciting mechanics, the majority of which provide performance-conditional feedback

for expressing knowledge, or – in one case – reward players based on a lottery payout. I also identify a game that rewards players based on a model (but is not a solution-based game) and an edge case whose rewards are partially based on an intersubjective consensus.

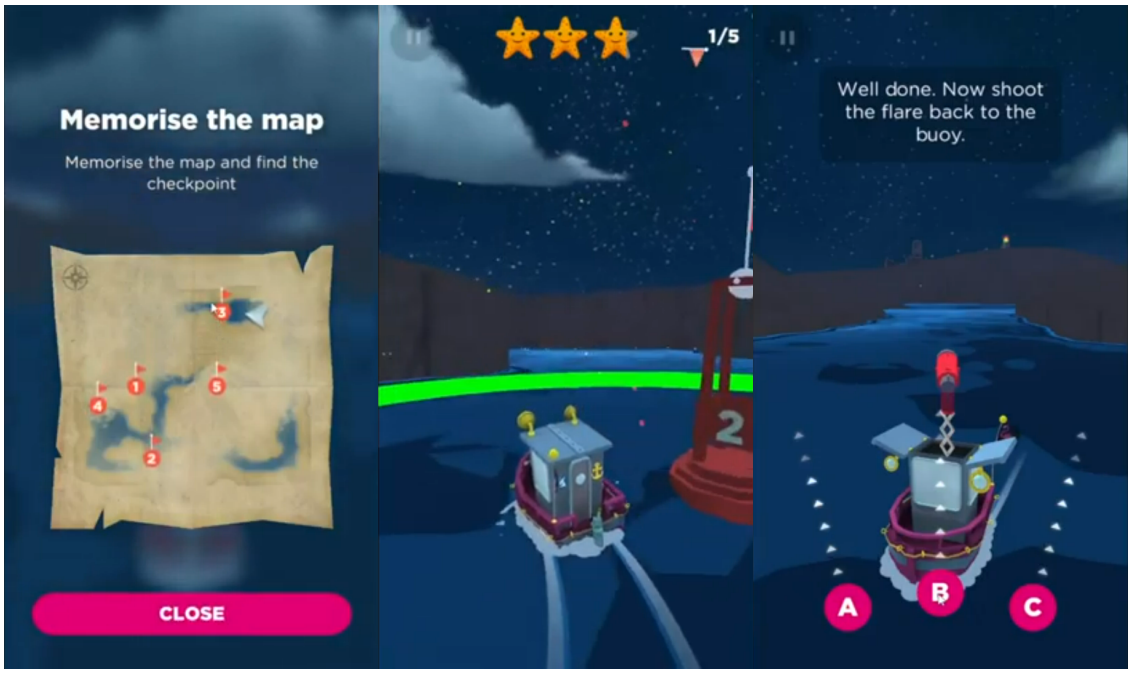
This survey is limited in two ways. First, many game methodologies are poorly described, which makes identifying and describing suitable examples challenging[4]. Second, where game use is not the focus of the paper, it is common that the use of a game methodology is mentioned only within the methods section, and such papers are thus harder to find in a literature search.

Task Performance *Sea Hero Quest* (Coutrot et al., 2019; Spiers et al., 2021) was a game that collected a mass sample of spacial navigation data from 4.3 million players

[4]While this may be because such examples originate in fields that are not interested in studying games themselves, this is a problem for reproducibility, and also – especially in the light of chapter 2 – conceals potential threats to validity. Studies that report using a game rarely clearly present the game *as a game*.

to contribute to dementia research (Alzheimer's Research UK, n.d.). In *Sea Hero Quest* players pilot a boat towards targets memorised from a map (Figure 1.3). The path that a player takes through the level is recorded. Additionally, some levels include a 'path integration' task where upon reaching a target, the player must judge the straight-line direction back to their starting point, out of a choice of three options. There is also a third type of level that does not collect data, where players pursue and photograph a sea monster in a *Temple Run*-style (Imangi Studios, 2011) game. Players are rewarded for efficiently locating all the targets and choosing correctly in the path integration task by a three-star rating system for each level.

Sea Hero Quest is an elicitation game as 1) significant investment was made to make it enjoyable; 2) the data it collects (e.g. players' paths through each level) is a reflection of players' individual competences; and 3) this data is remotely collected from the game.



(a) Level map (b) Arriving at a target (c) Path integration task

Figure 1.3: Screenshots of *Sea Hero Quest* in the Kano Reef levels. Most levels begin by players being shown a map of the level and locations of targets; in this level the map is partially obscured. Players navigate through the 3D environment to find a series of targets. In some levels, players must shoot a flare back to their starting point.[5]

The Great Brain Experiment (H. R. Brown et al., 2014) groups together four cognitive science experimental tasks within a smartphone game designed to be enjoyable and engag-

[5]Screenshots from video by *shachar700* https://youtu.be/k2MtVU4A0c4

ing. Three of these games: a working memory task, a selective stop signal task, and an attentional blink task, shown in Figure 1.4a-c, assess a player's performance at a particular task in terms of the accuracy they achieve. The tasks differ. The working memory task requires them to memorise and enter the positions of red circles in a 4x4 grid. The selective stop signal task requires them to tap the screen to catch falling fruit (but not tap if the fruit turns rotten as it falls). The attentional blink task requires them to identify an image that is presented in a rapid serial visual presentation. The target image is the second in a given category to be presented and it is this they must identify at the end of the trial from a grid of four options. In each of these games, performance-conditional feedback is provided to keep them on task. The working memory task rewards accurate recall, the selective stop signal task rewards responses within a time window, and the attentional blink task rewards getting the correct answer.

Similar to *Sea Hero Quest*, *The Great Brain Experiment* was designed to be an enjoyable game in order to achieve large-scale participant recruitment. It satisfies the other requirements of being an elicitation game: the data, being performances of typical psychological tasks, is clearly about the individual; and the data it collects is transmitted to researchers.

(a) Working memory (b) Attentional blink (c) Selective Stop Signal (d) Decision making

Figure 1.4: Screenshots of *The Great Brain Experiment*, showing each of the four minigames.

The game 'Number entry'[6] (Oladimeji et al., 2012), shown in Figure 1.5, requires the player to enter input (numbers or text) in a time-pressured context in order to observe

[6]Where a name is not given in the cited paper, a descriptive name is given in quotation marks.

error rates for different number entry interfaces and conditions. A game is desirable to motivate enough gameplay to observe infrequent errors. In the game, the player must keep hospital patients alive by entering correct numbers into multiple simulated number entry systems. The player clicks between different patients, each with a depleting health bar. For each patient, the player must set up an infusion pump by entering both a specified quantity and the rate into a simulated number entry device. Once successful, the patients health temporarily improves. Players are rewarded with score increases for correct entries with a bonus proportional to speed. Game features such as high scores and levels are included.

Again, this game is an elicitation game because it was 1) designed for enjoyment in order to maximise the overall data collected; 2) designed to collect data about the individual player: their own rate of making number entry errors and not, for example, their estimate of the average rate of the population; and 3) designed to record and transmit this data.

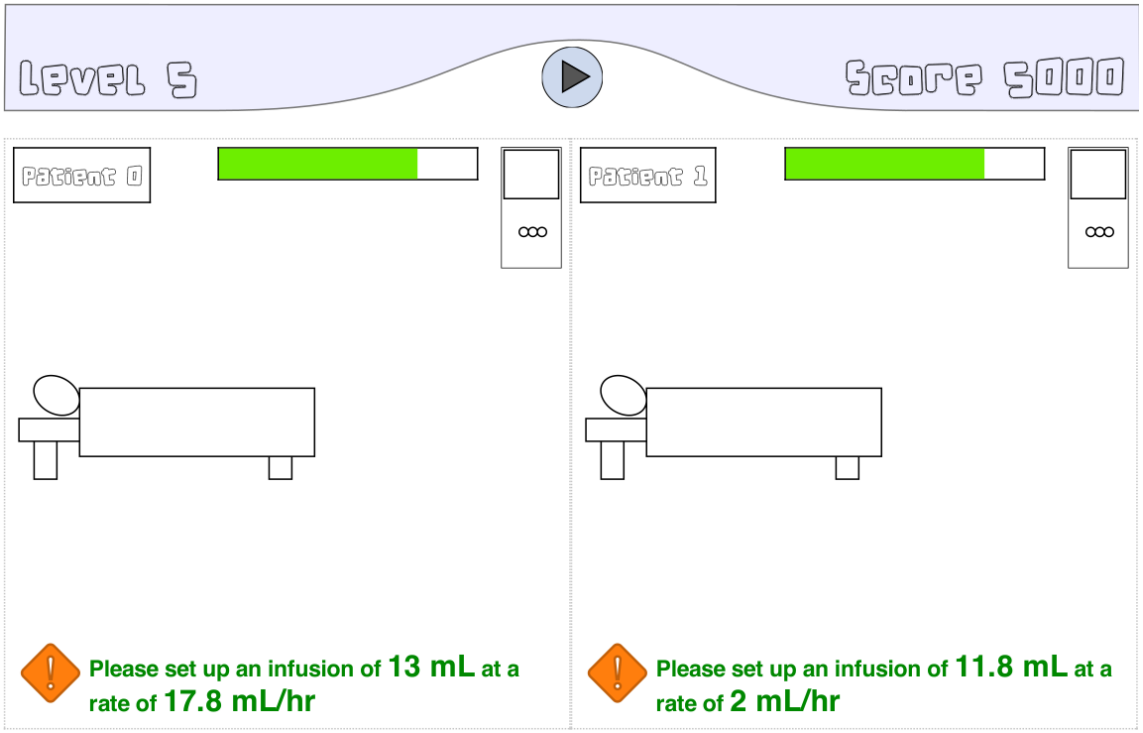


Figure 1.5: Image of the number entry game from Oladimeji et al. (2012). In this level, the player has to keep two patients alive in a virtual hospital. The green bars represent the time left to enter the values at the bottom of the screen into the simulated number entry device for that patient, shown in another game view. Fast accurate input scores points for the player.

The game 'Spaceships' (J. P. Spencer & Hund, 2002, 2003) – though not presented as a game *per se* – is a spacial memory task involving moving magnetic tokens on a physical

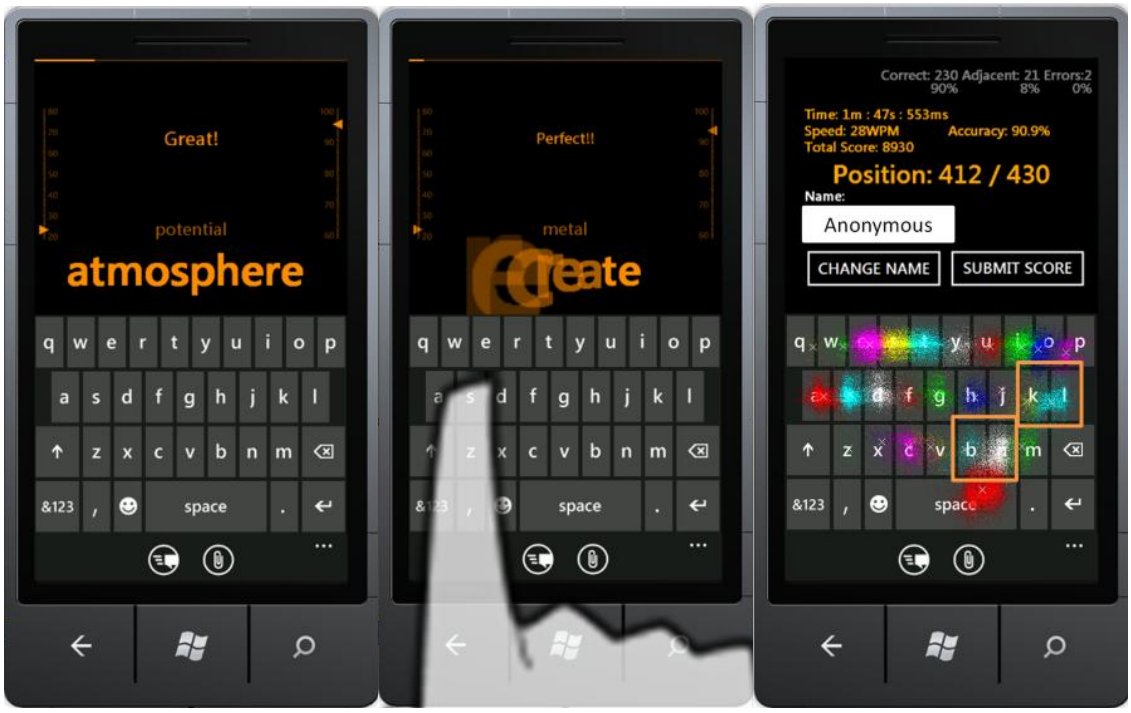
tabletop to the remembered locations of enemy "spaceships" (places on the table that were illuminated)[7]. Participants see an illuminated space on the table, then have to look at a fixation light for a period of time. Then on a trigger, they slide their magnetic token to where they think the light had been (where the spaceship was). Several kinds of game feedback were given to players, motivating it's inclusion in this survey: initiation time was graphically displayed; a score rewarding short initiations and high accuracy was given; total accumulated points over all trials, and a "flight rank" based on score, congratulatory messages on well performed trials, warnings on poor initiation times, and stars for every 80 points achieved. This is also an elicitation game as: 1) from the papers cited, this game appears to have been designed for enjoyment; 2) it is the individual's competence at spacial memory that is measured; and 3) the experimental setup was able to record participant data.

The smartphone typing game *Text Text Revolution* (Rudchenko et al., 2011), shown in Figure 1.6 models itself on other smartphone typing games, consisting of players typing a series of words using then on-screen smartphone keyboard. Thus it appears to have been designed for enjoyment, and collects individualised typing data. Players are shown a bar representing their current words-per-minute and their accuracy of hitting the correct keys. At the end of the game, players are shown a scoreboard with their personal best scores and the top scores of any user. Because the game knows which key the user is supposed to be pressing next, it can compare the location of the key to the actual screen touch points detected, for that particular user. This individualised touch point data is transmitted to the cloud for analysis.

All of these games reward performance at a task based on accuracy and/or speed. As the inputs made by the player can be compared against the inputs demanded by the game, the player's performance can be assessed. This performance-contingent reward motivates players to perform the desired task, be it number entry or spacial navigation, quickly and accurately, which is the task behaviour desired in the experiment.

Payout In the decision making sub-game in *The Great Brain Experiment* (Figure 1.4d), players start with 500 points. They spend these points in a series of choices between a certain payout and a lottery, and gain or lose as a result. In each case, players may choose

[7]While the cited papers contain diagrams of the physical setup, these do not show any of the game elements.



(a) (b) (c)

Figure 1.6: Illustrations of *Text Text Revolution* from Rudchenko et al. (2011).

to gamble a certain number of points for the chance of a greater reward (or a smaller loss). The number of points won or lost in the gambling game are accumulated by the player, whose goal is to get a high score. Here the data of interest are the choices that the player makes, under the assumption that these map to real-world risk-reward trade-offs.

This game also contains a self-report question "How happy are you right now?", which appears between every 2-3 game rounds. Players answer by marking a point on a line. Responses to this question are not rewarded nor is it integrated into the other game mechanics.

Knowledge The games *Dragon Master* (Metcalfe et al., 2009) and *Literate* (Lyytinen et al., 2007) collect data that assesses the correspondence between a player's knowledge and a ground truth[8]. In *Dragon Master*, this is part of an experiment to assess the effect of delayed feedback on learning – the ground truth is the words and definitions introduced by the game – where as *Literate* is an intervention designed to identify dyslexia and improve grapheme/phoneme recognition, for which the ground truth is standard Finnish orthography. *Dragon Master* gives the player a definition and they need to type the word it

[8]The papers cited provide no screenshots for these games

corresponds to. *Literate* presents a phoneme and the player must click the corresponding orthographic representation as it falls past the screen. In each case, accuracy to ground truth (entering the correct word, selecting the correct grapheme) is rewarded, and the successful player proceeds through levels of increasing difficulty.

In these two games, the human-subject data being collected is the individual's knowledge of a set of assumed facts, in contrast to the player's unbiased intuitions. For example, a player of *Dragon Master* presented with a definition such as "words or letters written, printed, or engraved on a surface" (Metcalfe et al., 2009, p.1081) might naturally respond "writing", but choose to instead respond with the expected "inscription" in order to earn more gold coins. Performance-based feedback motivates players to do their best, which is what is desired here. However, were the unbiased response desired, this would constitute a threat to validity. They satisfy the other requirements of being elicitation games by 1) being apparently designed to be enjoyable games; and 3) recording the player data for the researchers.

Conformity to model *BeFaced* (Tan et al., 2014) (Figure 1.7) is a camera-controlled tile-matching game for the iPad that is designed for collecting images of the players' faces and/or tracked feature point data (Tan et al., 2013) while they are making a range of facial expressions. Players see a grid of tiles. By swiping a player can swap the positions of tiles as in the game *Bejeweled* (PopCap Games, 2001). Once a group is made the player must make a facial expression matching the faces on the tiles to clear the group. A classifier is used to check that the player's facial expression matches the type of tile to clear. If the player's facial expression matches the target expression with sufficient probability according to the classifier, the tiles are cleared. *BeFaced* is discussed in more detail in chapter 4 where it is analysed using the theoretical framework developed in this thesis: Intrinsic Elicitation.

The feedback mechanism rewards players for producing facial expressions (in the form of images or feature point data) that match the classifier's model. Thus here we have an elicitation game that, similar to a solution-based game, uses a model to evaluate the quality of a players' provided data. *BeFaced* satisfies the requirements of being an elicitation game, as defined above. First, it is 1) designed for enjoyment, with gameplay inspired by the popular tile-matching game *Bejeweled* (PopCap Games, 2001). 2) It is designed to collect the facial expressions of individuals. Finally, 3) the facial expression data is captured from the camera video stream and labelled with the expression on the tiles. It then uploads

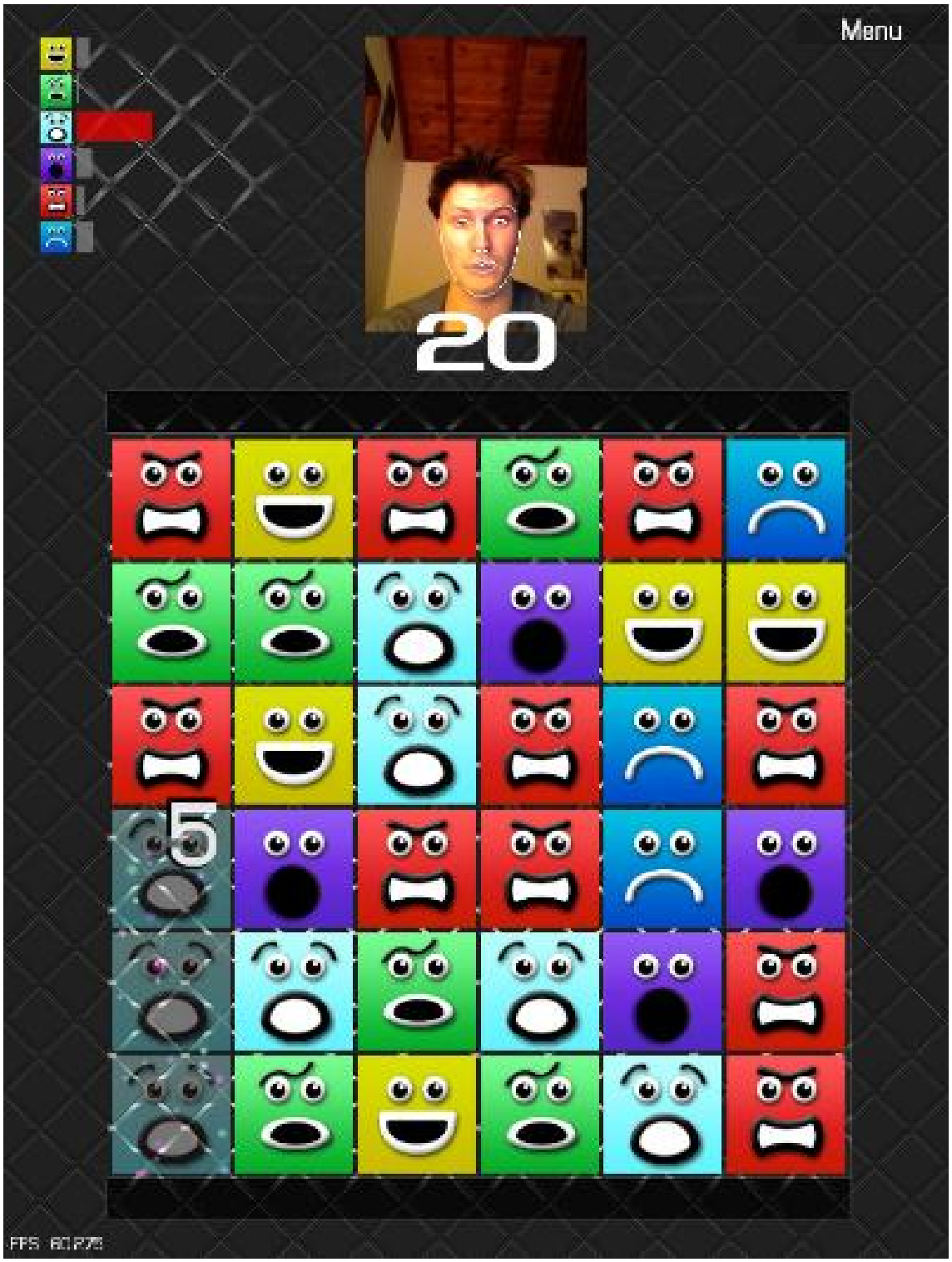


Figure 1.7: In *BeFaced* players match groups of tiles and then need to clear them by making a facial expression that matches the tiles to clear. Image from Tan et al. (2013)

these labelled images to a cloud database for subsequent collation and reuse.

BeFaced's use of a model to drive feedback should not be taken as evidence that solution-based game designs (similar to *FoldIt* (Cooper, Khatib, et al., 2010)) are generally

applicable to the design of elicitation games. Here it is rather that, by satisfying the facial expression classifier, the player must unavoidably provide additional data: their particular facial landmark data or photograph of their face. The model could not, of course, check that the image captured by the camera is actually *your* face. In analogy to other solution-based games, the classifier poses a 'problem', to which the player's facial landmark data is the particular solution they provide. While the classifier can check a player's solution is a valid solution to the problem (the expression matches the target expression), the data of interest is in fact *how* the player solved the problem, being the particular facial landmark data of their face. As discussed in chapter 4, the appropriateness of the solution-based template is limited to particular kinds and uses of human-subject data, which *BeFaced* happens to satisfy.

An Edge Case using Agreement *Apetopia* (Barthel, 2013) (Figure 1.8) is a game that collects data about how humans perceive the relative similarity of different colours. You speed through a gritty urban environment moving left or right to dodge bombs and heaps of rubbish and to collect coins to increase your speed. The goal of the game is to travel as far as possible. The game ends if you lose all of your health, which happens when you collide with obstacles. At regular intervals in the game, you are forced to choose between moving through one of two coloured gates. The instructions direct you to choose the coloured gate that most closely matches the colour of the sky. A 'correct' choice leads to a speed increase, as does collecting coins. If it is assumed that players follow the instructions of the game, players' choice of gate encodes information about which of the two colours they perceive as closest to a third colour.

Apetopia is discussed in more depth in chapter 4. For now, I will briefly summarise some key points. The main rewards in *Apetopia* are tangential to data provision: players are rewarded for collecting coins and punished for hitting bombs. Apetopia uses a heavily obscured agreement mechanic to reward players for the data they provide. Where a strong consensus exists amongst players, players are rewarded for conforming with this and punished for disagreeing. However, in part due to the infrequency that this happens, it is almost impossible for players to determine what causes the rewards or punishments. This means that this mechanic contributes little to the enjoyment of the game.

The purpose and analysis of the data is to build a single model of (apparently universal) human colour similarity (Barthel, 2013), not to model an individual's colour perception.

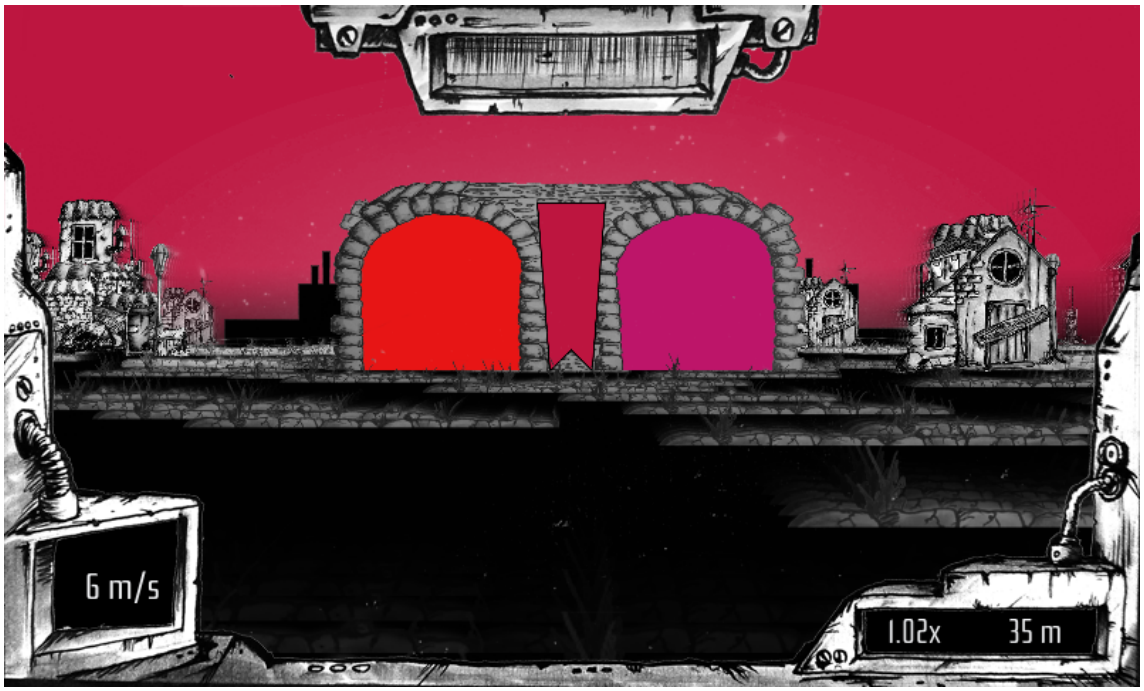


Figure 1.8: In *Apetopia*, the player has to choose between differently coloured gates.

Given the use of agreement mechanics it would seem most likely that this data is an intersubjective consensus (if the goal is a best consensus human perceptual colour model). However, the difficulty arises in that the agreement mechanism used in the game is obscure. If players remain unaware of this mechanism, clearly the data that is collected is also judgement about their individual colour perception, though if players become aware of the mechanism this claim is less convincing.

1.3.2 Examples that are not Elicitation Games

In order to more fully illustrate the definition of elicitation game given above, I will here give a number of examples of games that violate one of the three constraints, being that 1) the game is designed for enjoyment; 2) the game collects data about individuals; and 3) this data can be collected and transmitted to researchers. First I will contrast elicitation games with gamification.

Gamification 'Elicitation games', as I have introduced the term is intended to refer to whole games and not simply any use of game design elements, or any task that may be enjoyable. Non-game tasks that are enjoyable clearly do not involve game design and are thus easy to exclude. Gamification, on the other hand, *does* make use of some game design elements yet not in a way to give rise to a cohesive game as a 'whole art form'.

Bubble Trip (Levy et al., 2016) (Figure 1.9) is a gamified assessment task which adopts some game features without, in my opinion, constituting a whole game. Levy et al. (2016) describe the design of the game as being the result of extensive prototyping and playtesting to make an engaging game. However the game itself appears to be the surface addition of game features relatively unintegrated with its underlying questionnaire task. In *Bubble Trip* players control a fish. The player collects points by answering questions (regardless of their choice). They also get points for collecting bubbles. The players only meaningful choice is between whether to collect the bubbles or answer the questions – allowing the player to choose *not* provide data. A quote from the paper is illustrative: "some participants may have been more engaged with the game, spending time collecting bubbles and swimming around, whereas other participants may have been more focused on completing the task quickly, avoiding interacting with the game mechanics" (Levy et al., 2016, p. 12). It is notable that the paper does not report any evaluation of how enjoyable participants found the task, or whether participants perceived it as a game.

Figure 1.9: In *Bubble Trip* players control the fish to collect bubbles (left) avoid jellyfish (right) and answer questions by moving to one of the shells (top). Image from Levy et al. (2016).

The game is not designed for enjoyment as a primary goal Many experimental tasks collect data about individuals and are designed for enjoyment to some extent. While

I have already dealt with gamification above, some interventions might arguably constitute novel whole games, however well or poorly conceived. Considering the importance placed on enjoyment in the design provides an additional heuristic to compare borderline cases of experiment games. My hope here is to reinforce the ambition for elicitation games to be genuinely enjoyable games (and thus benefit from the assumed motivational benefits thereof).

In the definition of elicitation games given above, enjoyment must be a 'primary goal'. If this is the case, we should assume at least as much care has go into design for the enjoyment of the game as for the validity of the data. I suggest that this be tested as follows. Consider if there are any obvious design changes that would significantly improve the enjoyment of the game. If there are such, and these would have limited or no effect on the apparent validity of the resulting data, this suggests that enjoyment was not a primary goal.

Friehs et al. (2020) take a stop signal task and develop a version using 3D graphics, animations, theming and a consistent fiction. In this game-like task, the player is walking through a forest and needs to choose whether to go left, right or straight on at each crossroads by a keypress. At each junction a helpful fairy points in the correct direction, however the player is told that a beep preceding the fairy's appearance indicates that it is an evil witch in disguise. A correct or incorrect response makes no difference to the game. While there are rules to follow there are no goals and no sense of progress. There is no reward for correct or incorrect answering. Beyond satisfying the experimenter's request in the form of the initial instructions, there is little to keep the players providing the correct data.

The game was likely not primarily designed for enjoyment as there are various easy ways to make it more enjoyable even if distracting game elements like scores are to be avoided. First, the player could be rewarded (or punished) for each choice with sound effects (a witches cackle, for instance), animations, and other juicy feedback. A goal and sense of progress could be easily added. For instance, the forest could get darker and more menacing with each wrong turn, or lighter with each correct choice, with the player ultimately escaping or getting stuck.

The cognitive psychology experiment-cum-game *Ghost Trap Experiment* (Hawkins et al., 2013) (Figure 1.11 has the player catching ghosts by choosing each turn whether to check the ghost trap in the left of right room. The goal of the game is to catch as many

Figure 1.10: In the stop signal game from Friehs et al. (2020), the player (centre) chooses whether to go left or right at each junction. A fairy (top) appears to show the correct direction (here, pointing left).

ghosts as possible (letting as few as possible escape untrapped). When there is a power cut, the players must continue attempting to catch the ghosts, despite no longer being able to see their positions.

While designed for enjoyment and making use of a potentially enjoyable core mechanic, there is significant experimental framing and instruction around the game and this experimental context is integrated into how the game is presented. This suggests the integrity of the game as an (in principle) standalone artefact was not prioritised. For example the score and the progress through the game is presented with the experimental text, not the game view. By removing the explicit experimental framing and making some layout design changes (integrating the score, instructions, and progress into the game in a sympathetic way) it might be easily redesigned to prioritise enjoyment, and be thus considered an elicitation game.

The data is not about individuals *Peekaboom* (von Ahn, Liu, et al., 2006) is a game for collecting data for computer vision algorithms. It is a multiplayer game which uses an agreement mechanic. One player reveals sections of an image. The other player needs to guess the correct word. The revealing player can 'ping' points on the image which are highlighted to the other player.

While *Peakaboom* is 1) designed for enjoyment, it is not intended to 2) collect mean-

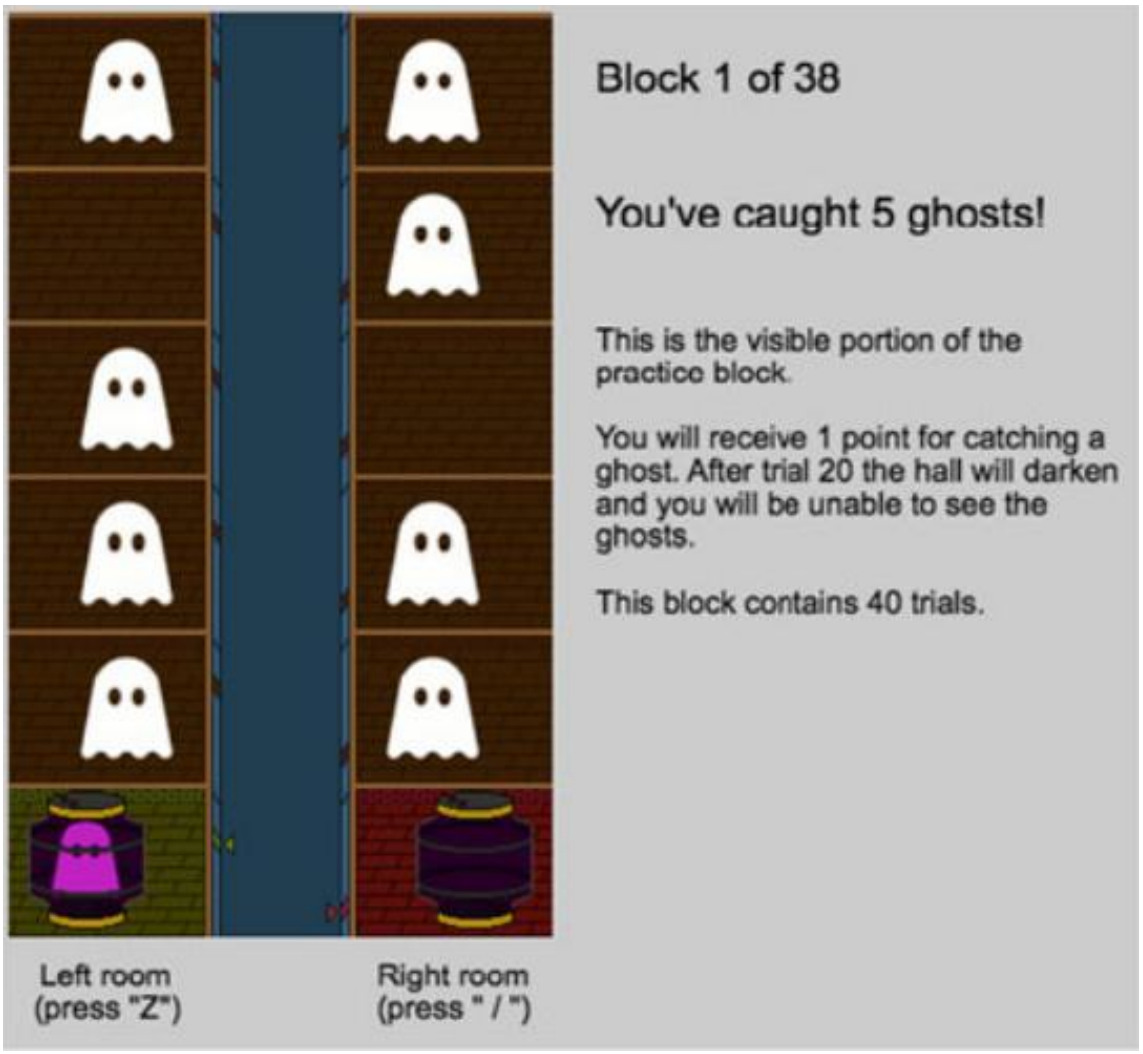


Figure 1.11: In *Ghost Trap Experiment* players need to check their ghost traps by choosing between the bottom left room and bottom right room. Ghosts move down through the rooms from the top. Image from Hawkins et al. (2013).

ingful data about its players as individuals. The data that it is designed to collect is a consensus between players. While it does collect data about individuals to construct this consensus (how individual players make guesses when shown sections of image, for example), this data is not itself the intended outcome of the game. Finally, 3) it is able to record and transmit this data to researchers.

Verbosity (von Ahn, Kedia, et al., 2006) is a game that uses an agreement mechanic for collecting common sense facts that players widely agree on. One player is trying to guess a secret word shown only to the other player. The other player completes fill-in-the-blank clues, such as "It contains a __", with any word that does not contain the target word. It is therefore optimal to give truthful clues. Similar to *Peakaboom*, while the game is designed

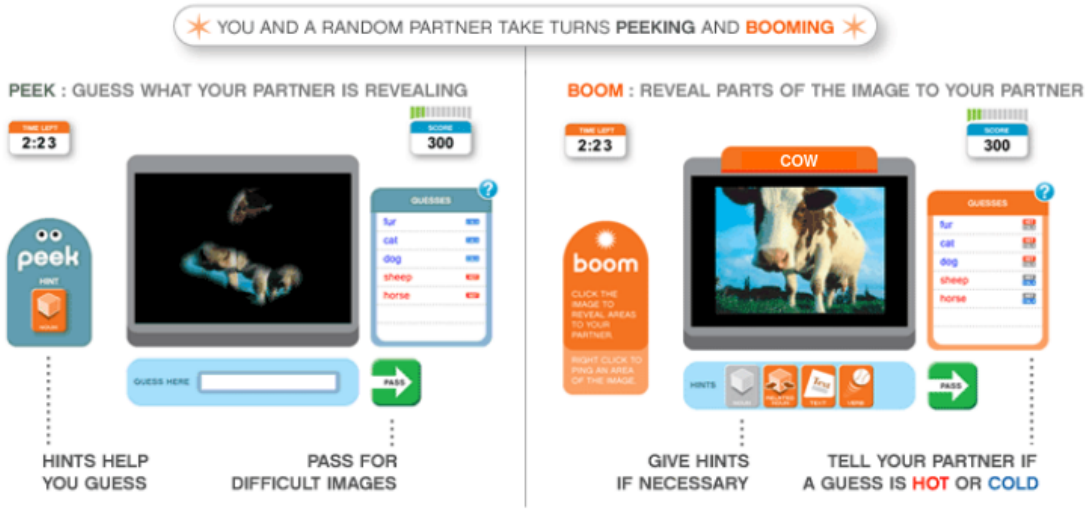


Figure 1.12: In *Peekaboom* one player 'peek' sees an image and tries to guess what it is, while 'boom' chooses parts of the image to reveal. Image from von Ahn, Liu, et al. (2006)

for enjoyment, and transmits its data to researchers, the data it is designed to collect is not meaningful data about the players as individuals, but about 'common-sense facts'. As with *Peekaboom*, it is unclear whether the data from an individual as being about that individual is meaningful. While it might be adapted by researchers interested in what individuals (*qua* individuals) think are common sense facts, this was not the intention of the designers.

The data is not recorded or transmitted The *Brain Age* (Nintendo, 2005b) or *Dr. Kawashima's Brain Training* series are collections of puzzle games for the Nintendo DS. These present tasks such as sodoku, performing a series of simple maths problems as fast as possible, memory tasks, and the Stroop test. *Big Brain Academy* (Nintendo, 2005a) is a puzzle game by Nintendo, similar to *Brain Age*.

For both of these games, and others like them, while they game can presumably be said to be 1) designed for enjoyment, and while they do 2) collect meaningful data about their players (which *Brain Age* presents to the player as their 'brain age' between 20 and 80, for example), these games do not (to my knowledge) 3) transmit this data to researchers. While there have been studies performed on the effectiveness of such games (Simons et al., 2016), such studies consider the game as a stimulus rather than as a measurement instrument.

Were these games to be extended with telemetry – or were this to be a design intention or consideration in its design – they could be considered elicitation games. If such data

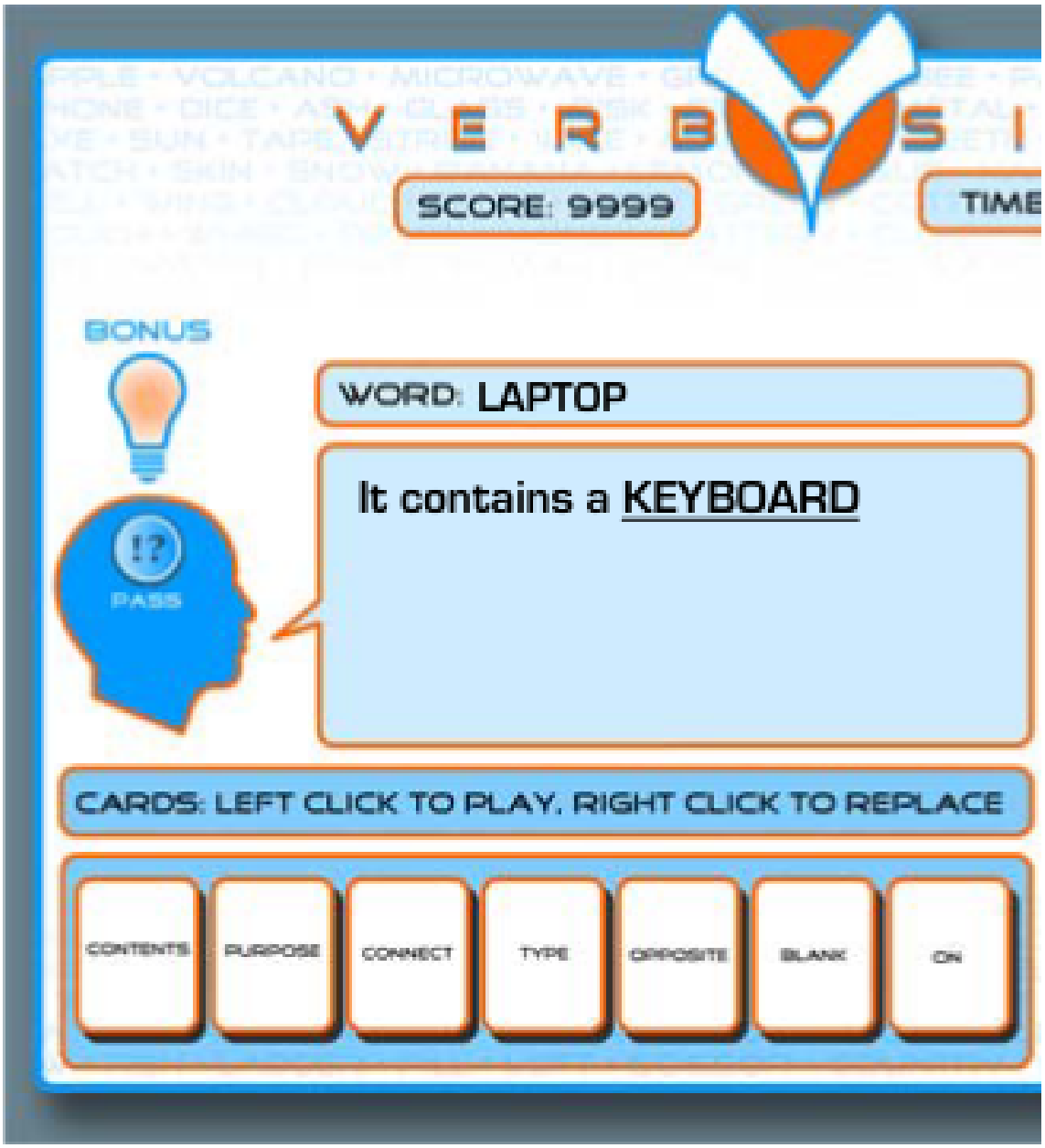


Figure 1.13: *Verbosity*. Image from von Ahn, Kedia, et al. (2006)

exists from these games and were to be applied to human-subject research, it could be considered applied gaming. The popularity of these games suggests that similar games designed as elicitation games could be successful.

1.3.3 Discussion

In surveying elicitation games, I have observed six ways in which games reward data provision. The most common is to reward task performance: directing players to act as fast as possible, or as accurately as possible. Similarly, elicitation games were found where

demonstrating the correct knowledge is rewarded, and one elicitation game was identified whose reward was a lottery payout, leading players to answer as strategically as possible to maximise their reward. I found one game (*BeFaced*) which rewards providing data that conforms to a model (facial expressions). One edge-case was found which rewards agreement with a consensus (*Apetopia*). Finally, I found one edge case (*Bubble Trip*), whose reward is independent of the data provided.

BeFaced and *Apetopia* will be analysed in detail in chapter 4, where it is unpacked why these are not generally applicable templates to the problem of human-subject data. For now it is important to note that these examples should not be taken to mean that solution-based or intersubjective-consensus game models are generally applicable to elicitation games. *Apetopia*'s reward mechanism is heavily obscured to the player, whose mental model of the game is unlikely to include any kind of consensus agreement. *BeFaced* compares player camera inputs against a model, but only in dimensions other than the that of interest for data collection: by rewarding matching a facial expression classifier, the player must unavoidably also give their specific facial landmark data or photograph. *BeFaced* does not compare this facial landmark data or photograph *per se* against a ground truth model.

Thus two main patterns have been observed. Elicitation games either reward some objective measureable outcome of a task (task performance, knowledge matching ground truth, lottery payout), or the dimension of reward is orthogonal to the dimension of interest (as in *Befaced*).

I additionally found a number of games that, with relatively minor changes to the collection of data could have been elicitation games, such as *Brain Age* (Nintendo, 2005b) and *Big Brain Academy* (Nintendo, 2005a). I also contrasted gamified data collection tools that do not satisfy the criteria for elicitation games and data collection games whose data is not about their players as individuals. While doing so, I have applied the criteria for an elicitation game. As given above, elicitation games are:

1. designed for enjoyment as a primary goal; and
2. designed as a measurement instrument for some meaningful data about their players as individuals, which
3. can be recorded and transmitted to researchers

A significant factor in these criteria is the intent of the designer: what the game is designed *for*. This arises from the conceptual situation of elicitation games as applied

games. Applied games as artefacts are the result of applied games *design*, a process which is inherently directed towards an outcome; Schmidt et al. (2015, p. 104) define applied game design as "the user-centric transfer and implementation of design concepts from the game world, in order to confer their individual, social and procedural qualities to a subject of interest, within its situated context, in order to pursue a defined goal." Moreover, as applied games are generally described – as in academic papers and publications – but are only infrequently accessible to actually play, it is generally easier to evaluate the intention of their design than to evaluate the actual game itself. In this situation is is likely to be easier to determine whether a game was intended to be enjoyable than to evaluate in any reliable manner whether it is *in fact* enjoyable.

One thing that has not been included in the above definition is a requirement of validity. This is on purpose. First, while validity may be a criterion of a *successful* elicitation game, it is not a requirement of elicitation game *per se*. Moreover, validity is hard to objectively evaluate, even more so than enjoyment or design intention. However, validity is an important goal of elicitation game design.

1.4 Validity

Whenever data is collected, the question of *validity* arises: To what extent does the data support the inferences we draw from it? Together with reliability, validity is a key construct and quality criterion of scientific research, with a long history in quantitative research (Jenkins, 1946), but also in wide use in contemporary qualitative research (Morse et al., 2008; Nahid Golafshani, 2003). Following Messick's popular conceptualisation, "[v]alidity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Messick, 1995, p. 471).

Beginning with Campbell (1957), researchers have developed multiple (general) typologies of validity and validity threats, usually structured around aspects of claimed causal relations between the treatment and outcome of an experimental study. In a classic typology, Shadish et al. (2002, pp. 38, 42–93) distinguish four kinds of validity:

- *Statistical Conclusion Validity*: Do the presumed cause and effect covary and if so, how strongly? Common threats to this validity types are statistical issues like low statistical power, inappropriate statistical methods (e.g. because the data violates

underlying assumptions like normal distribution), or uncorrected fishing for results.

- *Internal Validity*: Is the observed covariation between treatment and outcome actually due to their causal link or other factors in the research design? This connects to validity threats like confounds or selection bias. For instance, an observed negative correlation between age and gameplay performance may not be due to age, but due to a third factor (= confound) that covaries with age, such as gaming socialisation. Similarly, if gaming socialisation affects gameplay performance and researchers want to study the effect of alcohol consumption on gameplay performance, they need to ensure during sampling that treatment and control group don't have very different gameplay socialisations (= selection bias).

- *Construct Validity*: Do the particular samples, treatments, and outcome measures used in the experiment accurately operationalise the constructs that are studied? Many threats to validity in this category have to do with what researchers call reactivity (Shadish et al., 2002, p. 77). Participants actively make sense of and respond to the study they participate in: they may answer what they think the experimenter wants to hear, cheat to get a high score on an assessment, or are reminded by the study of stereotypes regarding their own aptitude in the activity studied, which may induce self-doubts and anxieties dampening performance.

- *External Validity*: Do observed correlations hold across other people, settings, treatments, and measures? In other words, do the study results generalise? Part of this is the question of so-called ecological validity: to what extent does the experimental situation reflect actual situations as people would experience them in their day-to-day life? For example, that people choose one game over another in a lab experiment where they only have two games to choose from may tell us little about how people actually choose games in their everyday life, where they are faced with hundreds of thousands of games in an app store.

Various validity threats from using games have been suggested: the use of different games as experimental conditions that vary in more than the targeted aspect (Ferguson, 2015); failing to recognise differences in gaming expertise linked to different game genres (Latham et al., 2013); or the high cognitive load of games, making it difficult for players to play a game in a 'natural' way *and* report on their gameplay experience, as standard think-aloud methods in HCI ask for (Hoonhout, 2008). However, these and similar observations

have remained scattered and piecemeal, and chiefly concern entertainment and educational games research, not the use of games for other research questions (see e.g. Louvel (2018) for a survey of ecological validity issues in lab-based games user research and S. P. Smith et al. (2015) for a meta-analysis of data collection forms in serious games). In fact, while authors like Williams (2010) and Deterding (2016b) have called for systematic research programmes on the correlation between people's in-game and out-of-game behaviour, there has been little integrative, let alone systematic work in this area – one exception being a narrative review of issues in gamified surveys by Keusch and Zhang (2017).

What we can be sure of is that the use of applied games for data collection in human-subjects research will require justification in terms of validity. First, for their data to be trusted, such games *qua* measurement instrument must be justified in terms of construct validity. In other words, they must measure what we think they are measuring. For the metamethodological research presented here, this is can also be understood in terms of ecological validity: such games must be ecologically valid *as a methodology for collecting their desired data*. That is, the data they collect – whether about rainfall or butterfly sightings, must be, in principle, meaningful or useful. Further, to be 'whole games', such games must be ecologically valid as an enjoyable/motivating game. Finally *studies* using applied games must be further justified in terms of internal, external, and statistical conclusion validity, as with any experimental study.

1.5 Problems to Address

Overall, there seems to be interest in developing games for experiments, at least in cognitive science, HCI and education. It is notable, however, that most of the examples of elicitation games identified are structured around tasks for which performance-conditional feedback can be provided. This might be the accurate performance of an experimental task, lottery payouts, or expression of accurate knowledge. In each of these cases, players can be motivated to perform the task demanded by the inclusion of surface game features to existing experimental tasks, primarily by including a score and a goal, or game-like graphics and a simple game narrative.

Those elicitation games that place greater emphasis on game design seem to be motivated by a desire for longer, more voluntary, and ongoing engagement. For example, as a participant recruitment strategy (e.g. H. R. Brown et al., 2014; Spiers et al., 2021), maximising data per participant (e.g. Oladimeji et al., 2012) or ongoing engagement for

learning (e.g. Lyytinen et al., 2007). Still, despite successful examples such as *The Great Brain Experiment* from 2014 (recruiting over 20,000 volunteer participants (H. R. Brown et al., 2014)), approaching a decade later, such games are still scarce.

At the start of the this chapter I identified several motivations for using games to elicit human-subject data. However, only a handful of examples of such games were identified. Conceivably, this might be limitations in the survey – for instance, there may be more enjoyable games used in lab experiments that do not describe themselves as such – yet that would not explain the absence of methodological work on the subject. On the assumption that such games are relatively rare, the question that then arises is: why? What are the issues preventing the use of elicitation games more widely, and for more kinds of human-subject data?

As it stands, we don't well understand how to design and evaluate elicitation games for human-subject data. The familiar templates used by Applied Games for data collection to validate the data collected do not apply. While there are taxonomies in human computation (Quinn & Bederson, 2011), and advice for matching game mechanics with data to collect in games with a purpose (Galli, 2014). We don't have comparable models to work from to collect human-subject data. In order to develop an understanding of how to design and evaluate elicitation games, we need to understand validity issues in games and what motivates people to provide valid data.

Validity issues in games For most games developed for experimental research, experimental validity is a concern of highest importance. Due to this, there may be a perception that prioritising participant enjoyment through game design is incompatible with rigorous experimental design, or such an experimental design would not be convincing to others in the field. As games seem to be inherently complex and variant, this may well often be so. However, currently such evaluation is ad-hoc, and not supported by any systematisation or review of the particular threats to validity characteristic of games. The development of design guidance may aid the design of better elicitation games without threat to validity, as well as justifying their use.

Motivation for providing data While research has looked at motivations to play games (Ryan et al., 2006) or participate in research (Crowston & Prestopnik, 2013; Curtis, 2015; Iacovides et al., 2013; N. Prestopnik et al., 2017; Tinati et al., 2017), it has not formulated this question in terms of what motivates participants or players of applied games to provide

valid data. Experimental research has largely taken such motivation as the default and concerned itself with the reverse question: what motivates participants to *not* provide valid data, for example, demand characteristics (Orne & Whitehouse, 2000). What is not often considered is the motivation for the experimental participant, at a particular moment in an experimental study, to choose to give valid data.

Finally, in order to design games to collect (valid) human-subject data, these insights on validity and motivation need to be integrated into a design model.

1.6 Research Questions

Elicitation games, as defined above, sit uncomfortably between the concerns of experimental research and games design. The principle concern of human-subjects research is validity and about this a great deal of knowledge has been amassed. However the field does not address itself to the very real design choices that are essential for creating a successful elicitation game. The applied games field, on the other hand, has considerable resources for game design and using games to structure and motivate tasks, but this has not been integrated with a concern for validity. Indeed, as characterised above, the predominant approach within the use of applied games for data collection is gamification+validation where the two are seen as essentially separate concerns.

When the gameplay *is* the data, as with elicitation games, the twin concerns of game design and validity cannot be separated. Instead, what is needed is an integrated model or understanding that can bridge these two fields: to draw on both the knowledge of validity from experiment research *and* the knowledge of game design from the field of applied games. There are no existing templates or taxonomies within the applied games field that apply to elicitation games, nor methodological guidance within experimental research that engages with the reality of game design. This thesis aims to fill this gap.

The overall research question of this thesis is:

> **How do we design games that motivate their players to produce valid data?**

I will structure my research to answer this question through addressing four sub-questions:

RQ1 What do we know about validity in games? A survey identifying the key ways in which games *qua* games threaten validity is necessary for grounding this research and would contribute to assessing the validity of experiments that make use of games.

RQ2 What motivates players to choose to perform actions that provide data within a game? An account of player motivation to take particular data-providing actions in a game would provide guidance on how to design for particular behaviours, necessary for understanding how to motivate accurate data elicitation. In contrast, most theories of game motivation treat motivation at the level of the game as a whole, not individual actions.

RQ3 How does game design motivate players to provide valid data during gameplay? A model that explains this would support the design and analysis of elicitation games as well as give a common language in which to compare elicitation games. It may further generalise to other uses of games in experiments and games for eliciting types of data other than human-subject data.

RQ4 What are practically relevant threats to validity in elicitation games? An identification of the threats to validity of greatest practical impact to the design of games for experiments would allow game design and analysis to be based on empirical result, rather than speculation alone.

1.7 Outline of the Research

To address these questions, I undertook a programme of research. To best understand the structure of this research, it is helpful to understand something of the historical circumstances from which it developed. As such, I next present a brief historical anecdote which I hope will further motivate this thesis. Here I will also explain the focus on speech and language data in this thesis. Following this, I will present a step-by-step outline of the contents of this thesis.

1.7.1 Historical Account of the Research

Some writers have a particular reader in mind when they write. If the present thesis is addressed to anyone in particular, it is addressed to myself when I was beginning to

research my original PhD topic. My hope is that the tools it provides will prove useful to someone in a similar position.

I began my PhD research with the aim of developing a game to collect spoken linguistic data. I had a background in formal linguistics and I was motivated by the particular benefit that game-based data collection might provide for linguistic research. My goal was to exemplify how applied game design can be applied to transform expensive in-person lab experiments in linguistics into scalable tools. For instance, to observe the character of child language acquisition (for instance, whether learning is gradual and statistical or a series of triggered changes (Yang, 2004)), instead of record the changing speech of children once a month for an hour in a lab, could a game record 5 minutes every day? This would provide a kind of resolution of data previously impossible.

The particular challenge I knew I would need to overcome was the difference between existing game-based methods for data collection in other fields (which, as seen above, are the use of relatively straightforward performance-contingent rewards to motivate data provision) and the relative impossibility of such designs in a linguistic context. Indeed, linguistic data is descriptive. It shows how an individual speaks (signs, writes, etc.), but it does not take any particular production to be 'better' than any other. Nor need any individual's production correspond to standardised forms such as standard English. What is desired is simply a spontaneous expression of that individual's linguistic ability: diversity in this is itself interesting data. To motivate such data, performance-contingent rewards are therefore impossible and a different kind of data collection game is required. Yet, as already introduced at the start of this thesis, suitable templates do not exist.

I attempted to explore specific solutions to this problem in the context of language. However, such directions were obscure and their reasoning based upon assumptions. Testing these assumptions directly in the context of language would be trivial to the linguist, yet obscure to the applied games researcher. Thus, I determined that to progress I needed to generalise to the wider class of human-subject data which shared the fundamental issues of interest. This I hoped would allow me to develop the tools needed to more successfully return to the issue of developing a speech elicitation game.

This history is responsible for the use of speech and language data as a case study, which runs through the whole thesis. In particular, I began the research behind chapter 3 when I was still aiming to design games specifically for eliciting speech data, as might be deduced from its narrow focuses on motivations for speech in games. Ultimately, I believe the

narrowed focus on a single domain made my results in that chapter more concrete, though it also made the route to a generalised model in chapter 4 more circuitous. Helpfully, speech and language is an appropriate case study of human-subject data. One of the benefits of adopting it is that I have not been tempted to restrict the scope of this research to only the 'easy cases', kinds of data where data where performance conditional rewards are possible. If this thesis has taken a novel approach to the problem of applied games, it is in large part the use of linguistic data as a case study that is responsible.

1.7.2 Thesis Outline

This thesis is presented in 8 chapters. This first chapter has motivated the research and situated it within the wider fields of applied games and human-subject research with games. It has presented the four research questions that guide the following work.

Chapter 2 The next chapter presents a narrative literature review on validity threats characteristic of the use of games in experiments. Much of the work in this area is speculative, there being little empirical research quantifying the degree of practical importance of these, particularly for games. This narrative survey similarly proposes potential threats to validity where there seems to be a compelling reason for them. I will later draw on this chapter when developing the Intrinsic Elicitation model later in the thesis.

Chapter 3 Next I address motivation for data provision within an applied game. There is little literature that addresses motivation for taking particular actions, or motivation on a moment-by-moment basis. Therefore, I will adopt a qualitative methodology to systematically generate theory grounded in the design of games. I will do this for the case study of speech data. I build a grounded theoretical understanding of what is that motivates participants within those games to provide the particular instances of speech they do. This leads us to design principles that can help the design of games to elicit speech. For this thesis, it provides a model that informed the development of my design approach for elicitation games.

Chapter 4 Having grounded my discussion in an understanding of validity and motivation, I go on to develop an integrated model of validity and motivation in elicitation games, called *Intrinsic Elicitation*. This generalises the understanding of motivation developed in the preceding chapter to apply to more types of data. The chapter exemplifies the model as

an analytical tool and explores the breadth of applicability of this model beyond elicitation games with a sample analysis of the games *Apetopia* (Barthel, 2013), *BeFaced* (Tan et al., 2014), and *Urbanopoly* (Celino et al., 2012).

Chapter 5 The theory I develop relies on incorporating relevant threats to validity as factors within its Rational Game User model. Virtual utility and mechanic actuation effort are two factors suggested to be of prime importance. This chapter presents a pair of controlled experiments to determine whether controlling for these factors is sufficient to maintain accuracy in an elicitation game in comparison to a practice-as-usual experiment. To do this, I introduce a novel game for eliciting adjective order data, *Adjective Game*, which due to the reliability of language between speakers allows us to compare accuracy rates for a human-subject data. I test self-report enjoyment and accuracy, finding a trade-off between these.

Chapter 6 While the development of *Adjective Game* adhered to the design principles of the model, there was still a decrease in the accuracy of data elicited between game and non-game conditions, suggesting further factors may need to be integrated. One such factor that emerges as a likely explanation for the differences in accuracy observed between 'game' and 'experiment'-appearing conditions are social norms, understood in experimental research as demand effects. This might suggest the task of taking an experiment out of the lab and into a game will inevitably decrease accuracy. Motivated by the results of the previous experiments, I explore whether demand effects in the form of explicit instructions or the metacommunicative framing of the experimental situation are responsible for an increase in accuracy. To do this, I perform two further controlled experiments using the same game. The results show that explicit instructions do have an impact on accuracy but I am unable to detect any effects of metacommunicative framing.

Chapter 7 From here I draw together the work of the thesis in a discussion. First I review the research questions above and what has been learned towards addressing them. Second, I evaluate the Intrinsic Elicitation model and identify areas where further testing is needed. In particular, I address the fact that while suited to analysis, it is less applicable to design. I give directions for future development of the model and future research in this area.

Chapter 8 The final chapter summarises the main contributions and limitations of the thesis and sketches the main directions for future work.

1.8 Ethics and Open Science

Scholars have raised ethical questions with the growing use of data collection through games, applied games, and gamification (e.g. Deterding et al., 2015). When games are used for data collection, it is far from clear whether players understand the scale of the data that is, or can be, collected about them, particularly given advances in the field of machine learning. Indeed, the research presented in this thesis assumes that it is possible to measure something about the player's latent attributes, such as preferences, beliefs, or competencies based on interactions with a game. No one can predict how some piece of information that seems innocuous could be used in the future. I believe the appropriate response to this is to rigorously minimise and anonymise the data that such applied games collect. I have attempted to do this in my own work.

1.8.1 Participants

The experiments reported here were designed to avoid causing harm or discomfort. The game participants were asked to play was an abstract game about shapes and language. It did not contain violence or contentious social themes. Participants were not exposed to any risks beyond ordinary play of a (potentially buggy, perhaps boring) casual game.

Participants were from the online panel provider Prolific (Palan & Schitter, 2018). Prolific provides anonymity for participants from researchers. Furthermore, each experiment collected the minimum of data about participants required. Details of anonymisation are given below. Participants were presented with study information about the activities they would be performing and the data that would be collected about them. As the study was run online, participants came to their own decision about whether to participate. Participants were paid for their time based on the expected completion time of the study. Participants were paid whether or not the data was recorded due to the possibility of game error. Participants who withdrew their submissions and contacted me regarding a bug preventing them from completing the experiment were paid in full.

1.8.2 Open Science

The data and analysis scripts for all of the experiments reported here are available through the Open Science Framework[9]. Preregistration was attempted for all of the experiments reported here. The first two studies did not have a published preregistered due to my own misunderstanding and a technical error, but for both a preregistration document has been archived on the Open Science Framework timestamped before data collection began[10]. For the remaining two studies, preregistrations were published before data collection began[11].

The published data for this thesis is provided in the format in which it was analysed, which is an anonymised form. Anonymity is a hard problem. Advances in machine learning mean that you can infer a lot from supposedly anonymous data sets. Furthermore, other organisations beyond my control may have recorded and stored datasets that, if exposed could be combined with my published data to identify an individual's data. As such, I erred on the side of caution in anonymising the data. Any part of the data that could conceivably be matching with another dataset (e.g. through timestamps) was disassociated from the rest of the data. Similarly, as few participants reported non-binary genders, and/or would be of a particular age, these values were disassociated from the rest of the dataset. The anonymisation process was performed (almost) entirely automatically. If one wishes to check that no substantive changes were made to the analysed data, the anonymisation script used is available for scrutiny on the OSF repository.

[9] https://osf.io/jac6s/, https://osf.io/w6xqj/
[10] Study 1: https://osf.io/hab82/; Study 2: https://osf.io/sg3uk/
[11] Study 3: https://osf.io/ek64h/; Study 4: https://osf.io/yq6ua/

Chapter 2

Threats to Validity

In order to understand elicitation games – whose substantive purpose is to collect human-subject data that is of scientific value – we need to understand how games characteristically can affect validity. What is it about games that typically mean that data collected using them might not be valid? Are these things essential, and if not, can we design to avoid them?

This chapter surveys known validity threats in the use of games for data collection, as well as highlighting potential, as-of-yet unstudied threats where there is a compelling argument for them. In interest of space, I constrain the discussion to well-established validity issues of quantitative research that are particularly pertinent to games. I organise the discussion along three key features of games that appear to underlie most validity issues I identified, namely *systemic complexity*, *variance*, *framing*, and *player-related factors*.

Complexity The complex systemic nature of games and gameplay make it hard to isolate and manipulate individual game elements without inadvertently affecting other properties as well, leading to confounds threatening *internal validity*.

Variance Games and gameplay are high in diversity and uncontrolled variance. This means that given the same effect size, larger samples are needed to make valid inferences about true effects, and increases the likelihood of detecting false positive effects (type 1 errors) as well as failing to detect true effects (type 2 error), impacting *statistical conclusion validity*. Furthermore, high variance threatens *construct validity*, as it becomes hard to hold everything but the operationalisation of the construct in question constant.

Framing Gameplay is a very particular kind of social situation or frame that differs both from experimental setups and other types of social situations we wish to make inferences about. This threatens the generalisability of findings, and with it, *external validity.*

Player-related Factors Despite the mainstreaming of video gaming, player demographics of particular games still can deviate from the general population, likewise impeding *external validity* of data collected via games. Interpersonal differences in play styles and genre expertise may confound results, while turning an activity into a game may lead players to deviate in strategic ways from how they would spontaneously behave in a non-game version of the same activity (i.e. cheat or 'game the system'), with negative ramifications for *internal, external, and construct validity.*

I arrived at this systematisation through a combined bottom-up and top-down approach: Bottom-up, I conducted an opportunistic literature search across major relevant databases (Web of Science, Scopus, ACM Digital Library, Google Scholar), searching for 'validity' and 'games', resulting in 10 relevant papers (see References). Top-down, I used the typology by (Shadish et al., 2002) to systematically ask for each validity threat they classify where and how it may manifest around digital games. Clustering reported and hypothesised validity threats, I arrived at a smaller subset of threats that I then tried to variously organise, arriving at three high-level characteristics of games that appeared responsible for them. In response to reviewer suggestions[1], I factored out player-related threats into a separate fourth category. Table 2.1 provides a schematic overview of game characteristics and validity implications.

For each of the four groups, I will first introduce underlying characteristics and then report observed and potential validity threats. In the discussion and conclusion, I will draw some overarching observations on the opportunities and challenges of using games and game design in data collection and outline areas for future research.

2.1 Systemic Complexity

Many even simple games form *complex systems* (Salen & Zimmerman, 2004), meaning that they are made of a network of many constituent parts which interrelate, mutually depend,

[1]This chapter previously appeared as David Gundry and Sebastian Deterding. 2018. Validity Threats in Quantitative Data Collection with Games: A Narrative Survey. Simulation & Gaming

Table 2.1: Overview of game characteristics. Shaded cells highlight relevant validity implications

Characteristic	**Validity Implication**	**Validity Threat**			
		Statistical	Internal	Construct	External
Systemic Complexity					
Games are rich, complex stimuli	Cognitive load and induced arousal can interact with measurement or confound		■		
Games and gameplay are complex systems	Manipulation may have unexpected emergent effects		■		
Games are novelty-based and learned	Learning effects over repeat measures		■		
Variance					
Games are divergent	Different games as treatment/control can differ on more than desired dimension		■	■	
	Data from one game may not replicate in others				■
Game setups are divergent	Measurement errors and confounds when different participants use different setups		■	■	
Game content is varied	Uncontrolled, non-random variance in stimuli and conditions	■	■	■	
Gameplay is emergent and varied					
Commercial games are not fixed					
Framing					
Play frame may differ from target situation	Gameplay behaviour may not generalise				■
	Gameplay may motivate dishonest responding		■		■
Research studies may differ from play	Forced gameplay may turn game preferences into a confound		■		
	Knowing alternative game conditions may produce resentful demoralisation or compensatory rivalry			■	
Players					
Gamers are not the general population	Games may attract a biased sample				■
	Games may differentially produce stereotype threat, evaluation apprehension, and social desirability in (non)gamers		■	■	
Gamers are diverse	Chosen game genre and platform may attract a biased sample				■
	Difference in player types or genre expertise may confound results		■		

and interact in ways that are hard to analyse, model, manipulate, or predict in isolation (Auyang, 1999). This manifests particularly in *gameplay* and *player experience*, the process in which players interact with a game and the experiences they have of this process (Hunicke et al., 2004). Both strongly *emerge* from the interaction of game system and players in ways that are *nonlinear* and *time-dependent*. Changing even a small parameter of a single game mechanic, such as drawing not one but two cards in a card game, can make one in-game action more powerful than another. This in turn can change what winning strategies are, how long the game takes to play, how satisfying its challenges are, or what kinds of information need to be communicated between the players. Notably, the emergent properties of particular design changes or elements do not necessarily hold across games. For instance, adding a time constraint to chess matches turn chess into "speed chess", a game with such recognised differences in gameplay and player experience that it warrants its own name. In contrast, adding a match time limit to the real-time game *Quake* (id Software, 1996) changes gameplay and player experience only minimally. Game development is therefore highly iterative, continually building and playtesting design changes to assess and tune their actual emergent effects (Hunicke et al., 2004).

What this means for researchers is that isolating and manipulating constituent parts – let alone psychologically "active ingredients" (Michie & Johnston, 2013) – of a game is inherently challenging. But without isolating features we risk confounds in our experimental manipulations – third variables that provide a competing explanation for our results. Confounds threaten internal validity, meaning that we may not be able to justify that the treatment has observed an effect.

Within this section, I identify and address three characteristics of games that are involved in their complex systemic constitution:

1. Games are rich, complex stimuli
2. Games and gameplay are complex systems
3. Games are novelty-based and learned

2.1.1 Games are Rich, Complex Stimuli

Stimulus complexity *per se* is not unique to games: traditional experimental materials may be more or less complex, ranging from the relative sparseness and uniformity of e.g. standardised paper questionnaires to the richness of social psychological experiments like

the famous bystander study in which confederates role-played a seizure during a discussion to assess whether the presence of other bystanders affected participants' responses (Darley & Latané, 1968). Compared with other typical media psychological materials like text, imagery, music, or film, games sit on the high end of stimulus complexity and richness, combining the above into a 'total art work'. Apart from the resulting stimulus variance across and within games (see below, *variance*), this introduces significant potential confounds of its own, namely arousal and cognitive load. Gameplay often induces arousal (C. A. Anderson & Bushman, 2001), which is a well-known potential confound particularly in self-report studies, leading to e.g. selective attention (Pham, 1996). Gameplay is also often immersive and engaging, which increases cognitive load (Schrader & Bastiaens, 2012). Extraneous cognitive load in turn is known in games-based learning to impede learning (Kiili, 2005, p.21). The cognitive load of educational games can interfere with their pedagogical effectiveness, and confound studies that compare game and non-game conditions without controlling for cognitive load (Wouters et al., 2013). Similarly, the high combined cognitive load of gameplay and certain data collection methods like think-aloud (Hoonhout, 2008) may overload participants, such that measurement interacts with and potentially confounds gameplay and player experience.

2.1.2 Games and Gameplay are Complex Systems

The systemic interrelation and interaction of elements of games makes it is hard to change one aspect of a game in isolation Kim and Shute (2015). In psychological parlance (Littman & Rosen, 1950a), changes on the *molecular* level of individual game elements or player actions tend to play out as wholesale changes on the *molar* level of the whole game or gameplay – often in emergent, nonlinear ways. Conversely, the same molar *gestalt* – a particular player experience, gaming strategy, or gameplay dynamic – may be realised through many divergent molecular game features and player actions. Communication research has developed some theoretical paradigms that acknowledge such systemic dynamics (Früh & Schönbach, 2005), and some theoretical models do describe games, gameplay and player experience on different levels of organisation (e.g. Klimmt, 2006). Still, much design research and guidance in applied gaming for data collection and other purposes focuses on individual, molecular game features or elements (Deterding, 2015), and the complex systemic constitution of games runs counter to the *de facto* linear, reductionist assumptions embodied in standard experimental research designs.

For game-based data collection, I see two particular ramifications. First, any manipulation (e.g. imported from a standard non-game study design) may generate unforeseen interactions and emergent dynamics in addition to the causal effects it is hypothesised to produce. Therefore, researchers should prototype and pretest manipulations to check for these confounds. This means that researchers need a clear idea of what confounds to look out for. For example, Carnagey and Anderson (2005) used two modes of the same racing game to manipulate whether violence is rewarded in the game. In one condition, players were rewarded for killing pedestrians and race opponents. In the other, they were prevented from doing so. While most differences between the conditions were controlled by using the same game, the difference in used game mode still arguably had an unintended knock-on effect on competitiveness (Adachi & Willoughby, 2011a). Concretely, there were different numbers of things to compete on in each condition. The condition where players were rewarded for destroying race opponents had two sources of competition: winning the race and surviving the free-for-all. This second source of competition was not present in the control condition where destroying race opponents was prevented. To me, this suggests that researchers should incorporate game design expertise in pretests, as these kinds of unexpected dynamics are likely more readily apparent to experienced game designers.

Relatedly, the development process itself has a potential significant impact on the conditions designed. In games, it is typical (and practical) to develop a first part of the game ('first playable', 'vertical slice') in high detail to establish the game's core mechanics and gameplay. All subsequently developed parts or levels are effectively variations and extensions on this core gameplay. Hence, designers are constrained in the development of subsequent content (or conditions) in a way that they are not in the first, and there will often be a general difference in quality, be that in terms of fun or balance, between initial and subsequent conditions developed in this way. As McMahan et al. (2011, p.4) assert for experimental designs in general: "in our experience, researchers may spend more time or effort implementing a condition that they subconsciously (or consciously) favor, biasing the study toward that condition."

2.1.3 Games are Novelty-based and Learned

Novelty is the property of an experience being new (Silvia, 2006). Simply put, the first experience of a stimulus or measurement instrument is different from subsequent encounters. In situations where novelty features strongly, the validity threat of *learning effects*

arises: participants perform differently when they had past experience with a stimulus or measurement instrument (Shadish et al., 2002).

This relates to two time-related characteristics of games. First, interest and curiosity are two major intrinsic motives and sources of enjoyment in gameplay. Games feature properties like dramatic conflicts, novel content, puzzles, or randomness to afford uncertainty that draws attention and is satisfying to resolve (Costikyan, 2013). If a part of a game entails largely the same outcomes and experience on replay, less uncertainty, curiosity, and interest are likely to arise. This is of particular relevance to games focusing on narrative: after a first play-through, the novelty and dramatic tension of their plot is largely exhausted (Roth et al., 2012). As a result, if players play the same section of a game twice – once as a treatment and once as a control – the diminishing uncertainty, curiosity, and interest may become major learning effects.

A second time-related game characteristic is that players over time learn to play them well. In fact, a large part of their enjoyment arises from the competence experience of learning to master the game (Deterding, 2015; Koster, 2005). Most games are therefore designed with a careful scaffolding of required and taught skills and knowledge, increasing difficulty in lock-step with growing player skill (Chen, 2007). If a player returns to replay earlier game sections they've already mastered, they are likely to find it easy to overcome its challenges, and will thus likely experience mastery or learning – again, a strong possible learning effect. Another potential learning-related confound is the difference between learning a new skill and performing a mastered skill, which can express itself in e.g. error rates, time taken, exploratory versus goal-oriented behaviour, and the like. Players' game knowledge may even threaten construct validity. For example, if multiple play-throughs allow players to memorise puzzle solutions rather than solving puzzles anew, game performance may be indicative of short term memory rather than problem-solving abilities. The amount of time participants have to learn a game before engaging with the game section that constitutes the experimental manipulation/control may also confound results. Too little time and players lack of skill may prevent them from effectively completing the experiment. Too much time may lead to ceiling effects or converging upon optimal strategies. Finally, which section of the overall sequence of a game players play may significantly impact what level of difficulty and required skills they encounter.

These strong potential learning effects and other confounds are of particular concern in within-subject, repeated measure designs where the same subject is presented with control

and experimental condition in sequence. Apart from choosing different study designs, one common mitigation is to vary the sequence of control and manipulation conditions as part of randomisation. A second mitigation strategy would be to use techniques like adaptive procedural content generation and pre-testing to ensure that game content in each condition is equally novel and difficult.

2.2 Variance

Next to complex systems, another popular way of framing games is as a possibility space (K. D. Squire, 2008). Games open a space or tree of possible states and partly relinquish control over in-game events to extraneous influences like player choice or randomness. Control, however, is the *sine qua non* of experimental design. It ensures that all participants experience the same manipulated or control condition as reported – essential for construct validity. It also minimises unwanted variance between conditions. Where such unwanted variance is randomly distributed, it statistically obscures true effects, requiring the use of larger samples. Where it is non-random, it becomes a confounding variable. Research games are thus faced with somewhat conflicting requirements to be both a controlled research tool and a game offering possibility spaces. This dilemma manifests itself around at least five different forms of unwanted variance:

1. Games are divergent
2. Game setups are divergent
3. Game content is varied
4. Gameplay is varied and emergent
5. Commercial games are not fixed

2.2.1 Games are Divergent

Games are a highly diverse medium, with different interfaces and controls (e.g. desktop monitor plus mouse and keyboard versus mobile touch screen versus motion control plus VR headset), different social contexts and configurations (e.g. public competitive Esports play versus private cooperative or competitive multiplayer gaming versus solitary play), and different genres (e.g. open world exploration, casual puzzler, idle game, RPG, first person

shooter), each affording different demands and experiences. No single game can therefore be taken as representative of all games. The selection of a particular game, including its interface, controls, social context and configuration potentially threatens internal validity if different games (or game configurations) are used for different experimental conditions, and impinges on external validity in terms of how well or widely any findings generalise. Indeed, the use of different games to operationalise different experimental conditions has been cited as a critical internal validity issue in the violence in video games literature (Adachi & Willoughby, 2011b; Ferguson, 2015). Because the games used differ on more dimensions than just violent content, these dimensions present potentially confounding variables (Elson et al., 2013).

In response, some scholars have suggested ways to match games on certain criteria (Adachi & Willoughby, 2011b), such that key features considered relevant to the investigation vary only in desired respects. For example, research on violence in video games has adopted this approach to match violent versus non-violent treatment and control games on competitiveness (Adachi & Willoughby, 2011a), difficulty of controls (Przybylski et al., 2010), and frustration (Przybylski et al., 2014). Two difficulties arise with this matching strategy. Firstly, games may not successfully be matched on the given factor. Secondly, games may remain divergent on unmatched factors that reveal themselves to pose confounds. For instance, C. A. Anderson and Dill (2000) selected the two games *Wolfenstein 3D* (id Software, 1992) and *Myst* (Cyan, 1993) as violent treatment and nonviolent control because they matched for "blood pressure, heart rate, frustration, difficulty, action pace and enjoyment" (Adachi & Willoughby, 2011b). Yet in this, C. A. Anderson and Dill missed that the games also differed in competitiveness, which proved to be a significant confounding variable (Adachi & Willoughby, 2011b).

Another approach is to adapt a single game to provide treatment and control conditions, a so-called modified game paradigm (Hilgard et al., 2017). This adaptation can often be achieved through modding an existing game (e.g. Elson and Quandt, 2016; Engelhardt et al., 2015; Mohseni et al., 2015), or developing bespoke games (e.g. Zendle et al., 2015). This allows researchers to control the experimental conditions far better, and avoids the potential confounds of different games. The challenge, as explained earlier, remains that a small manipulation within a game can have unforeseen and undesired emergent systemic effects on gameplay and player experience.

2.2.2 Game Setups are Divergent

Related to the diversity of games, the way digital games are technically delivered and instrumented can vary greatly. This leads to potential measurement errors or confounds, as the means for collecting, transmitting, and recording data are subject to error, interference, or unwanted variance. This issue is particularly acute in remote/online designs, where researchers have less control over gaming hardware, controls, and networks. Variance in participants' computers, controls, or networking bandwidth can cause issues with tasks and measures such as those involving reaction times (Hilbig, 2016; Reimers & Stewart, 2007).

2.2.3 Game Content is Varied

Even within a single chosen game, the interactivity of games means that the actual content (levels, puzzles, rewards, challenges) players experience will vary between players and game sessions, often significantly and to a not fully predictable extent. This threatens construct validity, as it may be hard to ensure that or discern whether players experienced the desired stimulus, and to ensure that or discern whether they experienced other, undesired variance in stimuli. If game performance is also used as a measurement instrument, such as in educational assessment, chance variation may overwhelm the meaningful information it contains. On the structural level of a game's design, there are at least three sources of emergent variance.

Games Provide Player Choice

Games relinquish significant control to the player (Klimmt et al., 2007), supporting choice in what character they embody, what goals they pursue, what strategies they use and what actions they take. Not only is such meaningful choice directly fuelling engaging and enjoyable autonomy experiences (Deterding, 2016b): it can regularly lead different players to perform different actions and experience different outcomes. Within an experiment, in contrast, tasks performed are usually strictly controlled, and undesired variance in outcome minimised. While almost all games offer some degree of player choice, the amount of choice offered differs markedly between game genres and games. Where some 'rail-road' players along the same trajectory, other open 'sand-boxes' with a wide range of possible goals and actions.

Games Often Include Random Events

Particularly so-called games of chance relinquish control over key game events to randomness. In other games, randomness is incorporated in the design through procedurally generated content to afford replayability: a random seed is used to e.g. generate a different game map every time. Where such events are *truly* random, they merely require a large enough sample to detect true and reject false covariance. However, they often are *pseudo*-random and thus potential confounds, e.g. using pseudo-random seeds or biasing randomness in desired directions. The dropping of in-game relevant rewards or 'loot' in role-playing games for instance is known to have carefully crafted chances to optimise player engagement and not negatively affect the in-game economy.

The Starting Situation Can Vary

The starting situation of a game can be fixed or variable. For instance, many games allow players to customise their characters before start, configure controls, or set the game difficulty. These configuration options are one of the easier variables to control. For instance, a save game or starting setup can be prepared and loaded to ensure consistency across participants. The important thing is to account for this potential variance, especially in remote designs. However, some game details change separately and automatically, for instance the game-wide unlocking of new items, levels, or achievements, which affects the possibility space of all subsequent players.

2.2.4 Gameplay is Emergent and Varied

By relinquishing control over the game state to player agency, games open the systematic possibility that players act differently with every game session. Game theoretically, one can model this possibility space of actions as a decision tree (Elias et al., 2012). If players were fully informed and rational actors in rational choice terms, solely motivated to win the game, one could calculate and predict the strategically optimal move, and such game theoretic calculations are indeed a common tool among game designers (J. H. Smith, 2006). However, especially in games with two or more interdependent actors (human or artificial), possible choices and game states quickly compound to a point where calculating the optimal move becomes humanly impossible: three pairs of turns into chess, there are 121 million possible game states, for instance. Nevertheless, game sessions display higher-level dynamics and player communities and expert players evolve higher-level strategies

and heuristics to reduce this complexity, which again interact and change in hard to fully predict ways (Elias et al., 2012). In active player communities around games like *League of Legends* (Riot Games, 2009), for instance, shared views about optimal high-level strategies like character choice ('the meta') are in constant flux. Complicating the picture further, player actions are regularly shaped by more concerns than mere winning, as shall be seen in later chapters. Overall, this means that especially in interdependent multiplayer games and other games with so-called emergent gameplay, gameplay actions and experiences are hard to control and predict on a low level and showcase emergent but again not fully predictable nor controllable patterns on a higher level of organisation.

2.2.5 Commercial Games are Not Fixed

Current trends in game development mean that even an individual game's variance in content is not necessarily stable, but may change between sessions and over time.

Game Updates and A/B Tests

The widespread availability of high-speed internet connections has enabled a new development and business model usually called 'games as a service', where games are increasingly provided as a continuing online service. As a result, there is often no single, canonical version of a game that holds constant across studies. Rather, continuous game changes have become commonplace, including patches and bug fixes, downloadable content (DLC), seasonal in-game events, and more. Even worse, developers now make frequent use of A/B or even multivariate testing, serving different players different versions of the same game in parallel to assess which works better. Researchers working with existing entertainment games, especially online ones, therefore run the risk of unintentionally or even unknowingly serving participants different game versions at different points, a so-called history effect threatening internal validity.

This risk can be easily avoided in games purpose-made for research. When using existing entertainment games, researchers should decide on a canonical version of the game for the purposes of the study wherever possible, document the version of the game used, and ideally make a copy of it available for future researchers interested in replication. In some cases, a canonical version may be safeguarded by downloading a local copy, disconnecting the game from the internet, or disabling updates. Where a stable version cannot be ensured and a game is updated during an experiment, the researcher should consider and report

any potential impact this may have had.

Adaptation and Content Generation

Many games tailor the experience to the individual player. Single-player games commonly use techniques like dynamic difficulty adjustment (Hunicke, 2005) to give players a satisfying experience by adjusting difficulty based on their past gameplay performance, or even procedurally generate whole levels to keep content novel for players (Shaker et al., 2016). While these systems aim to provide players with an overall evenly enjoyable experience, they reduce control and predictability of actual moment-to-moment game content for researchers. Whether or not to choose a game with dynamic difficulty adjustment or procedural content generation thus becomes an important research design consideration: if an even higher-level player experience is desired, such games may be the best option (but need pretesting). If low-level control is needed, they are to be avoided.

Multiplayer Games

Other players in multiplayer games are sources of variation. For example, one participant may face an easy opponent, while another faces an experienced opponent. While competitive multiplayer games use ranking and matchmaking systems to provide players with an overall even, 'fair' experience (Wardaszko et al., 2019), individual match experiences still differ in many respects. Similarly, online player communities may differ between servers and change over time in their size, activity, demographics, norms, and practices (Bartle, 2004). Thus two players of the same multiplayer game may have different experiences based on when they played and who they played with.

This variance of multiplayer games can be somewhat controlled by using confederates or bots. However, this may in turn produce history effects (when confederates tire out over multiple plays) or threaten ecological validity, as scripted human or bot play may differ from 'spontaneous' play. Where multiple study participants play together, it may be appropriate to use a group randomised trial design to adjust for intraclass correlation (Murray, 1998).

2.3 Framing

People's everyday life is organised into different kinds or types of social situations that each come with particular roles, norms, and expectations: going to the movies, shopping at a store, giving a lecture, etc. During socialisation, people learn what kinds of situations exist in their society, and how to understand and act appropriately within them. Sociologist Erving Goffman (1986) first extensively studied these kinds of situations, calling them *frames*. Whenever an activity is transplanted from its naturalistic situation into a different frame, like an experiment or game, this different framing may have a significant effect on people's experience and behaviour. In psychology, the class of validity threats called 'demand characteristics' refers to exactly this "totality of cues and mutual role expectations that inhere in a social context, (e.g., a psychological experiment or therapy situation), which serve to influence the behavior and/or self-reported experiences of the research participant or patient" (Orne & Whitehouse, 2000). For instance, by taking on the role of a good participant in the frame of an experiment, study participants may act in ways they hope help the researcher accomplish their study goal, rather than how they would spontaneously act in a different situation. Conversely, games research has highlighted that many kinds of behaviours that would be inappropriate in everyday interaction become acceptable or even desired in a gaming frame, such as aggressively competitive and strategic behaviour, bluffing, or teasing and taunting (Deterding, 2014). These frame differences may threaten *construct validity* when demand characteristics become unwittingly part of the treatment. They can threaten *external validity* in that behaviours and experiences occurring during play may not hold outside of the play frame.

2.3.1 Play May Differ from the Target Situation

One oft-mentioned feature of digital games is that they allow to *simulate* parts of the real world with great verisimilitude, especially when employing contemporary immersive technologies like virtual reality. Yet no matter how 'realistic' the simulation, *playing at* an activity is always socio-materially different from performing the activity without a play frame: social and material consequences are usually muted (the virtual lion doesn't really bite, the as-if-breakup is not a real breakup); and norms and expectations for behaviour framed as play differ from those for behaviour framed as earnest (Deterding, 2014). Thus, simply *calling* an activity 'a game' can already change the experience and observed behaviour (Lieberoth, 2015). This raises the question of mapping: for what kinds of behaviours and

contexts does in-game behaviour correlate with behaviour in the real world? (Deterding, 2016b) for instance, while aggregate economic behaviour in online game auction houses mirrors economic behaviour in real-world auction houses, communication norms in online game chat markedly differs from those of everyday face-to-face conversations (Deterding, 2016b). While scholars like Williams (2010) raised this as a basic meta-methodological question of games-based research, over a decade on, there is still little if any work on this issue. Instead, I here want to highlight only two obvious differences as starting points for future work.

Games Mute Socio-Material Consequences

As mentioned, in-game events usually have lowered practical and symbolic consequences compared to their non-play-framed counterparts. While gambling and the rise of real-money trading and microtransactions around in-game items provide plenty counterexamples, in many games, there is no bodily or economic risk involved in game outcomes. Players may therefore be more risk-taking in games than they would be in the real world. There is rich related debate in economics on how much participants need to be paid and what 'hard' payout consequences there need to be for participant decisions during a study for the study to count as ecologically valid (Camerer et al., 1999). Several studies on economic games find differences in player choice when monetary incentives are added (Schlenker & Bonoma, 1978).

Games Invite Strategic Action

Games are one of the few social contexts in which 'ruthlessly' rational, strategic, self-interested action is allowed and even desired: a player who doesn't try hard to calculate and take optimal moves in order to win would be considered a spoilsport (Deterding, 2014). This norm of gameworthiness is counterbalanced with norms of playworthiness – having fun together –, which may result in 'suboptimal' behaviour like self-handicapping. However, different game contexts and genres come with different norms how 'ruthlessly' one is allowed and expected to play (Deterding, 2014), and these norms may differ from the situational norms of the activity of context that one wishes to collect data on. For instance, if a game is designed to elicit people's preferences about different flavours of ice cream, and there is a strategic in-game advantage to answer 'chocolate' even if one *actually* prefers strawberry flavour, the game's design will confound the responses (see chapter 4 for

a detailed discussion and design guidelines to mitigate these effects). The focus on winning the game may also override participants' desire to be a good study subject and lead them to cheat. In games, cheating is the use of mechanisms unintended or forbidden in the game to achieve in-game advantage (Consalvo, 2007). In experimental terms, cheating occurs when participants use means that were not intended to complete an experimental task. For instance, in an online game eliciting players' hand-eye coordination speed, very motivated players may change their screen contrast or use different controllers like an auto-fire mouse to score higher than they would under 'standard' conditions.

2.3.2 Research Studies May Differ from Play

Like play, research studies also constitute a social frame with norms and expectations of their own – what psychology calls demand characteristics (Orne & Whitehouse, 2000). These pose game-characteristic validity threats where they interact or clash with the norms and expectations of gameplay now being re-framed as a research study.

Experiments May Force Gameplay Against Player Preference

First, in leisurely gameplay, players expect autonomy over what game they play when and how long (Deterding, 2016b). However, in research studies, participants are generally assigned to predetermined gameplay conditions. This may lead to frustration as participants may be made to play games they would not usually choose to play and may not like (Ferguson et al., 2017). Certain games may be more widely acceptable than others. For example, players of first-person shooters may be happy to play a casual puzzle game, whereas the reverse may not be true. More generally, genre, controls, or required energy and time to learn may all present differential barriers to engagement with different games (E. Brown & Cairns, 2004). Thus, they may all interact with player dispositions (genre preference, controller familiarity, etc.) that may covary with other player features (age, gender) to produce patterned differences in play outcomes, engagement, and the like that may confound the treatment-outcome correlation under study.

Game Conditions May Differ

If study participants learn about the differences in treatment and control groups, they may adjust their behaviour accordingly, a validity threat that is usually discussed under the labels of compensatory rivalry and resentful demoralization (Shadish et al., 2002).

Compensatory rivalry is the phenomenon wherein a control group puts in extra effort in order to compete against a group receiving an intervention. This may be exacerbated in game-based research given the competition-embracing social norms of gameplay (Deterding, 2014).

Resentful demoralisation is the opposite effect, where one group is demoralised by being put in the inferior condition. Games are generally expected to be fun, but often two experimental conditions cannot be equally fun. Participants who view their game condition as inferior may be subject to resentful demoralisation, e.g. if their condition is excessively difficult or particularly easy. Similarly, if participants are recruited on the basis of playing a game, they may be demoralised to find the game is different to what they anticipated, or they are in a non-game control condition.

2.4 Player Factors

In the preceding sections, I discussed principled issues that arise from the constitution of games, no matter the *particular* participants that engage with them. In this section, I summarise player-specific threats to validity, that is, issues that arise due to the constitution of video game players as participants, and issues that arise from interpersonal differences between players:

1. Gamers are not the general population
2. Gamers are diverse

2.4.1 Gamers are not the General Population

To draw inference from a sample to a wider population, the sample must be representative of that population else the external validity of the study is threatened. This is a particular concern for games-based research in that gaming is a voluntary pursuit: individuals self-select to play games, which may lead to sampling bias. Problems arise when characteristics of being 'a gamer' moderate study outcomes.

For some, gaming is part of their identity, while others regularly play games without self-identifying as 'gamers'. Yet no matter if they self-identify as 'gamers' or not, people who play games often share certain characteristics such as *gaming capital* (Consalvo, 2007), their accumulated knowledge, experience, and attitudes towards games and gaming. This includes gaming literacy, meaning the degree to which an individual feels confident in

understanding games and how they are played. Such concepts, skills and knowledge can be highly transferable. For example, participants who have played a game with a particular control scheme (keyboard and mouse to navigate a first-person shooter) are likely to have an advantage at similar games with similar control schemes over someone who is not familiar with them. Similarly, playing shooter games has been found to improve spatial cognitive abilities (Granic et al., 2014).

It seems likely that gamers will be more interested in taking part in a study involving games compared to non-gamers. This is evidenced in online surveys, where proficient players may be over-represented (Khazaal et al., 2014). As a result, games-based research may attract a participant sample with particular skills and knowledge that deviate from the general population. Relatedly, gamer identity and gaming capital may interact with using a game for data collection. Participants who do not identify as a gamer (e.g. older adults McLaughlin et al., 2012) may suffer stereotype threat: by anticipating that they will perform badly, their actual performance is decreased (J. L. Smith, 2004). Similarly, expectations about gameplay may heighten evaluation apprehension. While self-identifying gamers may find it socially desirable to perform well in a game and exert extra effort, non-gamers still often view games as a waste of time. Thus, non-gamers may want to downplay their investment in and performance at games, unless the gameplay has a socially acceptable justification (Deterding, 2017). Put differently, social desirability may produce significant performance differences between participants identifying as gamers or non-gamers.

Finally, the use of games as a research instrument may have differential effects on attrition, the drop-out of study participants over time. Some degree of attrition is common with long-running studies, and game-based interventions are in fact sometimes use to promote long-term behaviour change, such as *Zombies, Run!* (Six to Start, 2012) or *SuperBetter* (D. Johnson et al., 2016; Roepke et al., 2015). However, self-identifying gamers or participants with high gaming capital might be be more likely to persist with a game-based experiment or treatment, or in contrast, abandon a study earlier because they have 'higher' expectations of game design or perceive the intervention as less novel.

2.4.2 Gamers are Diverse

The 'gamer' stereotype of a white heterosexual male teen (Shaw, 2012) doesn't reflect the growing diversity of people who play games (Williams et al., 2008). That said, playing

a game is generally seen as a voluntary activity that individuals self-select into based on their preferences (Deterding, 2016b). Existing gaming literacy, socio-economic status, and the like may also affect what kinds of gaming devices and games people access. Hence, certain player characteristics may therefore be over- or under-represented among the users of certain games, genres, or platforms. While recruiting from a console multiplayer first-person shooter may result in a more white, male, gamer-identifying 'core gamer' sample, recruiting from a mobile casual puzzle game may result in an older, more female sample. These demographic characteristics all present potential confounds.

A related problem frequently raised in the literature is that players typically show different degrees of expertise in different game genres (Latham et al., 2013). Games may differ substantially in the skills they involve, one 'action' game may train twitch skills, while another may train executive processes. Differential effects of training with different video games were identified by Subrahmanyam and Greenfield (1994). Put differently, gaming literacy is not a unitary construct. Researchers should control for genre-related expertise to ensure differential representation across condition doesn't confound results.

A third common difference among players are so-called *player types* (Hamari & Tuunanen, 2014). Different people display different stable preferences in play activity and style: they may prefer exploration, socialising, winning, or something else altogether. These interpersonal difference again may confound results if not controlled for.

2.5 Discussion

Games, including elicitation games, are commonly used to collect data for research. While there has been some work on the validity of games-centred media effects and learning research, little has been done on the validity threats of using games to collect data for non-game-related research questions. I therefore offered a systematisation of potential validity threats of game-based research. While game-based research is just as fallible to general threats to validity (Shadish et al., 2002) such as publication bias (Ferguson, 2007) or issues surfaced in the current debate on reproducible research (Munafò et al., 2017), I focused on validity issues characteristic for games and their players.

This systematisation found that such threats to validity arose within four categories: systemic complexity, variance, framing, and game players. These are characteristic properties of games. As such, these threats are to an extent unavoidable with the use of games. While there are more complex, more variant, and more strongly framed games, no game is

entirely without them.

2.5.1 Moment-by-moment Validity

The presence of such characteristic threats to validity in games is a challenge to the traditional approach within experimental design, which is to rigorously minimise such complexity and variance by exerting strong experimental control. The familiar way to resolve this conflict is to make games more like familiar controlled experiments. This obviously has an impact on the design of novel games for data collection. A number of particular ways to address specific threats have been introduced throughout the chapter. However, I suggest that such approach can be taken too far. The typical end results of this are games played in most ungamelike situations and tasks that, while they share properties of games, are not 'whole' games (Washburn, 2003).

To insist that the games used are whole games, as with elicitation games, is a departure from this norm. Yet, to make enjoyable applied games such as elicitation games we might demand that, for the games we use, enjoyment remains a primary goal of game design. If so, and if the categories of validity threat described above are essentially unavoidable in 'whole' games, it is clear we must find a different way of conceptualising validity; one that is both convincing to researchers and compatible with the messy reality of games.

It cannot be that a satisfactory framework for the design of elicitation games simply enumerates game features to include or exclude in a modular fashion: the same characteristics that give rise to enjoyable gameplay can also threaten validity. As such, validity considerations must be integral to game design choices (in contrast to the gamification+validation approach). Similarly, consideration of validity must be contextual to the specific game used and gameplay situation. It clear that we cannot think of *games* as an artefact being valid, but rather only *instances of gameplay* being valid (for a particular purpose). Furthermore, due to the variance *within* a single gameplaying session, I suggest that validity can only be poorly understood as a molar property of gameplay as a whole, and is better understood at the molecular level as the aggregation of potentially diverse *moments of gameplay*. It is these moments that need to be designed with validity in mind, as well as enjoyment.

This moment-by-moment focus inverts our notion of experimental control. While we demand less control over the overall game experience – such as what challenges the player will face in what order – we need to know much more about what constitutes the individual moments of gameplay from which we derive our data. For instance, instead of presenting a

series of game levels identical in size and complexity, we might instead design to ensure that the differences in size and complexity have no effect on the moment-by-moment provision of data in the dimension of interest to us.

2.5.2 Validity Threats of Games

Throughout this chapter I have focused on validity issues characteristic for games and their players. The latter are maybe the most straightforward to address. Games and especially particularly game genres still attract particular populations with particular preferences and abilities. To some extent, this issue is self-correcting as game-playing becomes evermore prevalent and normalised across populations. Remaining bias can be controlled for with relatively standard research design measures, or simply documented as a limitation.

A less straightforward validity issue is that games are complex systems from which gameplay and player experience emerge non-linearly. This makes it fundamentally difficult to manipulate just one game parameter (as an experimental treatment) without potentially also changing many others, producing potential confounds that threaten internal validity. And because much of the enjoyment and engagement of games revolve around curiosity stoked by novelty and competence fuelled by experiences of learning, games can show strong maturation and attrition effects: playing the same content twice just isn't as fun or challenging as the first time around. The standard methodological responses are larger sample sizes with between-subject designs or within-subject designs with randomised ordering of conditions (Shadish et al., 2002). Another solution may be to pretest manipulations prior to the actual study, involving game design expertise to ensure no unexpected emergent confounds manifest.

Furthermore, I found that games, game content, and gameplay are highly varied, and researchers often have relatively little control over what players do and experience in a game and in what order. This makes statistical testing more challenging (or at least often requires larger sample sizes), and threatens internal, construct, and external validity. Different games vary on many dimensions, including crucially game genres. This cautions against operationalising different conditions of constructs as different games, or generalizing findings from one game to another, let alone other game genres. In educational research, authors like K. Squire (2011) have therefore called to replicate studies on the effects of particular design or instructional strategies across *multiple divergent games* before making any more general claims as to their effectiveness. Because games are far less standardised

than e.g. psychometric tasks and instruments, researchers should also document the games used in any publication in as much detail as possible, including screenshots or video figures demonstrating gameplay, the version and section of the game played when using existing entertainment games, and ideally, an executable mirror of the actual game. Otherwise, reviewers and readers will have difficulty assessing the validity of the reported findings, as will researchers interested in replicating the study.

Beyond such first stabs at mitigating strategies, I think the issues of systemic complexity and variance point to a more fundamental research need in games that is as much theoretical as methodological. Games and gameplay can be described on multiple levels of organisation (Klimmt et al., 2007). Developer experience suggests that some or even most of the socially and psychologically functioning mechanisms are located on 'higher', molar levels of organisation (Hunicke et al., 2004). For instance, providing *personalisation* during character creation should in aggregate increase autonomy need satisfaction, even if each individual player may have interacted with the personalisation interface and therefore have experienced a different moment-to-moment chain of actions and screen events (Turkay & Adinolf, 2015). However, there is little if any consensus nor methodological good practice on what level of organisation to study and manipulate games, or how to theorise and identify causally active mechanisms as constructs. Arguably the most progress in this respect has been made in gamification research, but even here, researchers are mainly pointing out a massive agenda of future desiderata (Deterding, 2015; Landers et al., 2018). By comparison, health sciences have now engaged in a decade plus of work modelling and identifying 'behaviour change techniques' as the psychological active ingredients in health interventions and are still far from any comprehensive consensus (Michie & Johnston, 2013). Even if a consensus construct (let alone taxonomy) existed, we would still need methods to quickly and reliably identify which kinds of active ingredients a given game held and what subcomponents of said games are involved in each.

And yet all this future work would only address the variance and emergent complexity of *games*, not *gameplay*. What differences in gameplay actually make a difference? When and why can we disregard 'low-level' differences in player behaviour because they all instantiate the same molar types of action (W. M. Baum, 2002)? What kinds of higher-level dynamics does gameplay reliably gravitate towards (Vahlo et al., 2017)? We are arguably even further from being able to answer these questions.

A final game characteristic threatening validity I identified was social framing: research

studies and gameplay are both very particular types of social situations with very particular orderings, norms, roles, and expectations, which may already be triggered simply by verbally and visually labelling a situation as 'a game' or 'an experiment.' While some evidence suggests a close correlation of in-game and real-life behaviour, some suggests marked differences (Deterding, 2016b). Scholars like Williams (2010) have therefore called for a systematic research programme on the mapping of real and virtual worlds. Ten years later, we are still dearly in need of such a concerted effort. By highlighting two characteristic features of play situations – lowered consequence and a license for strategic action – I hope to have given some starting points for it.

2.6 Conclusion

If the elicitation games we design are to collect valid data we need an understanding of the threats to validity that they are likely to face as games. Here I have identified four categories of threat that are characteristic of games. The categories are systemic complexity, variance, framing, and game players. Within each I discussed specific threats to validity highlighted in the literature. While one response to such threats is to avoid experimental tasks with these properties, minimising and controlling the complexity or variance of tasks for example, this is not compatible with fully embracing the use of 'whole' games designed for enjoyment as a primary goal.

In order to design elicitation games that collect valid data I suggested that we need a perspective on validity that lets us work with the messy reality of games *qua* games. The variance inherent in games and gameplay suggests we might look to the moment-by-moment level. Here we would consider the factors that will threaten the validity of data elicited in a given moment of data provision. One such factor must be motivation. How an individual is motivated to act – what data they are motivated to provide – will significantly, perhaps principally, impact the validity of the data collected in a given moment. Thus we need to understand motivation for (specific) data provision. What are the motivations acting upon the participant at each moment of gameplay? How does this affect what data the player provides (or whether they provide data at all)?

In the next chapter I turn to developing a theory of motivation for producing (specific) speech data in a game. To do this I ask the questions *Why does a player speak to a game?* and *Why do they say what they say?*. This is an attempt, in a concrete case study, to identify those factors that affect motivation for specific data provision. Such motivation

will be inextricably bound up with validity of elicited data. The perspective developed in the present chapter and the chapter that follows will help to form the integrated model of motivation and validity presented in chapter 4.

Chapter 3

Motivating Speech Data in Games

Settlers of Catan (Teuber, 2020) is a complex multi-player board game with dice rolling, building and trading mechanics. In *Catan* players compete to be the first to achieve 10 victory points by building villages, roads, and cities over a hexagonal grid of tiles. Through rolling the dice, villages and cities allow you to gain resource cards which you can trade or spend to build more villages, roads and cities. *Catan* has 4 pages of rules (Teuber, 2020), and only once, in the section on trading, does it mention speaking[1]. Yet during play, if you take the time to listen, you will observe a wealth of utterances – speech data – throughout the whole game. What is it about *Catan* that gets people to speak?

Figure 3.1: The game *Settlers of Catan* during play[2]

[1] *"You can announce which resources you need and what you are willing to trade for them. The other players can also make their own proposals and counteroffers."* (Teuber, 2020, rulebook p. 4)

[2] Photograph by Matthew Batchelder www.flickr.com/photos/borkweb/355808183/

On the face of it there are many answers: *because it's fun*, *because we like to socialise*, or perhaps *because it makes it easier to win.* These are all valid reasons that play a part in addressing why people speak. *Catan* is a multi-player game, so playing it is a social experience, and social experiences generally involve talking. Trading resources would certainly be harder without speaking, and resource trading is a valuable mechanic of you are trying to win the game. And it's easy to imagine that a completely silent game of *Catan* might not be very fun: no bragging, no boasting, and it being hard to follow what other players are doing.

The elicited language in a game of *Catan* is very different from that in a game of *Snap*, a children's game in which players take turns to reveal playing cards and shout the word 'snap' if a card is identical to the one before it. Why this difference? It might be that different kinds of players tend to play these two games, and thus the speech is not controlled by the game, but rather by the player. So what about *Snakes and Ladders*, another children's game? Different language is elicited again. Now imagine two different games of *Snap*, one played with a child and another with a colleague. While there will be many differences, we expect that at least some of the speech present would be more or less the same. In both cases, players will be saying the word 'snap', quite possibly with some speed in order to get in first. Clearly it is not just the players that determine the speech observed. Intuitively, something about the design of *Snap* predicts a 'core' of language that will be elicited. What are these design properties, and in what ways can they control the elicited speech?

The goal of this chapter is to develop a novel theoretical model to account for *why players engage in specific speech providing interactions with a game*, and explain *why they say what they say.* The place of this model in this thesis is to provide a foundational case study to understand motivation for specific data-providing actions in a game. The model can also facilitate the design and analysis of games for eliciting speech data. Linguistic data is selected as a case study that will be adopted throughout this thesis. As a model of motivation that is intended to inform design, it is desirable that variation in speech should be accounted for by those properties that can be controlled by the game designer, and not properties of the player or the context that the game designer would be unable to control. Thus I seek to model the link between formal, structural properties of game designs with the speech that is elicited.

In section 3.1 I motivate the present chapter within this thesis and the relevant liter-

ature. Section 3.2 then describes the methodological approach adopted for this chapter and give an account of the theory development process. The next two sections present the two submodels developed: Actuation in section 3.3.1, and Communication in section 3.3.2. Section 3.4 is a general discussion of the models. Section 3.5 concludes the chapter with a summary of takeaways for the rest of the thesis.

3.1 Motivation

Elicitation games seek to elicit data from their players that is useful for some purpose. For such games to be successful, the players must consent to provide this data by choosing to interact with the game in specific, data-providing ways. For example, players of *The Great Brain Experiment* (H. R. Brown et al., 2014), if they are to provide useful data, must engage in the four mini-games rather than spend all of their time in menus. When playing the mini-games they must take actions, rather than idle. If not all parts of a game provide data, it is not enough for a game to merely motivate play *per se*, rather it must motivate the kind of play that provides the desired data. To design successful elicitation games we must understand why players are motivated (or unmotivated) to engage in such data-providing interaction, and how this relates to the data they provide (and ultimately its validity). I adopt linguistic data as a case study throughout this thesis, and so here I will consider speech data and thus the design of games for the elicitation of speech. For this question, the existing literature on motivation in applied games, and speech interaction in games proves insufficient.

3.1.1 Speech in Games

Somehow, game designers regularly succeed in motivating players to speak, whether by accident or design. The social nature of many (particularly multiplayer) gaming contexts has always presented an occasion for speech in the form of socialising (Drachen & Smith, 2008; McGee et al., 2011; J. H. Smith, 2006; Stenros et al., 2009). Further, games have explored a wide variety of communication and speech interaction mechanics that not only passively facilitate, but actively motivate, speech between players. For example, the board game *Taboo* (Hersch, 1989) is not only an occasion for socialising, but incorporates rules that both require and restrict what is allowed to be said. Similarly, single player digital games have experimented with speech interaction ever since the technology existed to

support it (Allison et al., 2020). Such examples presumably encode design knowledge about player motivations for speech.

Within the academic literature, multiple areas inform research on the motivations for speech in games: Accessibility within HCI, Applied Games, Player Communication, and Voice Interaction.

Accessibility Even though research designing games that incorporate speech interaction is common, the source of player motivation to speak in such a game has received little attention. Within the field of HCI, existing work on speech interaction with games has focused on primarily on designing for accessibility with evaluations of usability rather than motivation. For instance, work as been done on creating accessible versions of games (Grammenos et al., 2007), such as Space Invaders (Grammenos et al., 2009) and Sodoku (Norte & Lobo, 2008), and enhancing the accessibility of existing games (e.g. Derboven et al., 2014; Harada et al., 2011; Mustaquim, 2013). Such studies, adapting existing game designs, do not consider the motivations for speech beyond the useability of the game.

Applied Games Speech interaction has been used in language learning games (e.g. W. L. Johnson, 2010; McGraw and Seneff, 2008; Silva et al., 2011 and games for speech rehabilitation and therapy, and emotion regulation (e.g. Cler et al., 2017; Duval et al., 2018; Fernández-Aranda et al., 2012). Among these applied games of language learning, two also act as a tool for collecting speech data (Gruenstein et al., 2009; McGraw et al., 2009). However the question of motivation for speech interaction as distinct from gameplay in general has not been significantly addressed.

Player Communication Studies have looked at the role of communication (predominantly text communication) in games for e.g. affording social interaction (Lazzaro, 2004; McGee et al., 2011; Stenros et al., 2009) or enabling coordination and collaboration (Manninen, 2004). The use of communication for both instrumental and non-instrumental purposes is commonly identified (J. H. Smith, 2006, p. 169): while some players want to win – often engaging in surprisingly little non-instrumental social interaction (Muramatsu & Ackerman, 1998) – others make creative and social use of communication methods (Wright et al., 2002). This distinction is reminiscent of Bartle (1996)'s taxonomic distinction between Achievers and Socialisers among players of MUDs. It has also been observed that player communication in parallel to gameplay can be sparse, perhaps because of cognitive

demands of gameplay (J. H. Smith, 2006).

Drachen and Smith (2008), whose research focused on roleplaying games, frame three hypotheses as to why players speak in games. He suggests speech may be *functional* (motivated by practical needs in the game for e.g. coordination, and as such likely to be determined largely by the game), *strategic* (directed toward furthering in-game goals, and a predictable result of rational player behaviour), and/or *social* (and thus only tangentially related to the game). J. H. Smith (2006) effectively tests the strategic hypothesis, analysing whether player talk is game-theoretically optimal for players. He codes transcripts of same-screen multiplayer gaming sessions for quantitative analysis. His quantitative analysis finds that a simple strategic account is insufficient to account for all of player speech. Drachen and Smith (2008) coded transcripts of pen-and-paper and multiplayer computer role-playing game play. They observe that difference in format leads to differences in the frequencies of different types of utterance, however much of this is due to the interfaces and contexts in which these games are played rather than generalisable properties of their game design. Both the models of J. H. Smith (2006) and Drachen and Smith (2008) are developed on few games; Drachen and Smith look at only a single genre, role-playing games. Finally, neither model is particularly suited to use in design.

Previous studies have been restricted to narrow domains and most of them have considered player communication via text rather than speech. While multiplayer roleplaying games have been frequently discussed (e.g. Drachen and Smith, 2008; Ducheneaut and Moore, 2004), overall very few genres have been studied. None of the existing studies have looked at games designed to elicit speech for some real-world use. No study has considered a broad spectrum of different types of game, nor has any attempted theoretical sampling to systematically identify different kinds of games or ways that games could elicit speech. Design principles recommended have been general and not directed at controlling the particular speech elicited (e.g. Ducheneaut and Moore, 2004). As such, a broad, purposely divergently-sampled, design-focused approach to identify general game design properties that elicit speech is warranted.

Voice Interaction Early research on voice in games focused on social and team dynamics and cooperative efficacy when using voice chat (Halloran et al., 2004; Salinäs, 2002; Williams et al., 2007), and the player experience of voice chat in multiplayer games (Wadley et al., 2015). Social and cooperative benefits arising from the richness of voice as a medium

for communication (Daft et al., 1987) are balanced against challenges in controlling self-presentation, background noise, speaker confusion, and the desire to role-play a fictional identity (Wadley et al., 2015).

There had been little research into voice interaction in *single-player* games before Allison's (2020) thesis on the subject. Players can experience a significant dissonance between player identity (speaking an instruction *to* a character) and avatar identity (speaking *as* the character) (Carter et al., 2015). This has been analysed with respect to the multiple nested social *frames* of such gameplay (Allison et al., 2019), which as understood in Goffman's (1986) Frame Analysis establish norms and role expectations of participants. Briefly, players switch between experiencing themselves either (a) as a character, (b) as performing strategic actions in a game, (c) as functionally controlling a system, or (d) as participating in a social interaction.

Allison et al. (2018) identifies game design patterns (Björk & Holopainen, 2006) for digital speech games, including single-player games, within a corpus of digital games. Such design patterns give common ways in which games are design to incorporate speech or voice control. For example, *who-what-where commands* allow the player to give instructions consisting of subject (who), action (what) and place (where) to computer-controlled agents; and *speak as a character* is a design pattern where players (to actuate some kind of mechanic) speak the words a character would use in the game. However, while these give specific, successful strategies that have worked in games, they do not themselves account for *why* they are successful at motivating speech interaction, nor what kinds of speech they motivate. Further, such a taxonomy doesn't help with designing novel interactions that might be necessary when attempting to elicit novel kinds of data with a game.

In summary, the existing research on the motivations for speech in games is piecemeal. While different work has addressed particular games and game genres, there has been no systematic overview of what motivates speech across a wide range of divergent games. Furthermore, such models and design patterns that describe what existing practice in entertainment games is not easy to generalise for creating novel design solutions for eliciting speech with applied games.

3.1.2 Motivating Behaviour in Applied Games

Research on motivation and motivating behaviour has come from three sources. In this section I first address research in applied games, particularly applied educational games (serious games) about motivating engagement to achieve outcomes. Second, I discuss general psychological theories of motivation that address why, and in what ways, people are motivated or unmotivated towards different tasks. Both areas of research have focused primarily on generalised motivation, and not motivation for particular action. Finally, from the game studies literature I discuss the Rational Player Model of J. H. Smith (2006), which is a rational choice model explaining motivation for particular behaviours.

Applied Games Existing approaches to the motivation of desired outcomes in applied games going back to T. W. Malone (1981) have analysed the game as a whole. In educational games, taxonomies enumerate 'game characteristics' whose presence will motivate play (e.g. Garris et al., 2002). They include 'instructional techniques' (Wouters & Van Oostendorp, 2017) to motivate and achieve outcomes. Models of serious game outcomes attempt to explain this as the result of game- and context-wide properties such as classroom structure and curriculum integration (Vandercruysse & Elen, 2017). The compelling concept of alignment (i.e. of gameplay with the intended outcome), as it has been formalised in serious games as Intrinsic Integration (the alignment of instructional content and gameplay) (Habgood & Ainsworth, 2011), is presented as one more property that a game (as a whole) may or may not possess. Within this perspective, research seeks to "[assign] credit to particular [whole-]game features in affecting outcome measures of learning and motivation" (Graesser, 2017, p. 210). Graesser goes on to lament that "the value-added of a particular feature is quite complex because of the complex correlations, interactions, and trade-offs inherent in the large set of factors" (Graesser, 2017, p. 210). Such complexity is hard to turn into clear and generalisable design guidance. We can expect that elicitation games would face similar challenges if we similarly adopted a whole-game perspective on motivation.

Psychological Theories of Motivation Theories of motivation in psychology may explain behaviour either at the molar or molecular level (Littman & Rosen, 1950b). Molar theories describe whole *behaviour acts* like playing a game or going for a walk, where as molecular theories describe individual units of behaviour, like pressing a button. A

molar view of game motivation dominates the player motivation literature, manifest in the descriptive levels of the two most frequently used theories: Self-Determination Theory (Ryan & Deci, 2000), and Flow (Chen, 2007). In contrast, molecular theories, such as behavioural approaches, have seen more limited use (Linehan et al., 2015; Linehan et al., 2009; Yee, 2001). While molar theories of motivation are suitable for understanding motivation for gaming overall (and are easily measured with post-test questionnaires), they can only make weak predictions when it comes to understanding moment-to-moment behaviour within a game (Kumari, 2021). I will make this argument exemplarily with a highly popular theory of player motivation: Self-Determination Theory (Boyle et al., 2012; Tyack & Mekler, 2020).

Self-Determination Theory (SDT) (Ryan & Deci, 2000) is a general theory of motivation very popular in games HCI (Tyack & Mekler, 2020) and gamification (Loughrey & Broin, 2018) research. It is one of a number of motivational theories based on need satisfaction (e.g. Bostan, 2009; Sherry et al., 2006). It posits three universal psychological needs that motivate behaviour: Autonomy, Competence and Relatedness. These three needs can be understood at motivating behaviour at different levels of abstraction: to an individual's general life outlook (global), or *causality orientation* (Deci & Ryan, 1985)), particular contexts within their life such as learning the piano (contextual), or particular situations (situational) such as a gameplay session (Vallerand & Ratelle, 2002). Activities that satisfy needs (e.g. if playing the piano gives a feeling of competence) are motivating.

SDT identifies a key duality in motivation between *intrinsic* motivation – performing an activity solely for its own sake – and *extrinsic* motivation – the kind that it is often assumed game elements such as points, high scores, and collecting powerful in-game items provide (Lafrenière et al., 2012). However, while if a player is motivated to play *in order to* attain such an item we can say they are extrinsically motivated, this is far from saying that encountering a powerful items in the game will necessarily lead players to be more extrinsically motivated. While extrinsic rewards harm intrinsic motivation in general (Deci et al., 1999), rewards interpreted informationally as feedback instead can enhance intrinsic motivation for an activity (Aronson, 1985). Intrinsic motivation is desirable in applied games design as it is likely to lead to enjoyment (Ryan et al., 2006).

While we would like our elicitation game to be intrinsically motivating to gain the *situational* benefits, including voluntary play, it is far from clear whether SDTs account of intrinsic motivation also applies to the *moment-to-moment* level of why particular data

providing actuations were performed at a particular moment of play. It is hard to draw concrete design direction from the psychological needs posited by SDT (Autonomy, Competence, and Relatedness) at the level of moments of gameplay. For example, would the pursuit of Autonomy require that providing particular data should always be optional? Would making providing data difficult lead to positive effects on Competence, or negative effects on Competence? These questions are hard to satisfyingly answer at the moment-to-moment level. Rigby and Ryan (2011) are typical in giving design recommendations that aim to keep the player playing the game as a whole, such as as providing optimal challenge and giving competence-inspiring feedback, and not providing direction for motivating the player to perform action A over action B. Indeed, the standard autonomy-promoting advice of giving players meaningful choices is likely to make it *harder* to control what specific actions a player performs.

Models of the Player At the other extreme, the Rational Player Model of J. H. Smith (2006) is a strong theory that predicts player actions within a game as precisely the optimal strategic move in a given moment. Significantly, the theory addresses the molecular level of moment-to-moment game actions in a way that affords making specific, concrete predictions about how players will act. This is certainly the kind of theory that we want for the purposes of designing novel elicitation games.

However, while a useful lens for design – and indeed literal textbook practice (Schell, 2009) – J. H. Smith's (2006) own studies found that it was not sufficient to explain real player behaviour. Naturally so, as by design it considers only formal game properties and doesn't take account of potential game-external motivations. Data provision is fundamentally a game-external activity as it must originate in the physical, real-world actions of the player (even if those are just pressing a button). As such, while we might expect the Rational Player Model to be a useful, if incomplete, model for analysing specific, moment-to-moment *strategic decisions* in games, it is likely too simple to model moment-to-moment *data provision*.

3.1.3 Summary

As it stands, no current model of motivation in the literature suitably explain why players choose to speak and say the things that they say when playing a game at a moment-to-moment level, let alone a model suited to the design of games intended for the purpose of

eliciting such speech. While we have design patterns for voice-interaction in digital games, these are only loosely concerned with *how* the player is interacting. While we might look to existing games for inspiration, there are relatively few examples of these. Those that have been created that collect speech data (Gruenstein et al., 2009; McGraw et al., 2009) do not present generalisable design knowledge. Non-speech eliciting games generally use designs were feedback can be given for user inputs. As such, they do not provide templates that can be readily adapted to novel data types. We do not have a readily available model of why a player provides the data they do to a game in terms of things the designer can control. J. H. Smith (2006) provides a starting point, but one that is as yet critically incomplete. In this chapter I develop a model for the specific case of games that elicit speech and interpret it in terms of the concrete design guidance it implies. My research question is therefore "What controllable factors motivate players to produce particular speech data in a game?"

3.2 Method

A qualitative study using mixed document analysis was performed following the principles of grounded theory (Corbin & Strauss, 2008; Glaser & Strauss, 2010), which is an appropriate tool for developing new theory where none previously exists. The goal of this study was to develop a model that could inform the design and analysis of games (analogue or digital) for the elicitation of speech. In the absence of existing speech elicitation games, my sample was based on the closest comparable phenomenon: entertainment games involving speech and microphone interaction.

Games were coded from one or more of: autoethnography, videos on the video-sharing site YouTube[3] (including Let's Plays), informal discussions with players, game descriptions and reviews on sites such as BoardGameGeek[4], forum posts, and official game rules. This was chosen as an economical method to achieve a very diverse sample.

A dataset of 344 digital games that include microphone input was generated semi-systematically by searching the video game database MobyGames[5], video game review site GiantBomb[6], and the online gaming portals Kongregate[7] and Newgrounds [8]. Games were

[3] www.youtube.com
[4] www.boardgamegeek.com
[5] www.mobygames.com
[6] www.giantbomb.com
[7] www.kongregate.com
[8] www.newgrounds.com

identified using the search terms and categories 'microphone' and 'mic'. A small further set of board games were identified when needed by searching the website BoardGameGeek by board game mechanic. This process sensitised me to the domain and provided a loosely categorised list within which to identify theory-generating examples, without undermining the grounded theory principle of theoretical sampling (Glaser & Strauss, 2010). In addition to this, I often identified other examples from my own experience (particularly where the games are not generally considered to be about communication). I also searched for theory-generating examples when it seemed plausible that there would be a game that challenged an aspect of the emerging theory, for example a game that specifically *discourages* speech (e.g. *Alien Isolation*, Creative Assembly, 2014), or a single player game that involves social motivations for speech (e.g. *Seaman*, Vivarium, 1999).

The grounded theory process began with a small selection of games that all used speech interaction (both player-computer and player-player) taken from this larger list. These were identified as a set that represented widely varying types of games. Initial codes were closely justified by the games *prima facie*, for example coding based on genre, medium (computer/board game), number of players, and input device (Microsoft Kinect, Nintendo 64, etc.). As this coding was applied to more games, it became clear that such codes would be little practical use. Coding proceeded to more interpreted properties, such as the type and manner of feedback, the 'situatedness' and representation of the language within the game world, goals, types of constraints, and so on.

I iterated on the codes and theory. Frequently the theory would evolve 'all at once', as a new code triggered a perspective on the dataset and many of the other codes had to be realigned with a new approach. Iteration in the coding process was rather flat with a great deal of revision of the base codes, which were themselves relatively abstract descriptions of gameplay. As the process progressed, games were split up into mechanics which were coded separately. Initially codes described properties of games or mechanics overall (e.g. style of feedback mechanism) but more progress was made once codes described moments of gameplay experience (e.g. the value of a particular utterance, or the effort of speaking).

I made some conscious choices during the coding process based on by research goals. The desire to attain a theory that would be useful to guide design was perhaps the strongest factor. I focused on properties of the game itself that could be manipulated by a game designer more than properties of the player or gaming context. I chose to keep the theory agnostic to gaming modality to avoid reifying merely contingent differences between e.g.

analogue and digital games that would be easy to challenge with novel games by digitising an analogue game or making an analogue adaption of a digital one. I did indeed observe significant differences between board games, digital games, and roleplaying games. By integrating these into a single model I hope to have got at more interesting and generalisable properties of the games' design. Finally, I did not transcribe game sessions and undertake line-by-line coding. This was economical in allowing me to look at a wider range of games in less detail and with a stronger focus on the causal influence of formal rules for the benefit of design. Indeed, due to the similarities between some of the games (and the simplicity with which many of them used speech mechanics), I found many of them could be understood after limited play and that watching 'Let's Play's and and reading reviews was a viable strategy for understanding them. Bartle (2010) describes this as 'groking' a style of game.

One of the primary data collection tools I used was autoethnography, which in Ellis et al.'s (2011, p. 273) definition is "an approach to research and writing that seeks to describe and systematically analyze personal experience in order to understand cultural experience". Within HCI and games research autoethnography has been used to understand users and players as a relatively quick and inexpensive alternative to ethnography which avoids the typical challenges of data collection from participants (Cunningham & Jones, 2005; Rapp, 2018) while having the potential to generate empathy and reveal user experiences (O'Kane et al., 2014). My goal in adopting an autoethnographic approach was to access the experience of moment-to-moment motivation to speak or, significantly, to *not* speak while playing various games. Autoethnography provided a method for me to directly interrogate this experience, which is rarely documented and was not present in my other sources. The products of this work were memos containing my reflections on my gameplay during the period of the study. In writing these memos I focused on perhaps only one or two interesting moments rather than attempting an exhaustive description. In this way my memoing approach was similar to that from the Glasserian variant of grounded theory (Cole & Gillies, 2022). These memos served as data that informed the theory development process.

However, to reflect on a text requires critical distance yet authentic play requires immersion (Bizzocchi & Tanenbaum, 2011). Bizzocchi and Tanenbaum's (2011) response to this is to suggest that researchers oscillate between the two 'states' of player and scholar. This well describes my practice. As player I adopted different play styles but largely tried

to immerse myself in 'normal' play. As scholar I introspected about my thought processes, observed the state of the game, or carefully studied the result of taking particular actions. I was guided throughout by concepts from my developing theory and particularly by what 'felt interesting' as this reflected something that I could not yet explain. If my theory would suggest I should be motivated in a particular way, I would check if this felt true to my experience. I ran ad-hoc experiments: What happens if I (make myself) say something different? How do I feel? How do the other players react? Similarly I observed and questioned my co-players in an ad-hoc, opportunistic way.

Playing a game for research is further problematic because games are procedural and non-linear texts, meaning that different play styles can give rise to different experiences (Bizzocchi & Tanenbaum, 2011). My autoethnographic focus meant that I was not seeking to play in a standard or generalisable way but rather prioritised authentically observing my own play and experience. However, in analysing the games I did find it informative to contrast this with playing (and speaking) as I felt the game expected. This style of play is described by J. F. Van Vught's (2016) cooperative playing strategy. Through this lens, a game provides a range of cues to the player of appropriate ways to play. Concretely, games with speech mechanics often cue how the speech mechanics should be appropriately actuated through instructions, user interface design, and narrative framing. When this style of play conflicted with my natural inclination it tended to highlight ways in which the speech mechanics were uncomfortable or unnecessary. Finally I engaged in free, exploratory play (J. Van Vught & Glas, 2018) in order to understand how the speech mechanics worked.

Finally, beyond determining there was limited existing literature in the field, I did not substantially engage with the speech interaction literature until towards the end of the grounded theory process (in fact, much of the literature on voice interaction cited in this chapter was published once the model was largely complete). I note several correspondences with the literature in the discussion at the end of this chapter.

3.3 Results

Players are motivated to speak and to say what they say *at* a particular moment of play and *by* the context of that moment. That is, motivation for speech is not a property of a molar *game characteristic* (understood as a property that holds of a game as an artefact), nor a property of the player (as in popular trait-based models of player motivation, e.g. Bartle,

1996; Yee, 2006), but the outcome of narrowly situated moment of gameplay containing game, player, and wider context. In this sense, it does not make sense to talk about *games* motivating speech (except in aggregate). Rather games give rise to *moments of gameplay* which motivate speech[9]. This means that, to understand motivation for speech in a way that supports design, we need to reorient ourselves away from the popular *molar* models of motivation to consider particular utterances as *molecular* behaviours. The primary shortcoming in the literature here is the inability in models such as Self-Determination Theory to compare between multiple competing action candidates (at all, let alone at a molecular level). While Autonomy might be embodied by the choice between two ways of speaking at a particular moment, it cannot exist in the utterances themselves. While choosing the right utterance may inspire feelings of Competence, this again motivates the act of choosing, not the utterances themselves.

Motivations to speak are contextualised at the level of conversational interaction and game interaction feedback loops. These correspond to the Communication and Actuation submodels introduced below. Properties of the game or gameplay as a whole (among the first codes to be identified from the data) were found to have little utility for describing speech motivations. For example, that a game was multiplayer or had communication mechanics was found to have little necessary relation to whether player speak and what is said.

An early challenging example was the game *Mao*. *Mao* is a multiplayer card game with communication mechanics (specifically, the need to say certain things at certain times, such as "have a nice day" if you play a seven of any suit[10]). However, due to highly restrictive speech rules (speaking out of turn results in drawing a card as a penalty), *Mao* elicits relatively little speech compared to most other games. Yet it doesn't consistently elicit *no* speech. Rather, there are occasions during play when speech is elicited and occasions when it is not. However, while *Mao* starts with several speech rules, the dealer in a game of *Mao* is able to introduce arbitrary new rules to the game. Due to the relatively high share of speech rules in the base game, new rules are not unlikely to involve speech. They

[9]This reframing has already been introduced in the last chapter because it came to play a significant role in the research of thesis. It originally emerged out of the work described here as an unexpected but helpful way of restructuring the developing theory.

[10]There are many sets of rules for *Mao*, which contain similar requirements such as announcing the name of the game when you have one card remaining, saying "ace of spades" whenever you play an ace of spades, etc. The above rule was included in the games I have played, as well as in the rules as described here: http://www.mu.org/~doug/mao/maorules.html Accessed: 2022-02-11

can supersede existing speech rules. Moreover, all game rules are originally secret and discussion of the rules is expressly forbidden (although at least one player must know all the rules to play the game). Thus in *Mao*, players attempt to learn and abide by a changing set of (speech) rules, taking their cues only from punishments for infractions handed out during the game.

Rules do have a significant influence in encouraging or preventing speech. However, the dynamic changing of speech rules, and the speech they elicit in *Mao* is hard to descriptively capture at the level of the whole game[11]. In contrast, player motivations seem simple: say what you think follows the rules at each moment in the game. The problem is, at a macro level, this is underspecified: to understand what a player thinks follows the rules at each moment in the game, we must break apart the game as a whole into smaller units of analysis. Specifically, we see that the effect of both the communication mechanics and the speech restrictions in the game are dependent on the (player's mental models of the) formal state of the game in a moment of play.

Two complementary analytical lenses were found to account for such motivation for speech at a molecular level. These correspond to two complementary ways in which speech can be seen. First, speech can be seen as the communication of information. Here we analyse speech as if within a conversational turn, peculiarly constrained by the gameplay context. Second, speech acts describe the performance of speech as an action enacting changes in the world (Searle, 1965). Here we analyse speech at the level of a moment of interaction. Thus an account of the motivation for speech in games can be given either in the effect it creates (Actuation lens) or the information it conveys (Communication lens).

3.3.1 Actuation

Games are commonly described in terms of their mechanics, often defined as the 'verbs' by which the player can act in or on the game (Sicart, 2008). For example, a platform game might have running and jumping mechanics, allowing the player two means of moving their avatar through the game world. Such mechanics only exist by virtue of there being a process by which the mechanic is triggered, which we will here call *actuation*. If there is a jump mechanic in a platform game, it is by virtue of there being a button on the controller

[11] To introduce an "arbitrary rule" code as a mechanic by which any speech could be necessitated or prohibited (or both, at different times in the same game) is about as minimally descriptive as can be imagined.

that causes the jump effect. In this example, we can speak of *an instance of pressing* that button to be an *actuation* of the jump mechanic. If one could also jump by shouting the word 'jump', such an act of shouting would be another actuation of the same mechanic.

Actuation (hereafter capitalised) is one motivation for speaking to a game. In particular, Actuation motivates instances of speech to games when that speech (or voice-input more generally) constitutes an actuation of one or more mechanics. This is transparent when voice input is a direct replacement for another input device (e.g. Harada et al., 2011; Sporka et al., 2006). Actuation is the motivation arising from the fact that speaking 'to' the game effects a change in game state. However, such motivation is not a generalised property of the game (where games that generally allow speech-control afford 'greater Actuation'). Rather, Actuation is understood with regards to a particular utterance situated and contextualised within a particular moment of gameplay.

To give some examples, in *Tomb Raider: Definitive Edition* (Crystal Dynamics, 2014), a third person action game, the player controls Lara Croft who employs a variety of weapons to kill enemies. When played with the Microsoft Kinect's in-built speech recognition, spoken commands may be used to change weapons and reload. In *Ryse: Son of Rome* (Crytek, 2013), the player may issue commands to their allies by speaking key words and phrases. In *Dead Rising 3* (Capcom Vancouver, 2013) and *Splinter Cell: Blacklist* (Ubisoft Toronto, 2013) (Figure 3.2), key phrase detection triggers the avatar to taunt or alert enemies. In each of these cases, players will periodically find such outcomes desirable in the game. As such they may choose to actuate these mechanics using speech rather than an alternative button press. If they do so, we can say they are motivated by (the components of) Actuation.

Non-verbal voice controls lend themselves to being motivated by Actuation. *Blow Boat* (Edvin & cuber3, 2005) typifies a type of games which are characterised by a 'noisemaking' mechanic. *Blow Boat* (Figure 3.3) is a 2D side-scrolling game in which you race your sailing boat to the end of the level, while avoiding obstacles, by blowing into your microphone. Similarly, to disperse fingerprint powder in *Apollo Justice: Ace Attorney* (Capcom, 2008), cool cooking in *Cooking Mama* (Office Create, 2006), or to play a pan flute in *The Legend of Zelda: Spirit Tracks* (Nintendo, 2009), players blow to actuate the mechanic. In hobbyist games that use microphone intensity to fly a plane (*Jet Pass*, rigel2010, 2009), control a pong paddle (*Micropong*, karaokeparty, 2010), or take on the role of a death metal singer (*Deathmetal Sim*, killthemouse, 2005), the player's only means of control is

Figure 3.2: The game *Splinter Cell: Blacklist* (Ubisoft Toronto, 2013). The game has just detected the voice command "Hey You" and a nearby guard is about to leave their patrol route to investigate. Such an outcome can motivate speech or demotivate it if the outcome is unwanted.[12]

making noise into their microphone. Igarashi and Hughes (2001) describe three interacting techniques for non-verbal voice control in games: continuous voice, rate based control by pitch, and discrete control by tonging. Such vocalisations have no semantic content, but the information they do contain: intensity, pitch, and frequency, can be readily detected by the game in a way analogous to the position of a joystick or the pressing of a button[13].

Mechanics in analogue games can similarly be actuated, and thus motivated by Actuation. The only difference is that the feedback mechanism happens to be situated within the collective minds of the players, rather than in game code. For example, announcing a clue in a game of *I Spy* (e.g. "I spy something beginning with 'a"'), suggesting an answer ("Apple?"), and providing feedback ("Yes, it was Apple") can all be understood as transforming a formal game state, even though we cannot point to a representation of this game state. *Mao* is a similar example. Declaring "have a nice day" is (part of) an action, and its formal importance can be seen in whether or not a card has to be drawn as a penalty. These examples can be understood as actuations.

The Actuation model contains several components. Together, these provide an answer

[12]Screenshot from *DanQ8000* https://youtu.be/OWTeu4ystGc

[13]In a similar way, consciously controllable breathing patterns can be interpreted as actuations of a mechanic (Sra et al., 2018; Tennent et al., 2011)

Figure 3.3: *Blow Boat* (Edvin & cuber3, 2005), an example of a class of simple games using microphone noise to control the game. In this game the player blows into the microphone to push along their sailing boat in order to win a race.

to the questions *Why does a player speak to a game?* and *Why do they say what they say?* Consider a moment in the play of a game like *Tomb Raider: Definitive Edition*. Within the game, circumstances might lead the player to want to reload their weapon. To do so, only a proportionally small range of inputs will actuate the reload mechanic. This range of inputs is defined by the *Rule-Based Constraints* of the game. Meanwhile, the player is playing within a particular environment, perhaps with other people who might be disturbed. Our player's (possible) actuation is at the mercy of two things: *Efficiency* (what actuations are more or less efficient at desirably changing the game state), and *Appropriateness* (what actuations are more or less appropriate within the sociomaterial context of play). What is efficient or appropriate may be informed by the Rule-Based Constraints of the game. Desire for *Performance* adds a third influential factor on the speech elicited. This is illustrated diagrammatically in Figure 3.4.

3.3.1.1 Rule-Based Constraints

A game's rules indirectly – but significantly – determine what counts as a meaningful or *effecting* utterance in the game. While some games do not formally constrain what types of actuations are allowed, it is common in many games to have a grammar within which utterances must be constructed. This grammar might be reading aloud pre-written fill-in-the-blank sentences from cards as in *Cards Against Humanity* (Dillon et al., 2011). In *Cranium Hoopla* (Alexander & Tait, 2002), when giving comparison hints, players should give utterances of the form "It's bigger than blank but smaller than blank". A more complex example is the game *There Came an Echo* (Iridium Studios, 2015), where 'grammatical'

Table 3.1: The Coding Hierarchy: Actuation. The component dimensions of actuation affect the choice of whether and how to actuate a speech mechanic.

Code	Description	Example
Rule-Based Constraints	Constraints imposed by a formalisable grammar	Clues given must match template in *Cranium Hoopla* (Alexander & Tait, 2002)
Efficiency		
Optimality	The utility of an actuation towards in-game goal	Saying the keyword to reload weapon when out of ammo in *Tomb Raider: Definitive Edition* (Crystal Dynamics, 2014)
Effort Minimisation	The minimisation of effort/difficulty required for the actuation	Blowing, rather than speaking, to trigger an input based on microphone intensity level in *Blow Boat* (Edvin & cuber3, 2005)
Appropriateness		
Fair Play Norms	Constrained by voluntarily adhering to conventions of how to play	Speaking coherently in *Mario Party 6: Fruit Talktail* (Hudson Soft, 2005)
Situational Norms	Constrained to avoid actions inappropriate within wider sociomaterial context	Avoiding saying things offensive to people nearby in *Cards Against Humanity* (Dillon et al., 2011)
Performance	Actuations are sometimes enacted in an performative way	Desire to sing well in *Rock Band 4* (Harmonix, 2015)

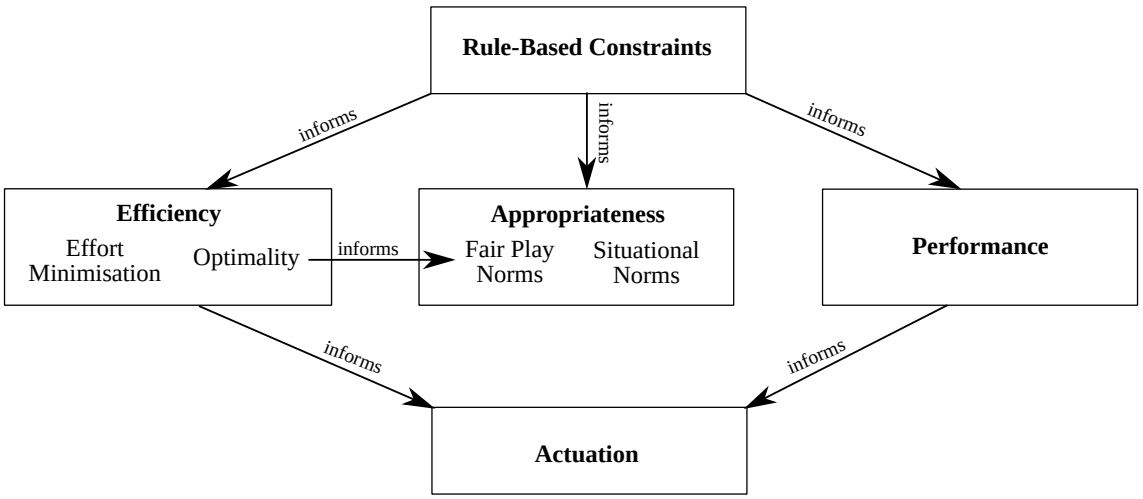


Figure 3.4: The Actuation sub-model. A actuation of a speech mechanic is informed principally by three groups of factors: Efficiency, Appropriateness, and Performance. The first two of these are informed by the Rule-Based Constraints of the input mechanism required for the game, if specified.

Figure 3.5: The game *There Came An Echo* (Iridium Studios, 2015) in which the player gives commands to a team of NPCs using voice commands. In this screenshot the player's squad is attacking from both the top and bottom of the screen. The labelled locations (Alpha 1, etc) make it easy to direct characters to move to particular locations or attack particular groups of enemies.[15]

actuations can be complex, such as "Corrin and Miranda, move to Delta 2 on my mark"[14] (Figure 3.5).

A common game design pattern in larps (a neologism for Live Action Roleplaying Games) is the use of 'effect calls'. When a player wishes to enact an effect, such as causing damage with a weapon, or casting a spell, they announce the effect for other players to hear using a string of key words. The other players then react appropriately to the effect based on their knowledge of the call system. For example, 'single' might cause a single point of damage the individual it is directed to, or 'knockback' require them to step or jump backwards to simulate being hit with a powerful force. Referees also make use of calls to manage gameplay, or use safety calls to pause or stop the game. Calls can be combined in specified ways. For example, in the UK festival larp *Empire* (Empire Wiki, n.d.), 'mass' is a prefix call used to direct an effect to everyone in a 20-foot cone in front of the player calling it (e.g. 'mass repell'). Such calls are actuations of mechanics. Because they rely on shared agreement on how the call system works (codified in the rules), players cannot

[14]https://youtu.be/mq72AOmtYb8
[15]Screenshot from PlayStation https://youtu.be/Zi1vgFh5BtQ

invent new calls. This system of permissible words and their meanings are what we here call Rule-Based Constraints of the game.

Speech recognition can define Rule-Based Constraints. Games with keyword-detection speech control, such as *Tomb Raider: Definitive Edition* (Crystal Dynamics, 2014) or *Dead Rising 3* (Capcom Vancouver, 2013), detect a set range of commands. In *Tomb Raider* this includes the names of weapons and 'reload'. Games with speech recognition that embed the who-what-where commands design pattern (Allison et al., 2018), such as *There Came an Echo* and *FIFA 2014* (EA Canada, 2013), detect a large range of phrases that fit a particular grammar.

Rule-Based Constraints inform how players speak to actuate mechanics in the game, but they do so shaped or enforced via two further factors. In solo-play contexts, the Rule-Based Constraints appear to be mediated by Efficiency (how these rules, implemented within the feedback system give reward certain behaviours). These could be called the *actual* Rule-Based Constraints. When other people are present, there seems to be an increasing influence of Appropriateness (players agreeing to abide by the set of rules as part of the social norms of play). These could be called the *ideal* Rule-Based Constraints.

3.3.1.2 Efficiency

Players typically construct actuations to be a maximally efficient (and effective) means of effecting a desired change in the game. Thus efficient actuations are preferred to inefficient ones. In the absence of other factors (i.e. Appropriateness), this leads to a convergence of player behaviour on the actual mechanism by which the desired change is effected (the actual Rule-Based Constraints). As solo-play contexts have minimal influence of Appropriateness, they thus lead to actuations determined solely by Efficiency. Two sub-dimensions of Efficiency were identified: *Optimality* and *Effort Minimisation*.

Optimality A significant influence on the efficiency of a (potential) actuation is the expected strategic utility of that actuation within the game. Different mechanics (and different actuations of the same mechanic) can lead to divergent outcomes. In the simple children's game *I Spy*, a player is given the first letter of the name of a nearby object. Their goal is to respond with the correct word. The player knows the word must begin with a particular letter, and so does not (except by mistake) respond with any word not beginning with the correct letter. Similarly, the player in *Tomb Raider: Definitive Edition*

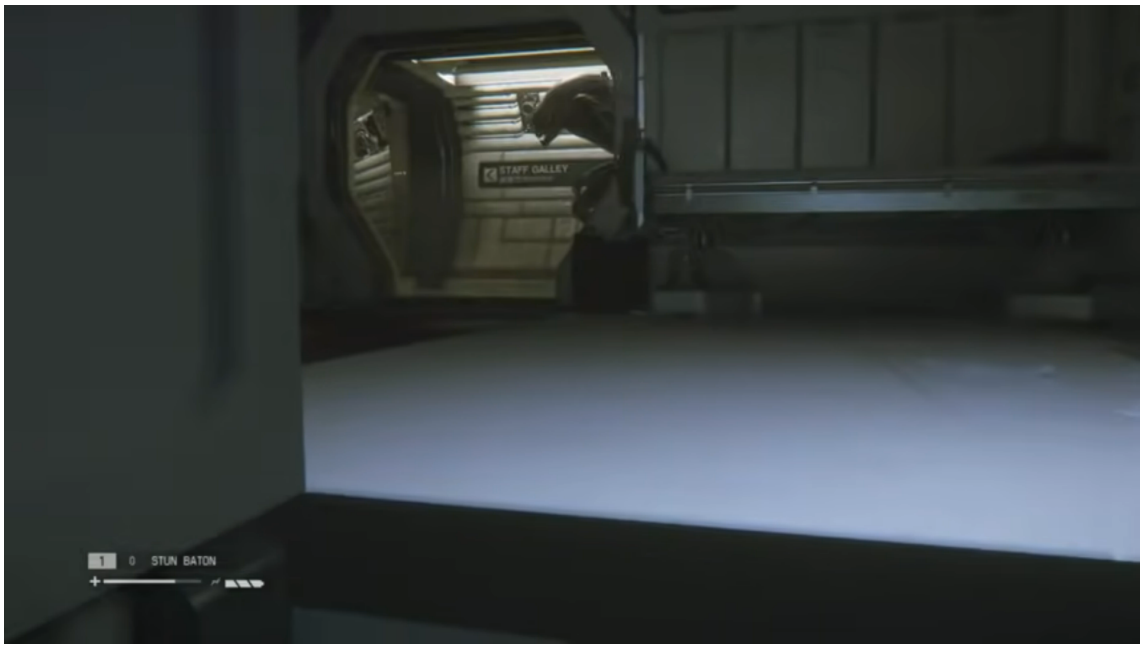


Figure 3.6: The game *Alien Isolation* (Creative Assembly, 2014). The player is hiding under a table in a medical bay. The alien, searching for them, is about to enter the room. When played with microphone input enabled, any sound the player makes at such a moment is liable to lead to their discovery.[17]

(Crystal Dynamics, 2014) will only say "reload" if using the reload mechanic is strategically desirable at that moment.

On the other hand, if the effect of a given mechanic actuation runs counter to a player's goals, it will be avoided, as in the stealth horror game *Alien Isolation* (Creative Assembly, 2014) where noise detected from your microphone can alert the alien in the game to your presence (Figure 3.6). Players attempt (not always successfully) to be silent when the alien is present:

> "I was hiding from the alien (successfully, I should add) under a desk in the hospital. I have sound detection turned on because I think that's awesome. Anyway, I'm holding my breath (in real life) and sitting super still when my sleeping Great Dane wakes up and decided to bite the hell out of his loudest squeak toy."[16]

The optimal input is not always the one intended by the game design or expected by other players. The game *Mario Party 6: Fruit Talktail* (Hudson Soft, 2005) is an asymmetric party game for the Nintendo GameCube where one player using a microphone to identify hexagonal platforms in a grid (each is labelled with a picture of a fruit) (Figure

[16]Reddit post by ChanceTheDog https://www.reddit.com/r/LV426/comments/2injw3/anyone_using_the_kinect_for_alien_isolation/ Accessed 2022-02-11
[17]Screenshot from *Cerebrophage* https://youtu.be/fQXxzf9jJHs

Figure 3.7: The game *Fruit Talktail* from *Mario Party 6* (Hudson Soft, 2005). One player names a fruit and the other players need to jump onto that tile before the other tiles fall away. The fruit 'grapes' has just been named and this has been detected by the game's speech recognition. The other tiles have dropped slightly. In a moment they will fall away entirely.[19]

3.7). The other players must jump their characters on to the selected platforms before the other platforms drop away. Hearing what their opponent has said gives them a chance to react more quickly. However, the forgiving implementation of the speech recognition allows an alternative strategy, described in the quote below. This strategy gives the speaker an advantage, motivating its use (note however that some groups consider this a violation of Fair Play Norms, introduced later).

> "Anywho–one time I was playing it with my friends, and the mic-wielder decides he wants to make the game a little more interesting so he shouts random combinations of fruit at the television. You know, like "Bananamelon." This made it really hard for everyone else to know where to run since there was no way to predict which platforms would fall!"[18]

Effort Minimisation Speech actuations are typically experienced to be more effortful than using standard input methods such as a controller, though different possible speech actuations vary with regards to the degree of effort they require (Igarashi & Hughes, 2001;

[18]Comment by Hail-NekoYasha on their webcomic https://www.deviantart.com/hail-nekoyasha/art/Fruit-Talktail-23365971 Accessed 2022-02-11

[19]Screenshot from Nintendo https://youtu.be/4rZgF4JBtPI

Sra et al., 2018). For example, singing is commonly perceived to be more effortful than humming, and talking as more effortful than noisemaking. Effort Minimisation motivates against effortful actuations. For instance, while 'noisemaking' mechanics triggered by microphone intensity can be actuated equivalently by humming or talking in full sentences, players almost always prefer the low-effort method: humming, blowing, or vowel sounds.

Non-speech alternatives can be seen as the natural or inevitable choice in practice due to their lower effort. A player of *Fruit Talktail* responding to a player using speech input, commented as if it should have been obvious: "if you press somethings,it will give you a list, no need to talk"[20]. I found myself defaulting to the use of keyboard and mouse while I was playing through every voice-controlled game on Kongregate[21] and NewGrounds[22]: their voice interfaces were largely similar and once I had a feel for them, it was easier to explore the rest of the game with keyboard and mouse inputs.

On the other hand, there is no reason why, if speech were *easier* than using a controller, it could not be preferred for the same reason. Effort Minimisation may work in favour of speech in *There Came an Echo* (Iridium Studios, 2015):

> Strategy is also where voice commands really prove their worth in There Came an Echo. Wishnov [Lead designer] agreed with the idea that for simple commands like 'yes' and 'open fire,' nothing is faster than pressing a button. But what voice control allows you to do is cue up commands and string together series of directions. [...] while it's totally possible to do it on a controller, it's arguably easier to pull off and more natural with voice — not to mention that it makes you feel like a bad-ass military strategist. (S. Sarkar, 2015, para. 7)

However, Effort Minimisation does not always have a strong influence when the actuation is motivated by more than just Optimality. In multiplayer games, the requirement to maintain Appropriateness (introduced below) limits the extent to which Effort Minimisation is possible. For example, in *Dixit* (Roubira, 2008), players will go to (reasonable, though perhaps not extreme) efforts to provide a good (optimal, performative) spoken clue. Performance (introduced below) is another motive that can motivate effort not necessary from the viewpoint of Optimality, such as in singing games. Yet even when the high-effort actuation is motivated, players can fall back on lower-effort actuations in situations when the level of effort is too great or unachievable. For example, while players of *Rock Band 4*'s (Harmonix, 2015) singing mode generally prefer to sing to the game (see Performance

[20] Comment by long42950 on video https://youtu.be/nLTGrbVsQww
[21] www.kongregate.com
[22] www.newgrounds.com

as a motivation for Actuation), the difficulty of successfully controlling the game in this way leads some to switch to humming instead:

> "It still kinda goes to the humming way on certain songs, i tried to do good 4 u last night and shit man, that song has some aggresive tone changes that require you to humm the song, also because the song is kinda fast, so it's a bit difficult to actually sing each part"[23]

Together with Optimality, Effort Minimisation (in the absence of other motivations) leads to a convergence on the minimum-effort actuation required required to successfully control the game in the most optimal way possible. That is, the player converges on Efficient inputs.

3.3.1.3 Appropriateness

Actuations are always performed in a sociomaterial context and are interpreted normatively by players and other response-present participants within the context of gameplay and of the wider situation (Allison et al., 2019). Different contexts both alter the sociomaterial norms (even playing with different boardgames groups), and the extent to which the player is concerned by them (playing alone in an empty house vs. playing in the presence of others). There were two primary sub-dimensions I found for Appropriateness: *Fair Play Norms* and *Situational Norms*.

Fair Play Norms Speech actuations can be influenced by normative expectations about what sort of actuations are or are not 'fair'. It is a widespread norm of (contemporary, competitive) gameplay that players should not gain an unfair advantage in the game or put another player at an unfair disadvantage (Huizinga & Hull, 1949). Further, often there is a perceived 'correct' way of playing the game, and not playing in this way is seen as 'playing wrong' (though is often more optimal) (Sniderman, 1999). These norms differ between games and gameplaying situations and, as in my experience of larp in particular, negotiating what the Fair Play Norms are can be the subject of many (sometimes long and tedious) discussions (Bergström, 2010; Hughes, 2005). Fair Play Norms are informed by wider cultural norms, but they are affected by the Rule-Based Constraints which establish what actuations should have effect, as well as players' beliefs and expectations about the particular game.

[23] Comment by SkullMan140 on the post https://www.reddit.com/r/Rockband/comments/pwcmkq/i_know_this_is_a_weird_gripe_but_does_anybody/ retrieved 24 November 2021

In multiplayer games, players largely confine themselves to rule-abiding speech actuations. 'Playing by the rules' is one common expectation about fairness, and violating this in a multiplayer game is widely (though not always) condemned as cheating. Take the cooperative card game *Hanabi* (Bauza, 2010), an unusual game because players hold their cards backwards so that only *other* players can see them: players are constrained in what information they are allowed to share with one another. Players obey these rules despite their inefficiency as to do otherwise would violate the spirit of the game. Similarly, 'playing to win', i.e. trying to do one's best to achieve the goals as stated in the rules (or aiming for Optimality), is a common Fair Play Norm (Bergström, 2010).

Fair Play Norms extend to unspoken agreements about appropriate play (Sniderman, 1999). A common example I observed is that players should speak in an audible way. Larps often allow players to make 'calls' (e.g. damage calls), requiring other players to respond appropriately (e.g. losing hitpoints)[24]. Players expect one another to only make calls they are entitled to make within the rules of the game (playing by the rules). Moreover, calls are expected to be made clearly and not at an excessive rate (more optimal, but seen as not in the spirit of the game). Yet it often happens that – especially in intense combat – these norms are violated, sometimes resulting in bad feeling between the players.

Player expectations about the 'correct' way to play the game can have some influence even in a single player game. Two quotes from a discussion of the game *Rock Band 4* (Harmonix, 2015) (Figure 3.8) show how humming, instead of singing, is perceived by some (though not all) players as violating Fair Play Norms. A player of the *Rock Band* series says:

> "While it may be accepted by the game and register when a person hums, it really does feel against the spirit of the game as a whole."[25]

However, it's notable that the following commenter is still happy to resort to humming if necessary to perform well at the game (Optimality), despite perceiving it as not the 'correct' way to play. This example shows that in this case Fair Play Norms exist, but in this case of solo play, they have limited effect.

> "But I always feel a little dirty humming or slurring the words or using the harmonics

[24]Here I have drawn on my personal experience of playing and running larps as part of a University of York student society and playing *Empire*, a UK festival larp.

[25]Comment by *itsiank* on Reddit post https://www.reddit.com/r/Rockband/comments/5tor82/far_from_perfect_but_way_more_fun_than_humming/ Retrieved 24 November

Figure 3.8: The game *Rock Band 4* (Harmonix, 2015) being played in vocal mode with three player harmonies. A karaoke-like game where players sing the words scrolling past on the screen. An line indication of the target pitch for each player is shown. Only the correct pitch is required to score points.[27]

> in freestyle, I don't do it unless it's like the fifth try at nailing a song."[26]

Situational Norms Actuations are constrained by wishing to avoid violating the sociomaterial norms of the surrounding situation within which the gameplay is situated. Gameplay doesn't take place in a vacuum: there can be people who will overhear or be disturbed by actuation. Moreover, other players are also friends, colleagues, strangers, etc. in front of whom we are reluctant to act inappropriately.

The very act of speaking to a game can threaten Situational Norms. Research in HCI has found that the use of voice interfaces in general have a social acceptance barrier to adoption (Al Hashimi, 2007; Rico & Brewster, 2010). The use of voice-as-sound can be annoying for those nearby (Igarashi & Hughes, 2001). Carter et al. (2015) note that players of *Tomb Raider: Definitive Edition* (Crystal Dynamics, 2014) reported feeling like they were disturbing other people nearby when using speech recognition, and that it was 'uncomfortable' and 'embarrassing'. Similarly, Allison et al. (2019), found that players frequently felt uncomfortable if the vocalisations needed for the game sounded odd

[26] Comment by an unknown user on Reddit post https://www.reddit.com/r/Rockband/comments/5tor82/far_from_perfect_but_way_more_fun_than_humming/Retrieved24November
[27] Screenshot from *PMS Lammy* https://youtu.be/9VMhgBRInVk

in the wider social situation, imagining what some (hypothetical) observer might think. In particular, players expected that they will be interpreted as attempts at meaningful communication, rather than mere game controls. One player, responding to advice on how to gain confidence speaking in an online multiplayer game said:

> "I'll try the speaking out thing without speaking into the mic later, really like this idea. Pretty sure i'll get some strange looks at home but oh well haha." [28]

Players are hesitant to violate Situational Norms and may need some degree of license to do so (Deterding, 2017). One such game that licenses the violation of Situational Norms is the card game *Cards Against Humanity* (Dillon et al., 2011), which involves combining sentence fragments to produce (often offensive) statements, which then must be read aloud (one is chosen as the best). Saying intentionally offensive things is usually seen as inappropriate in most social situations, but when you are all playing *Cards Against Humanity*, such norms are intentionally violated. In my own internal experience playing such games I have observed the conscious process of weighting up the conflict between Fair Play Norms (this is something I am supposed to say in the game), and Situational Norms (but it might be embarrassing to myself or perceived as inappropriate).

The balance between Fair Play Norms and Situational Norms is a contextualised decision about Appropriateness: given that I am in *this* situation, and I am playing *this* game (and *this* is the game state, etc.), what would be the maximally appropriate actuation for me to choose?

It seems likely that Appropriateness, as an influence on Actuation, is moderated by whether other people are co-present in the social situation. Appropriateness seems to obtain only when others are observing or playing. If there are people co-present but not playing, Situational Norms obtain but Fair Play Norms might not.

3.3.1.4 Performance

Actuations are not only enacted, but frequently also *performed*. They aim not only at effecting some change in the game, but doing so in a stylised manner: a performance either to themselves or to the other players. Hämäläinen et al. (2004) remarks how part of the appeal of voice interaction in games is the social spectacle it creates. Light et al. (2011,

[28] Comment by *shimmybee* on Reddit post https://www.reddit.com/r/OverwatchUniversity/comments/5v5hxl/how_to_overcome_my_big_fear_of_speaking_over_the/

p. 6) observed that "SingStar encouraged players to 'perform' in a variety of ways such as dressing-up or ad-libbing lyrics with localised or simply obscene variations of the original".

> "Often a dressing up box would be brought out at certain gatherings where wigs, gowns, feather boas and other items of clothing would be worn by singers and the audience. The dressing up box would usually lead to comedy performances whereby any desire to win by points or hit the right notes was overridden by the desire to make the performance as funny as possible." (Light et al., 2011, p. 6)

The following player of *Rock Band* was concerned that were the scoring system of the game made more strict (and challenging), Efficiency (Optimality/Effort trade-off) would overwhelm the Performance motive, making humming, rather than singing, the preferred choice and thus the game less fun overall:

> I think the reason the scoring system is the way it is really just boils down to keeping vocals fun. If the game was awarding points based on your performance on each individual note (whether you sang its entirety and with perfect pitch) the optimal high-score strategy becomes either humming the whole song to switch between notes quicker or trying to sing as robotically as humanly possible, both of which kinda take away the fun and the point of playing vocals IMO. With the way it is right now, there is still a lot of room to belt out your own cover of a song and still get 100% and a high score because there is a bit of leniency in how much of the song you need to sing correctly, and that keeps it fun.[29]

Calls in larp are explicitly not to be interpreted as something a character *says* in the world; they 'represent' the *effect* that they are calling or the event (spell, explosion, etc.) that triggered them. Yet my experience is that players often imbue the call with the character of the creature they are playing[30]. Similarly, in *There Came an Echo*, a reviewer remarks:

> "It's one thing to say open fire, it's completely another to bring out your best general voice and yell out 'Kill them all!"' (Bidwill, 2015, para. 6)

The three dimensions presented here: Efficiency, Appropriateness, and Performance, each contribute to a player's selection of an actuation. That is, collectively, they affect how (or if) a player chooses to speak to trigger a change in the game state. They may or may not be aligned. In many – perhaps most – games, players need to make ambiguous choices about how to maximally satisfy these conflicting demands. For example, a (particular)

[29] Comment by ihatecompvir on https://www.reddit.com/r/Rockband/comments/pwcmkq/i_know_this_is_a_weird_gripe_but_does_anybody/ Retrieved 24 November 2021

[30] I recall, at the University of York campus larp *Shattered Legacies*' summer event, a particularly nerve-wracking combat against the spear-wielding demon named Dragon, a significant antagonist in the game. Their terrifying attack calls were delivered in a shrieking cackle, very much adding to the atmosphere of the game, as our party of ten or so struggled desperately to defeat them.

performance may or may not be Appropriate: being seen as gloating, taunting, smug, or shocking in one's actuation might violate Situational Norms. Performances may be more or less Efficient. Other players might score the quality of the performance as in *Snake Oil* (Ochs, 2010) and *The Big Idea* (Ernest, 2011), aligning Performance with Optimality. If there is no such alignment, Performance may or may not be a strong enough motive to overcome Effort Minimisation. Finally, Efficiency might not align with Appropriateness: playing to win is not always playing fair.

We seem to consciously experience choices like these as a kind of psychological tension. Players reported sharing the feeling, which often arises in a cooperative game like *Hanabi* (Bauza, 2010), "I really want to say x but I can't, so I'll say y instead". There is a feeling of holding oneself back from doing what you *want* to do, because of the constraints you are imposing upon yourself. This is merely a suggestion, however, as the phenomenological experience of data provision is beyond the scope of this research.

3.3.2 Communication

Communication is a second motivation identified for speech, related not to the effect it achieves directly (as in Actuation), but to the information it conveys. This information can be valued in various ways, including for its utility in the game.

Imagine we are playing the word game *Taboo* (Hersch, 1989). *Taboo* is a multi-player team game where one player on a team is given a word that they must somehow get their teammates to guess, while avoiding a use of 'taboo' words. You might be given the word 'cat' to describe, but be forbidden from using the terms 'animal', 'pet', 'kitten', or, of course, 'cat'. I, as the other player, am unable to see the card myself, but I can listen to clues that you give and make guesses as to what I think the secret word might be. You might say 'cute furry mammal', to which I guess 'rabbit'. You indicate my guess is incorrect and try and give me some more hints. 'Not a dog', you say, to which I immediately guess 'cat', and we score points for our team.

What is going on here? You have been given some information by the game: the secret that is written on the card for only you to see. We agree that within the rules of the game you cannot simply show me the card. That would remove the enjoyable challenge of the game. The information is no use just staying in your head, however. In order to score points for our team, you need it to be in my head also.

Taboo's rules are designed to construct a *Perceived Information Asymmetry*. This

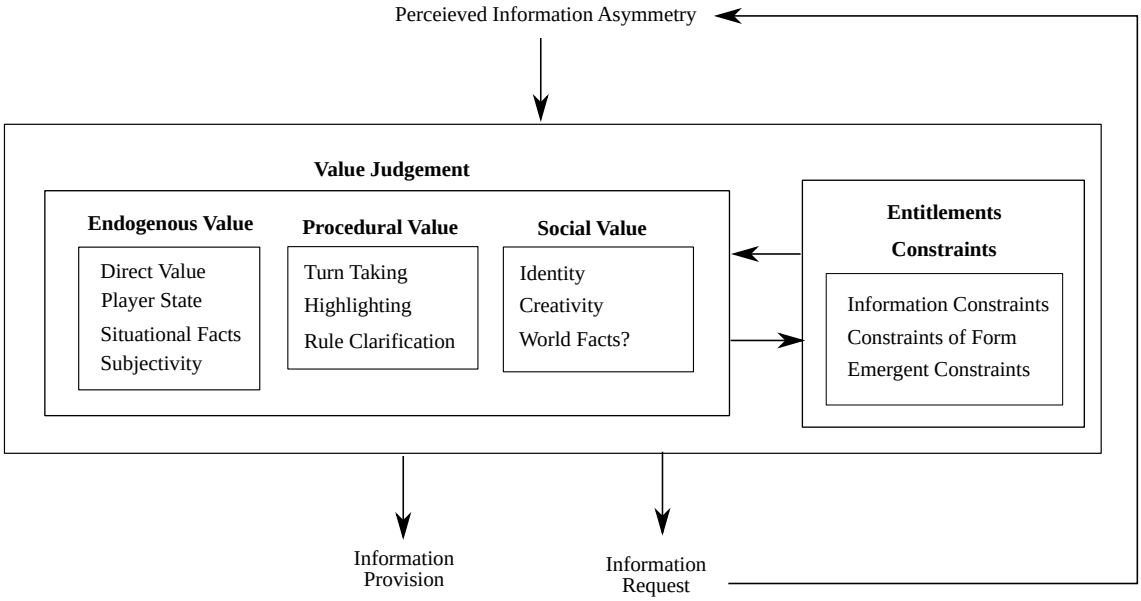


Figure 3.9: The Information Need sub-model. Perceived Information Asymmetries are evaluated by the interaction of Value Judgements for the information combined with the Entitlements and Constraints of the game to determine the Information Provision or Information Request that is spoken.

asymmetry is resolved, and points are scored when what I say matches what is on your card. The challenge and interest created by *Taboo* relates to how it impedes the process of resolving the asymmetry. Common words and associations that you would otherwise use are forbidden by the game rules. Such games take us out of *communion* and challenge us to find ways to reattain this communion by (restricted) *communication*. Herein lies the fun.

Taboo could be understood through the lens of the Actuation model with 'hint giving' being a main mechanic of the game that transforms the game state. Communication gives us another lens to understand this sort of situation. Communication, with its focus on information rather than action draws largely orthogonal distinctions to those within the Actuation model.

Perceived Information Asymmetry A perceived information asymmetry is an awareness arising during gameplay that some, but not all, players know a particular piece of information (i.e. an awareness of a particular information asymmetry). Two kinds of information asymmetry may be perceived. Either the speaker has some information that the other person is believed not to have, in which case the ultimate result of their decision may be an *Information Provision*, or the speaker does not have information that they believe

they other person does have, in which case the result may be an *Information Request*. Such an Information Request is likely to trigger another player to perceive an information asymmetry, to which they may choose to respond with Information Provision.

These perceived information asymmetries arise as a result of the dynamics of situated, moment-by-moment play. Information asymmetries do not necessarily last very long, perhaps only an instant in a fast paced game like the team first person shooter *Overwatch* (Blizzard Entertainment, 2016). They may be resolved (the information is shared and there is a resulting perceived symmetry of information), or forgotten. Alternatively, they may remain until they can be satisfied. Players often engage in contextualisation of events within the game retrospectively ("What were you hinting at there?", "Why did you take that move?"). At the end of a competitive game, players might ask about the strategy their opponent was performing, the (hidden) cards they had in hand, etc.

Perceived Information Asymmetries might arise as a result of the formal information properties of the game design. In particular, games that involve hidden information by definition involve information asymmetries. In a game like *Taboo* (Hersch, 1989), information is presented on a card to one player and hidden from the others. However, the game need not incorporate hidden information for a Perceived Information Asymmetry to arise. For instance, the limitations of human players create other information asymmetries. In board games, there is usually the expectation that information that is not hidden by the game rules should be accessible to all players. While the contents of a player's hand is generally hidden information, the arrangement of tokens placed on the board is usually public information. Yet in practice, players often cannot clearly perceive everything happening on the board. For example, this might arise where one player performs a rapid succession of public actions (as is the case in *Dominion* (Vaccarino, 2008), discussed below). Each *should* be visible to the other players to inspect, but in practice this is impossible and thus an information asymmetry (what were the actions just performed) is created.

However, the perception of an information asymmetry is not sufficient by itself to explain speech. The game of *Taboo* constructs several other information asymmetries that do not elicit speech. Only one player, for example, knows what the 'taboo' words are in any given round. Only one player can see the colour of the front of the card. Yet, on the whole, these sorts of topics do not arise during the game of *Taboo*. Speech only happens when the information satisfies an interaction between *Value Judgements* and the *Constraints* and *Entitlements* of the game.

Value Judgement The deck-building card game *Dominion* is not about communication. However its official rules contains the following line (emphasis added):

> "To play an Action, the player takes an Action card from his hand and lays it face-up in his play area. *He announces which card he is playing* and follows the instructions written on that card..." (Vaccarino, 2008, rulebook p. 5)

This is a simple form of speech rule: in a certain situation in the game, certain information (the card being played) must be *announced*, apparently whether or not anyone is interested in it. This injunction takes up half a sentence half way down page 4 of an 8 page rulebook. Is the existence of such a rule sufficient to elicit speech? By itself, it seems unlikely.

Dominion is one of the board games I have most experience in playing, and on the whole exactly what is described above occurs: players narrate the action cards they play. However, the existence of the rule does not seem to be what motivates the speech. On occasions this rule is, by unspoken consent, overlooked. The occasions when this rules is not followed seem to be when the following are true: 1) there is another conversation occurring which this rule would disrupt, and 2) the player's turn is relatively simple. On turns with a complex sequence of actions however, the rule is always followed. It is on such complex turns that the information being narrated can less easily be observed from the cards on the table: there is more value to its announcement. In other words, the speech is not determined by the rule, but a *value judgement* about the information at that moment of play. Indeed, rather than it being the *rule* that leads to the speech, the opposite is closer to the case: the value of the speech has been formalised as a rule.

The value of the information in *Dominion* is not the same type of value as in *Taboo*, where successfully communicating the information is worth points, but it is valuable to the other players for understanding what cards their opponents have and for keeping track of what is happening in a relatively complex game. Three types of value that information asymmetries have were identified. First, *Endogenous* Information Asymmetries are cases where the information has a value within the game itself (e.g. an opponent's cards in *Poker*, the secret in a game of *Taboo*). Second, *Procedural* Information Asymmetries occur when the information has procedural value (e.g. whose turn is it, describing what sequence of cards are being played in a turn of *Dominion*). Third, *Social* Information Asymmetries, where the information has value to the players as actors in the game as a social interaction. Information may have more than one type of value.

3.3.3 Entitlements

In the case of *Dominion*, described above, the value of information relates not to the speaker (who already knows what cards they have played), but to the other players. In fact, as it reveals to the other players what cards the speaker has in their deck (hidden information), it against the speaker's strategic interest to share this. This is where the rule quoted above does have an effect. The fact that a player must *announce* their actions equivalently means that the other players are entitled to that information. Thus it is the value of the information *to the entitled player* which is judged. It follows that even though a player might be entitled to information, if they (are perceived to) have no value for it, it need not lead to speech, hence why on occasion the speech rule in *Dominion* can be overlooked.

3.3.4 Constraints

In *Taboo*, the most valuable information is the secret hidden on the card. By one's teammate learning (and announcing) this, a team scores points. However, the rules of the game impose *Constraints* on what information may be shared. Certain words are forbidden (including the secret itself). There is a negotiation between the values of different pieces of information and the constraints defining to what information may be shared. While one clue might convey the most valuable information, an alternative, less informative clue might be the best that satisfies the constraints.

Information Constraints Some constraints limit what information may be communicated, but not necessarily how that information is communicated. For example, *Hanabi* (Bauza, 2010) is a cooperative game where players are not able to see their own cards, even though this would be useful to them. *Hanabi* includes the following rules, constraining players from freely resolving this information asymmetry by limiting what information may be conveyed:

> "Two types of information can be given and the player given the information chooses only one type of give.
> A. Information about one specific COLOR (and only one)
> B. Information about one specific VALUE (and only one)
> IMPORTANT: The player must give complete information. If a player has two green cards, the informer cannot only point to one of them, he must point to BOTH green cards." (Bauza, 2010, rule booklet)

A limitation on what information may be communicated may extend to all information

(or at least, this is the logical ideal toward which the rules are written). The cooperative real-time board game *Magic Maze* (Lapp, 2017) does not include communication mechanics, but it does include a rule *preventing* players from communicating. In *Magic Maze*, players must work together to move pawns around a maze before a timer runs out. As each player is responsible for all movement in a particular direction (e.g. north or east), it is essential for players to agree on where they are trying to move each figure in order to get anywhere. However, the constraint the game imposes on speech means that most of the game is conducted in silence. When this constraint is not in force (e.g. after a player has flipped the sand timer but before any further moves have been taken) there can suddenly be a burst of speech to strategise the next period of gameplay.

Constraints of Form Some constraints dictate the form in which information must be communicated. For example, in the cooperative game *Cranium Hoopla* (Alexander & Tait, 2002), there is a secret that one player must communicate via hints. In one mode of the game, the player who knows the secret must give a hint in the form of a comparison (e.g. "it is bigger than *blank* but smaller than *blank*"). Similar constraints of form exist in games such as *20 Questions*, where only yes or no questions may be asked (and the player is both entitled and constrained to the response 'yes' or 'no').

Emergent Constraints The practicalities of the situation impose constraints. These constraints might emerge from time pressure, as during intense combat in *Overwatch* (Blizzard Entertainment, 2016), or the presence of noise and other conversation, as is fairly common when playing tabletop roleplaying games, such as *Dungeons and Dragons* (Wizards of the Coast, 2014), with many players all needing to talk to a single game master. They might also arise from technical difficulties from the implementation of voice chat in digital games (Wadley et al., 2015).

Together, Entitlements and Constraints interact with the Value Judgement in a dynamic process of considering, evaluating, and discarding possible utterances. While satisfying the Entitlements and Constraints, players seek to maximise the value of the information they communicate when they speak. This value comes in the form of *Endogenous Value*, *Procedural Value*, and *Social Value*.

3.3.5 Endogenous Value

Players typically engage in goal-directed action in games. In other words, they want to win, or achieve whatever alternative goals they have set themselves. Information is frequently valuable for this purpose, be it the locations of enemies in *Overwatch* (Blizzard Entertainment, 2016), the contents of one's hand in *Hanabi* (Bauza, 2010), or the action intentions of one's team-mates in *Magic Maze* (Lapp, 2017). This information is valuable *within the game*, hence I will label it *Endogenous Value* In game design, the term Endogenous value is most often used for tokens that a player might possess (Schell, 2009). For example, in a roleplaying game, different items (weapons, armour, etc.) are more or less valuable to a given player, based on their in-game effect(iveness). Here I extend this usage to information. Information has (positive or negative) endogenous value when it contributes or harms a player's ability to win (or achieve their goals in) the game. Those who want to win (or otherwise perform well) will want to communicate information (in the form of an Information Request, or an Information Provision) because of its Endogenous Value. As the Endogenous Value of opposing players is usually harmful to one's own game goals, it also follows that in such situations players will want to avoid communicating information with Endogenous Value to an opponent.

Hanabi is a multi-player cooperative card game in which the players attempt to build sets of cards. Each set is of a different colour, and numbers must be played in sequence. In this respect it is like a multi-player game of patience. However, in contrast to most other card games, players of *Hanabi* can see only *other* players' cards. The player in *Hanabi* can perform three actions. They can play a card, discard a card, or spend a token (of which there is a limited supply) to tell one player a single piece of information about their hand. In *Hanabi* players start with no information about their hand. Each piece of information they are given helps them to identify those cards they can play and which they can safely discard. Outside the game of *Hanabi*, this information has no value, but within the game it is a precious limited resource. Players are careful not to squander any chance to give information, and carefully consider what best to say to gain maximal benefit. Information in *Hanabi* has high Endogenous Value.

Four types of Endogenous Value for information were identified. *Direct Value*, *Objectivity*, *Subjectivity*, and *Situational Facts*.

Table 3.2: Dimensions of Endogenous Value

Code	Description
Direct Value	Information, the possession of which is rewarded by the game
Player State	That sub-part of the formal game state that defines a player, including e.g. deck composition, loyalty cards
Subjectivity	Intentions, dispositions, and judgements of a player with regards to their options in the game
Situational Facts	Facts of game state independent of any player

Direct Value Some games directly reward knowledge of particular information. An archetypical example of this is a quiz. Each answer in a quiz has direct value by virtue of the score that is assigned to it. Where a team answers a quiz together (as in a pub quiz), team members who know the answer to a question naturally share this information with the person writing or submitting the answers if they might not know it (i.e. if there is a perceived information asymmetry). Yet very different games can also furnish information with direct value. For example, in the stealth action game *Deus Ex: Mankind Divided* (Eidos Montréal, 2016) (though it does not have speech mechanics), knowing a password might allow you to deactivate a security camera or acquire useful items.

Many multiplayer games based around direct-value information restrict what information can be shared between players. Information that has direct value but cannot be freely communicated might be called a *secret*. Often communicating a secret is the goal of the game, such as discovering the hidden word in *Taboo* (Hersch, 1989), *Cranium Hoopla* (Alexander & Tait, 2002), or *Almaniac* ("ALMANiAC", 1990). For instance, because *Taboo* is a team game, it is in the interest of the player who possesses the secret to communicate that information to their team-mates (though there are Constraints on how this can occur).

Player State Information in games frequently has value because it lets a player model hidden or hard-to-observe aspects of the game state relating to one or more players in the game.

For example, in *Hanabi* a player starts off not knowing what cards they are holding in their hand. Knowing the cards you hold allows you make effective use of your turns (and avoid wasting valuable cards). Such information, of the objective facts of one of the players (including oneself) is frequently valuable in games. In many card games and board games, knowing the contents of another players' hand is strategically useful. In competitive games,

rules sometimes force players to reveal information about themselves to other players, for example in *Go Fish*, players must reveal when they do not have any of the requested card (by saying 'go fish'). In *Mao*, players must announce when they are on their last card. In the cooperative game *Forbidden Island* (Leacock, 2010), players share with each other the actions that they can perform in the game. In some larp call systems[31] players must respond to a call with 'No Effect' if it does not effect them. This allows their attacker to learn information of strategic relevance. Players of *Dungeons and Dragons* (Wizards of the Coast, 2014) frequently ask one another for information (such as their current hitpoints) before deciding how to act (for example, whether to cast a healing spell).

Hidden identity games such as *The CopyCat* (Wu, 2016) assign each player a role which is hidden from the other players. In *The CopyCat*, players have to take it in turns to describe a particular topic. Such speech has value because it (in principle) lets other players distinguish who is human (to whom the topic has been revealed), and who is a copycat (who is pretending to be human, but does not know the topic).

Subjectivity Information can have value in that it reveals the (subjective) opinions and intentions of other players, allowing players to better predict or model the actions of others. Such potential actions have a knock-on impact one's performance in the game, such as by knowing what card a clue is likely to refer to in *Dixit* (Roubira, 2008).

In the game *Cards Against Humanity* (Dillon et al., 2011), one player makes a subjective decision about other players' actions (in this case, the cards they have played) which determines the points that player's score. To play the game to win, a player benefits from information that allows them to better interact with such subjective valuation mechanics. Because of this, the player reading out the options and announcing which one they think is best has Endogenous Value.

Many cooperative games, such as *Forbidden Island* (Leacock, 2010) allow players to coordinate their actions. For this they must share information regarding their abilities (Player State), but also their intentions towards particular actions (Subjectivity).

In *Mario Party 6: Fruit Talktail* (Hudson Soft, 2005), one player names a type of block, and all other players must jump onto that type of block before the others drop away (performed automatically by the computer using speech recognition). The signal requires or triggers reaction from the other players. Hearing this information has value in allowing

[31]The particular example here is from the University of York campus larp system *Shattered Legacies*

a player to react. Due to limitations in the speech recognition in Mario Party 6, non-target phrases are often interpreted as triggers. Players often see this as cheating, as it prevents them from receiving the information that they value (and consider themselves Entitled to).

Situational Facts Games frequently take place within a virtual world. Information relating to the situation within this world will often guide player actions towards their goals. For example, in *Overwatch* (Blizzard Entertainment, 2016), a first-person shooter, each player can only observe a small part of the game world. To effectively coordinate with team mates, players may share information relating to the situation of the world around them, such as the positions and actions of opponents. Players of *Dungeons and Dragons* listen to the Dungeon Master describe the situation they find themselves in, and remind each other of information they may have forgotten or not been present for. Drachen and Smith (2008) identified this sort of 'Environment-world description' as the most common type of speech in tabletop roleplaying games.

3.3.6 Procedural Value

Games are often procedurally complex interactions. As players, we desire for these interactions to be successful. In other words, whether or not we can play well, we at least want to *play the game*. Players will engage in speech if there is a perceived information asymmetry that has procedural value, leading to either an information request (e.g. "whose turn is it?") or an information provision (e.g. "it's your turn"). We value information that helps us achieve this for its *Procedural Value*.

For example, information about the state of gameplay is partially or completely hidden from some players, and gameplay is frequently too complex to be easily tracked. Players need to be reminded of things they should know, rules they need to follow, or when to take their turn. This is particularly in evidence when playing with players new to a game or with young children[32]. This information is so essential to the play of most games that constraining it can be the defining feature of *Mao*, a card game whose rules prevent almost all procedural information from being shared, including explaining the rules (especially to new players) and asking whose turn it is. This makes *Mao* a very particular gaming

[32]From my experience of school homework (4-5 years old), educational games to be played with children exploit this. For example, a bingo-style game teaching phonics requires both graphemes to be read aloud and the written form of phonemes to be located. A typical interaction sees the child asking for procedural information to continue play. Child: "What is this sound?", Adult: "It's a /d/"

Table 3.3: Procedural Value codes. These are ways in which information can have Procedural Value in a game

Code	Description
Turn Taking	Information relating to who should act (and what they should do) to maintain procedural flow
Highlighting	Voluntarily speaking information that is otherwise shared to make it more 'visible'
Rule Clarification	Speech to ask or explain how the game should be played

experience.

Players value procedural information because they want to be participants in a successful and functional gaming situation and avoid delay and dysfunction. Players want to feel like they are a competent player (even if they're not good at *winning* the game). Part of engaging in the game together is a joint interest in making the game run, even if you are losing. Being intentionally disruptive to the procedural flow of the game is considered bad sportspersonship. Providing procedural information (being helpful) may be seen as socially desirable.

Three sub-codes were identified for Procedural Value: Turn Taking, Highlighting, and Rule Clarification.

Turn Taking Players are expected to act in various capacities during a game. In some games, they need only take their turn (though may need to be reminded of when this is), but in others they may need to react in specific ways outside of the expected order. To address this, turn taking language naturally arises in most multiplayer games with some kind of alternation of who is supposed to be acting. Statements such as "It's your turn", or "Whose turn is it?" are frequent. The underlying perceived information asymmetry addressed is who is supposed to be doing what for the procedural flow of the game to continue, which may be completely orthogonal to what is happening *within* the game at that moment.

In the tabletop roleplaying game *Dungeons and Dragons*, players may take actions (e.g. cast a magic spell) to which other player must immediately respond (e.g. rolling a dice to avoid the effect, or making a note of a new bonus on their character sheet). It's common for a player to not only announce their action (an actuation which also reflects an information asymmetry with Endogenous Value), but how other players need to respond. For example "I cast *Fireball*, everyone needs to roll a reflex save".

Highlighting Some information is shared but is not easily accessible by all players of the game. This can lead to situations where players interrupt the game to ask for clarification on something they are entitled to know. To avoid this, players can Highlight information that is otherwise inaccessible by speaking it aloud.

Azul (Kiesling, 2017) is a competitive tile-placing board game with no explicit communication mechanics. All players have complete information about the tiles the other players have placed. Despite this, when a player places tiles in a way that will end the game at the end of the round, it is friendly (at least in the games I have played) to announce that you are ending the game. This avoids players taking moves that would not make sense if they knew what they were entitled to but might not otherwise notice. It allows the game to be played in a more relaxed way. Similarly, in *Dominion* (Vaccarino, 2008), the game ends when all the highest-valued victory cards (usually 'provinces') have been purchased from the centre. It is common to ask "How many provinces are left?" and for the other player to answer, honestly, if they are closer to the pile.

Dominion is a complex card game where one turn may consist of many actions. The rules include a requirement for players to describe their turns. This rule serves a purpose in the game, as it makes information about a player's deck visible to other players. To the extent that other players are interested in this information, there is an Endogenous information need. However, the sharing of hidden information is not the only reason for this rule. It allows the player acting to keep track of their turn and it allows other players to check that they are following the rules. Where neither of these needs are significant enough to disrupt other socialising, this requirement of turn description is sometimes overlooked.

Rule Clarification Most games (with the exception of e.g. *Mao*) and gameplaying groups allow players to clarify the rules during play, especially for novice players. Knowledge of the rules is information that often carries high procedural value as it is essential to playing the game. This happens frequently in games with extensive and complex rule sets such as the tabletop roleplaying game *Dungeons and Dragons*. Explaining such rules or common strategic considerations may be counterproductive with respect to Endogenous Value, but is normal that players should be helped to play the game (if not necessarily play the game *well*).

3.3.7 Social Value

As social encounters, games are full of information asymmetries, significant and insignificant, from what someone think of you as a coworker to what they had for breakfast. Games can make information of social or personal interest relevant in the context of gameplay and give (other) players an Entitlement to it. A prototypical example of such a game is *Truth or Dare*, wherein players take it in turns to either accept a dare or answer a question that is usually personal and sometimes embarrassing. Why are we interested in asking or answering these questions (and why do we sometimes wish we didn't have to)? Because of the (positive or negative) Social Value of the information to ourselves and to the other players.

While there are games where information is valued purely for its social value as in *Truth or Dare*, in many (perhaps most) cases social value is additive. In a game, the information we choose to communicate reflects not only on us as a player (e.g. what cards we have, what combinations we will prefer in *Cards Against Humanity* (Dillon et al., 2011)), but also on us as a person in the 'Social World' encompassing the game (Allison et al., 2019). Unsurprisingly, players have the tendency to 'elaborate': to go beyond what is strictly required for strategic play. Games involving Subjectivity-valued information lend themselves to such elaboration as asking or expressing game-relevant preferences, beliefs, strategies, etc. naturally aligns with sharing information about themselves as a person. While a player who explains their clue in *Dixit* (Roubira, 2008) may be giving information that might be strategically useful in a vague theoretical sense, it is the social value of that information that is much more immediate. Other times, the social sharing outweighs whatever minor cost to Endogenous Value that could be identified, like the player who reveals their strategy to an opponent ("I would get that card, but I want to try keeping my deck as small as possible." "Look, I drew my entire deck in a single hand, that's neat.')

Social Value can apply to information otherwise valued. However, I identified three ways game designs structure information asymmetries primarily around Social Value: through sharing information about *Identity*, by making relevant *Real-World Facts*, and by encouraging *Creativity*.

Identity Players often value sharing their identities with one another: how they (consciously or unconsciously) represent themselves in relation to other people and the world. They are both interested in requesting information about the identities of others and pre-

Table 3.4: Social Value codes. These are ways in which information can have Social Value in a game

Code	Description
Identity	Personal information and identity claims about the people you're playing with
Real-World Facts	Information about a topic of interest to the players external to gameplay
Creativity	Creative self-expression

senting their own identity claims. When understood as actions within the Social World, Endogenous and Procedural information requests and provision could be understood as entailing some kind of identity claim, even if it is just "I am a person who plays to win", or "I am a person who tries to be helpful". However, there are also games that are designed primarily for this social identity value.

The game *Truth or Dare* specifically licenses the asking and answering of personal questions. On their turn, a player has to choose whether to perform a dare or reveal a truth, based on the social contract that the other players will have to do likewise. There are various other icebreaker games based on the idea of identity sharing, such as *Never have I ever*, *Two Truths and a Lie*, *21 Questions*, and *Would you rather*.

Not all such games based around Identity are multiplayer. *Seaman* (Vivarium, 1999) is a single player game for the Sega Dreamcast wherein players raise a digital creature through successive stages of its life cycle in a digital vivarium (Figure 3.10). To encourage the Seaman to evolve, it is necessary to talk to it frequently and keep it happy. Once it learns to speak, it begins asking the player personal and philosophical questions about e.g. politics and religion. Answering these keeps the Seaman happy and lets the Seaman learn about you (within the limitations of the speech recognition system). The conversation has been described by reviewers as endearing (Provo, 2000). For players, answering these questions seems to be more social in motivation than strategic; a walkthrough author advises:

> "Also, be ready to undergo a psychological evaluation- Seaman will eventually want to know everything about you (and I do mean EVERYTHING), and he will offer some very interesting insight over the course of his lifetime. Try not to abuse this by lying to him, it can be much more interesting to tell the truth and perhaps learn things about yourself you hadn't realized before!" (Little, 2000, sec. 3, para. 10)

Identity construction is significant in the play of hidden identity games. Such games assign players different (hidden) formal goals depending on a (hidden) role they are allocated.

Figure 3.10: The game *Seaman* (Vivarium, 1999). The player talks with a fish with a human face and a sarcastic attitude. In this screenshot, Seaman has just said: *'...You humans are going to get lazier and less mobile and you'll forget how to deal with each other face to face. I think things could get pretty ugly down the road. Do* you *think the internet is dangerous?'*[33]

In play, this 'formal' identity merges and blends with social identity construction through roleplay. *Werewolf* (or *Mafia*) is a prototypical example: some players are secretly assigned to be werewolves and the rest villagers. The werewolves attempt to (secretly) eliminate the villagers, while each round the whole group votes to eliminate who they think is a werewolf. Here players present a fictional identity, often incorporating roleplay to do so. The werewolves want to be seen as villagers, and the villagers are concerned not to be mistaken for a werewolf. Players construct such fictional identities to an even greater extent in larp, where players establish enduring social relationships entirely in-character. Such identity construction can align with strategic play of the game: having in-character friends in larp is – especially in social games such as *Empire* – necessary to succeed in many of the political roles in the game. However, players also just spend time together in character as an end in itself.

Real-World Facts Players often make reference to information in the wider world, about which they may be interested (or wish to appear interested). This is most obvious within multiplayer quiz and trivia games. Team-based quizzes align social information of this sort

[33]Screenshot taken from a speedrun by *pile san* https://youtu.be/t6yIceuSBAg

with Endogenous Value, as players share their knowledge with each other to agree on an answer for their team. While structured in the mechanics of the team quiz, it is often the content or fiction of the game that is responsible for this kind of speech. Other games raise topics that players subsequently discuss, and in competitive games players frequently discuss topics raised, if only after it no longer can make a difference within the game[34].

Creativity Players want to express themselves creatively, make jokes, play with words, and have other people socially value our creations. Storytelling games such as *Black Rabbit Dice* (M. Anderson, 2014) provide a minimal game-like structure around this creativity, where as games such as *Cards Against Humanity* (Dillon et al., 2011) and *Snake Oil* (Ochs, 2010) align this creative motivation with Endogenous Value: competing creations are evaluated and the best scores points. Larps and tabletop roleplaying games can involve significant creative information requests and provision. Players of *Dungeons and Dragons* (Wizards of the Coast, 2014) prise describing the world and their actions in a creative and compelling way. Players in larp may engage in hours immersive creative discussion with very little content of direct Endogenous or Procedural Value, such as engaging in entirely specious philosophical or metaphysical debates.

3.4 General Discussion

I set out with the goal of explaining what motivates players to say what they say in a game; that is, what motivates players to take *particular data-providing actions* – within a linguistic case study. Two general motivations were identified: actuation of a mechanic, and the communication of information. These two models of motivation provide lenses through which we can analyse the provision of speech data: either instrumentally, as a speech act; or semantically, as a communication. These lenses afford different sorts of analysis and design.

Actuation, the model of motivation presented first, sees elicited speech as a lever to modify the game state, albeit a lever with perhaps many and complex dimensions: pitch,

[34]For example, in a round of the game *Dixit* (Roubira, 2008), which has stuck in my memory, one player gave the clue "One for the Howard League". The correct picture, as I recall it, was an image of the inside of a fairy-tale tower with an anthropomorphised egg escaping down a ladder propped at the window. As soon as the round was complete, we had one pressing question: "What is the Howard League again?" The reference was to the Howard League for Penal Reform, about which we engaged in a brief discussion before moving on to the next round.

volume, phonological correspondence to particular words, even language syntax and semantics as a format for constructing commands. Analytically, any such speech can be interpreted as an illocutionary act situated within both game (whether represented in data or in the minds of other players) and wider social situation. In terms of design, it places focus on the actuations available to the players and the effects of those actuations. In short: speech has to *do something*.

In contrast, Communication sees speech being afforded by perceived information asymmetries. The speech afforded is that which communicates the information necessary to resolve the asymmetry. Rather than being concerned with the effects of speech *per se*, this analysis focuses on the information to be shared and the situational value players ascribe to that information. Using this lens, designers should focus on how a particular gaming situation leads players to value certain kinds of information and ensuring the asymmetric distribution of that information; if this is achieved, the speech will take care of itself.

There are similarities between these models. Both centre around some notion of strategic action, corresponding to J. H. Smith's (2006) Rational Player Model. Around this, both models incorporate the influence of the wider social context. Further, each model can be understood in terms of the other. Communicative utterances might be seen as a kind of actuation that affects the (knowledge) state of other players in the game. Actuations might, more obliquely, be interpreted as informational instructions to address a (persistent) information asymmetry of the form "what do you want to do?". Significantly, both models are characterised by moment-to-moment rationality. Speech, whether an actuation or a communication, is a utility-maximising act, chosen at a moment of gameplay and weighted by a variety of factors.

They do not express all of the same concepts however. The actuation model (via Appropriateness) identifies the constraining effect of the wider social context on speech as an interaction. In contrast, the Communication model considers only the impact of the *information* being communicated. While it might explain why we wouldn't play *Cards Against Humanity* (Dillon et al., 2011) in church, it would not explain why we might feel limited playing *I Spy* with a noisy four-year-old in the quiet coach of a train.

It is plausible that with further theoretical development the two models could be combined into a single 'meta' model. However, I feel that it is preferable to have multiple models each providing a simple and relatively unnuanced perspective. Such lack of nuance is desirable for theoretical abstraction (e.g. to other types of data) and generating concrete

design guidance (Healy, 2017). Each model provides a lens, in the sense of Schell (2009). It is through juggling multiple such lenses that good design is best achieved.

3.4.1 A New Perspective on Motivation

The model presented here suggests a particular, novel perspective on motivation in games. Central to this is the idea that speech elicitation – or the motivation of behaviour in general – is determined by moments of gameplay. This is unlike popular molar theories of motivation, such as Self-Determination Theory (Deci & Ryan, 2000; Tyack & Mekler, 2020), which see motivation as a generalised property, perhaps at the level of a context (gameplay in general) or a situation (playing a game) (Vallerand & Ratelle, 2002) but not in practice more fine-grained than this. Such theories describe motivation for whole behaviour acts, such as playing a game. The type of theory developed here is geared towards moment-to-moment motivation for describing atomic actions within a game. In this way it is more akin to J. H. Smith's (2006) Rational Player Model, which specifies the particular action a player concerned only with maximising their score will take. While the account of motivation developed here is specific to speech elicitation and does not directly generalise more widely to motivations to produce other kinds of data, the general perspective on motivation developed here can be sketched in a general way to inform research more broadly.

To do this, we must first agree what we are motivating; the two sub-models here identify this differently: as actuations of mechanics, or communications of information. While I believe either could plausibly serve, I will adopt actuation based on the reasoning that, (most) games having mechanics, they will necessarily involve actuations of those mechanics, whereas many games would not naturally be thought of as having communication. Moreover this maps more closely to existing perspectives on motivation that consider behaviours rather than information. This decided, the essential perspective on motivation developed in this chapter is as follows:

1. The unit of behaviour that is motivated is the actuation. The concept of actuation need not be specific to a speech act. For example, imagine a webcam game for eliciting facial expressions where players effect changes in the game by means of a 'face act', or a typing game whereby players effect through a 'keyboard act'.

2. One's motivation for an actuation is inherently relative to the other possible ac-

tuations available at that moment in time. Motivation is therefore contextualised within a moment of gameplay. Generalised motivation is understood reductively as the aggregate of countless moments of motivation.

3. The evaluation of possible actuations integrates multiple factors. From both of the perspectives on player data provision presented in this chapter, we can see that both in-game concerns (e.g. Endogenous Values, Optimality) *and* out-of-game concerns (e.g. Procedural and Social Values, Appropriateness, Effort Minimisation) are jointly considered when deciding how to actuate a mechanic.

A motivation model that takes this approach can be used to predict concrete action-choices by players in narrowly specified situations. This perspective discards high-level, hard-to-operationalise needs such as autonomy (Ryan & Deci, 2000) or motives such as achievement, affiliation and power (Schultheiss, 2008) as too abstract for concrete interpretation at the zoomed-in level of individual actuations. Instead, such an approach would need to identify what moment-to-moment factors must be incorporated. It seems plausible that strategic performance in the game as in the Rational Player Model (J. H. Smith, 2006) would be one such factor regardless of the behaviour being motivated. Yet it is clear that this must be extended, at least with other out-of-game factors.

The other factors identified here for motivating actuations might plausibly generalise beyond speech motivation, which future research could confirm. Effort Minimisation could apply generically to any kind of actuation as all involve some physical and cognitive effort, however it is uncertain whether it would extend as a consideration to relatively effortless actions such as pressing a button. Considerations of Appropriateness might be more salient in speech, but non-speech actions might also be seen as violations of fair play or situational norms (e.g. in an anatomically explicit round of *Pictionary*). Finally, performative motivations for actuations need not in principle be specific to speech though some types of data may provide less opportunity for them. It is easier to imagine performative actuations in a webcam game or dance game than in a typing game, for instance.

3.4.2 Correspondence with Frame-Analytic Models

In both the Actuation and Communication sub-models, an integration is performed of multiple factors operating *at different levels* of the gaming situation. There are both factors *strictly within the game*: Optimality and Endogenous Value; there are factors that operate

at the *interface to the game*: Effort Minimisation and Procedural Value; and finally there are factors that belong to the *social context of gameplay*: Situational Norms and Social Value.

This identification of three distinct 'levels' of factors corresponds to existing frame-analytic models of gameplay. Frame analysis is a popular tool in game studies for understanding the nature of the gameplaying situation Deterding, 2013. It describes situations as belonging to one or more normative 'frames' such as 'lecture' or 'football match' Goffman, 1986. These frames inform the roles participants take (e.g. 'student', 'spectator') as well as directing their attention to different aspects of the situation.

The three levels of factors described above match those identified in Conway and Trevillian (2015)'s frame analysis of gameplay. They identify three nested frames, each structuring the player's behaviour and experience. Players experience gameplay within a *Social World*, a social situation constituted of the regular social norms and role expectations of such. However, players can also experience themselves also within an *Operative World*, whereby the player is an agent interacting with a system with rules; this brings with it a transformed set of norms and behaviours. This distinction corresponds to the separation in my model between motivational factors inside (e.g. Efficiency) and outside (e.g. Appropriateness) the game. Indeed, corresponding exactly with Allison et al.'s (2019) corroborating frame analysis, I drew a further distinction within these game-internal factors (Operative World). Allison et al. split the Operative World frame into two: the *Functional World*, relating to the use of the game system; and the *Strategic World*, where players focus on achieving their goals within the game; likewise, I distinguish between factors at the *interface to* the game (effort, Procedural Value) and *within* the game (Optimality, Endogenous Value). Within Frame Analysis, each frames introduces norms of experience and behaviour: in terms of speech motivation, I observed corresponding groups of factors influencing motivations for speech.

Finally, Conway and Trevillian (2015) and Allison et al. (2019) both (independently) identified a further frame: a *Character World*, whereby players experience themselves as the character they are controlling within the game. This does not clearly map onto anything in my model. There is some relationship here to Performance (within the Actuation model), which incorporates the characterful dramatisation of an actuation. It is plausible that in some cases this dramatisation may be experienced by players through a Character World frame, with the associated norms of such. While players may frame other performances

such as singing as within a character (e.g. taking on the character of the singer in the game, or the real-world performer), it is unlikely that this is always the case. This frame of player experience doesn't neatly map on to any of our groups of factors. As such, future work might look into whether any corresponding Character World-directed motivations have been missed in the present model.

As in the model presented here, these competing demands imposed by these nested frames are balanced by the players. For instance, Allison et al. (2019, p. 9) observed that players of noisemaking games sought a balance between control of the game via e.g. vowel sounds rather than sentences, and their maintaining the social congruence of their vocalisations. This maps on to the trade-off between Efficiency (control being a function of Optimality and Effort Minimisation) and Appropriateness (vocalisations being meaningful or congruent within a situation being encompassed within the Situational Norms of gameplay). However, the frame analyses don't provide a mechanism by which this balance is accomplished, as provided here.

Frame analyses are about the social norms, role expectations, experience and behaviours of players. While they are not theories of motivation, because frames structure experience and behaviour, we might expect that they can meaningfully categorise motivational factors also. Indeed, future work may look to further integrate these by beginning from the five levels of gameplay framing identified by Allison et al. (2019) and systematically considering motivational influences at each level. For example, in the Character World, how does the player's identification with e.g. an action hero give rise to motivations for speech? One such motivational factor may be avoidance of the experience of the character-player identity dissonance identified by Carter et al. (2015).

3.4.3 Similarities with Player Communication Studies

While much work in player communication is on text, two models of player spoken communication have been developed (Drachen & Smith, 2008; J. H. Smith, 2006). Both emerged out of an iterative coding process and were primarily used for quantitative analysis of speech primarily in terms of the semantic content of that speech. In contrast to the approach taken here, the data they used data were transcripts of game sessions and related to only three and two different games, respectively. These studies looked into the *role* of speech in the gaming situation where as I sought to identify *motivations* for speech. Despite this, there are many similarities between our results.

First, though not a result of their study, Drachen and Smith's (2008) three framing hypotheses are clearly reflected in the three values I found to influence a communicative utterance: players speak *functionally* where a potential utterance has procedural value; *strategically* when it has endogenous value; and they *socialise* when their utterances have social value. However, rather than discrete hypotheses for speech motivation, I found these factors are better understood as influences that additively determine the utterance selected. Indeed, this general reorientation applies to all comparisons between my own work and that of both J. H. Smith (2006) and Drachen and Smith (2008). Naturally enough given their aims, they both categorise utterances into their role, leading to a taxonomy of speech types. In contrast, I did not code utterances directly, and instead found a model that influences utterance construction. Any individual utterance may be influenced by all three factors. This general reorientation of how to view speech in games leads to a new perspective on the literature.

The studies treat the function of speech largely as communication. Both find patterns of quantitative differences in types of speech between the games they study. This pattern fits the explanation presented here that much speech in games is designed to resolve information asymmetries. For example, in pen-and-paper roleplaying, where there is need to jointly establish the world state (resolve information asymmetries between player and game master), a majority of speech addresses this; where such information is projected on a screen to all players, this type of speech disappears. In addition to communicative uses of speech, Drachen and Smith code some utterances as 'Character Action Description', which they identify with illocutionary acts. Such utterances correspond exactly to Actuations in my model. Other than this, many of the codes identified by both studies are easily understood as either information provisions or requests, e.g. asking for advice, help, information or receiving it, and many of these can be identified with likely corresponding values (endogenous, procedural, social). However, they do identify a number of more specific codes that my model is too abstract to interpret except in a general way, such as 'Apology' (Drachen & Smith, 2008) and 'Joke' (J. H. Smith, 2006) which might be related to Social Value. Such utterances seem to be not so much motivated by the game as by the wider social situation, which was not a focus of my theory building.

There are areas where the addition of my model aids interpretation to previous studies. We can assign likely motivations driving most of these: asking or giving advice, help, or information (about e.g. the game world) has likely Endogenous Value as knowing (or the

other player knowing) this contributes to the ability of the players to achieve their goals. Critiques (e.g. of self, of group, of game) could be understood as having multiple values: endogenous to the extent it is motivated by a desire to improve ones performance, and social to the extent is a discussion about a shared interest in the lives of the players. A more detailed investigation to nuance these, admittedly broad, categories, would need to consider both the game design and empirical data (e.g. transcripts) in a way that was not attempted here.

J. H. Smith (2006) finds that the simple game-theoretic strategic explanations of his Rational Player Model is insufficient to explain the quantitative differences in speech he observed between games. In particular, more help was given in a *competitive* game than a *cooperative* game. He suggests the difficulty or novelty of the interface better explains his results. The model presented here may aid interpretation. While the competitive situation would lead to helpful utterances having no endogenous value, a complex interface would mean that utterances to help the other players understand how to the game should be played (i.e. Rule Clarification) would have greater Procedural Value. Such utterances are thus motivated because of the desire to be successful in the game as an interaction (compare Allison et al.'s (2019) Functional World frame). J. H. Smith's Rational Player Model might not be incorrect so much as incomplete. Additional factors such as Procedural and Social Values might be helpfully integrated to attain a more explanatory model of the player.

Finally, one factor that seemed to be absent in both models is actuation effort, though this is unsurprising. First, the majority of the utterances were communicative, not actuations of mechanics. Second, if Appropriateness does overwhelm the influence of Effort Minimisation when in multiplayer situations, we would not expect it to be observed in the exclusively multiplayer gameplay sessions analysed. However, this suggests a benefit of analysing single-player speech interactions in player communication studies, as this may more clearly reveal influences upon communication likely to be obscured in multiplayer interactions.

Addressing the player communication literature more generally, I found a distinction between instrumental (Optimality / Endogenous Value) and non-instrumental (Appropriateness / Social Value) influences on speech motivation that mirrors much of the literature J. H. Smith (2006). Creativity applies not only to text communication, as observed in the FPS *Counter-Strike* (Valve, 2000) by Wright et al. (2002) but also speech, as observed

here in the Performance and Social Value/Creativity codes, though only in general terms. The complexity of textual creative communication suggests that performative and creative motivations could be analysed with far more detail than the abstract treatment here.

While these correspondences contribute to our understanding of players and their motivation, the main intention for the models presented here was to be directly useful in designing games.

3.4.4 Application to Design

As I discussed at the beginning of this chapter, existing models of motivation do not provide concrete guidance for creating a game to elicit speech. This motivated the development of the present model, with the intention that it would give concrete and subject specific guidance for design. The goal was further to identify general principles that would allow designers to go beyond existing known-good design patterns for voice interaction, such as those identified by Allison et al. (2018). While the model presented above is a theoretical account of motivation, and is not design advice *per se*, such a model can be used to support counterfactual thinking in design (Oulasvirta & Hornbæk, 2021). A particular strength of a model is if it deals in things that designers will deal with, is grounded, domain specific, and makes concrete predictions.

First, the concept of an *actuation* is itself helpful in designing games. While the concept of game mechanic is a staple of game design practice (Hunicke et al., 2004), *how* those mechanics are physically controlled is generally overlooked – perhaps because in the majority of cases this consists of simply pushing a button. Actuation, as a concept, allows us to refocus attention on the particular physical action, with its component dimensions and influences, that triggers the mechanic. Whatever data we want to collect must somehow be *encoded* within this actuation. Actuation, as a motivation, embeds this physical act within both the game space and the gameplaying context. We must look at an actuation in a duel way, as simultaneously a game act and a social act. The components of the Actuation model can be used in design to motivate or demotivate particular kinds of speech, as suggested in Table 3.5.

Optimality and Effort Minimisation are both within the control of the designer. First, the designer generally has control both over making certain actuations more or less effective in the game. Second, they can decide what input methods players will use to interact with the game, and what the mapping between input methods and mechanics will be. On the

Table 3.5: Design implications of the Actuation sub-model.

Code	**Design Implication**
Actuation	Make the desired speech input have a direct effect on the state of the game.
Rule-Based Constraints	The inputs parsed by speech recognition or allowable within the grammar and rules of the game should be a minimal superset of the kind of speech input desired.
Optimality	Make the desired speech input uniquely effectual at achieving desirable outcomes in the game.
Effort Minimisation	Align the desired speech with a low-effort actuation. This may mean removing alternate modes of input, such as controllers.
Fair Play Norms	Influence player's perceptions about what is the 'fair' or 'correct' way to play the game, such as by incorporating speech rules whose violation would advantage or disadvantage one or more players. Designers can ensure that *not* producing the desired speech would somehow disadvantage another player, thus generating pressure that they do so.
Situational Norms	Control where, and by whom, an elicitation game is played. The social norms of an experimental situation, for example, may be helpful to get players to behave as desired. The game might be released on particular platforms to take advantage of likely situational norms e.g. games console vs. mobile vs. public installation.

other hand, designers have less influence on Fair Play Norms, but they can still explicitly endorse or condemn certain behaviours in the game rules. Finally, designers have little influence on Situational Norms, though they may occasionally be able to control in which situations a game is played.

Approaches taken to developing accessible versions of games (e.g. Grammenos et al., 2009; Norte and Lobo, 2008) and enhancing the accessibility of existing games (e.g. Harada et al., 2011; Mustaquim, 2013) can be well understood within the Actuation model. The approach generally taken has been to enhance the games interface by adding voice control – whether non-verbal properties (Harada et al., 2011) or speech recognition (e.g. Grammenos et al., 2005) – to trigger existing mechanics in otherwise unchanged game designs, such as dictating a chess move (Grammenos et al., 2005) or naming a row and column in Sudoku (Norte & Lobo, 2008). As such, such a system could be profitably analysed in terms the Actuation model. While actuation effort is widely considered by accessible game designers, it is also relevant to consider how the voice control relates to contextual social norms as well as the potential for actuations being performances.

The Communication model lends itself to formal game design as information asymmetries can be readily formalised and constructed. Existing information asymmetries can be exploited or new information asymmetries might be constructed by the game. This might

be as simple as giving each player a hidden card, or a complex emergent result of gameplay dynamics. Such asymmetries can be foregrounded to the player, so they are perceived, perhaps by integrating them into the core dynamics of the game. Finally, existing games serve as examples as various ways in which three types of value can be given to that information: either through making the information valuable in the game, procedurally necessary to play the game, or of social interest in the context of gameplay. Formal game design tools such as rewards and goals allow imparting information with Endogenous Value. Consideration and playtesting of the procedural practicalities of play allows design towards Procedural Value. Finally, various social games can be mined for ideas about creating social value. Design implications of each of the components of Communication are discussed in Table 3.6.

3.4.5 Limitations

There are many different ways to approach the question of speech motivation, each potentially contributing complementary perspectives that allow us to understand the problem more fully. The two models presented here, understood as lenses in the sense of Schell (2009), could be supplemented by more through further study of the same phenomena from different starting points. For instance, I adopted an approach very different from the qualitative/quantitative methodologies of J. H. Smith (2006) and Drachen and Smith (2008). The correspondences between our results are all the more interesting because of the triangulation of different methodologies. Given that there is still much more to learn about speech motivation, further studies that challenge the assumptions made in this one are warranted.

From the start I worked at a high level of abstraction, treating whole games or game mechanics as a unit. This required a significant amount of 'conceptual chunking' to just get started. This chunking would have brought baggage from my own knowledge and experience of games and game design that will have directed and potentially limited the resulting theory. The significant role of autoethnography is liable to further entrench these biases as my personal self-reflected experience may not generalise to other players. Furthermore, working from game rules, descriptions, even performative play such as 'Let's Plays' would have obscured experiential factors that the use of e.g. interviews may have identified.

An alternative would have been working from transcripts or interviews. However, I feel

Table 3.6: Suggested design implications of the Communication sub-model. Note that not all of these suggestions are necessarily good ideas from the perspective of enjoyment or usability. If a type of speech is not desired, these give suggestions of designs to avoid.

Code	**Design Implication**
Communication and Information Asymmetry	Restrict certain players' access to game-relevant information, provide new information to only a subset of players, or identify information asymmetries the players are likely to bring to the game (e.g. personal information).
Direct Value	Create a goal of learning a particular piece of information. Make this information available to some players an constraint how it can be communicated to elicit the desired speech.
Player State	Give players hidden information, such as a hidden hand of cards, and make this of strategic value to other players. Introduce rules defining when players are entitled to know information about other player's hands.
Subjectivity	Make player intentions matter, either as team-mates or opponents. Introduce scoring dependant on subjective preference, or introduce mechanics based on agreement.
Situational Facts	Ensure strategic play requires integrating information from a complex world state over which each player has a limited view.
Turn Taking	Introduce complexities in turn structure, such as mechanics where a player can require another player to take an action. Make how a player needs to act dependant on the history of actions in the game.
Highlighting	Introduce rules that entitle other players to certain information during play (e.g. holding a single card in *Mao*). Ensure non-speech representations of game actions (e.g. cards placed on table) are sufficiently hard to follow in themselves that they require narration.
Rule Clarification	Use complex or dynamic rule systems, such as introducing context dependant rules or having a lot of special cases experienced players might memorise.
Identity	Introduce free-choice question-asking mechanics or prompt questions about the players. Require players to share aspects of their identity within the game, e.g. during character creation.
Real-World Facts	Raise topics of likely interest to the players, e.g. through the game's fiction or discussion prompts.
Creativity	Give opportunities for creative expression, such as through collaborative storytelling or roleplaying.
Information Constraints	Limit or mandate the communication of information that is likely to be expressed in the speech desired.
Constraints of Form	Constrain the form of allowable speech to that of the desired speech as far as possible.
Emergent Constraints	Consider how the dynamics of the game could be made to hinder successful communication, such as by time pressure, parallel conversations, game sound effects or music, or licensing players to engage in disruptive behaviour.

it would have shifted the focus away from a wide-scale survey of abstract design features of games that systemically elicit speech to spontaneous or naturally occurring motivations for speech in an inevitably narrower set of games, over which the designer may have less control. Part of what I brought to the analysis, including the autoethnography, was precisely my game design experience that allowed me to draw out meaningful concepts from game systems, rather than aggregating them from recorded instances of play. This was motivated due to my, at first, extremely pragmatic research goals: I wanted to work out how to design an elicitation game for speech data. In as much as it has resulted in a model with concrete (and, hopefully, accurate) design implications, I consider it successful. Still, I feel there is a need for further studies that challenge my approach.

The level of abstraction of my model means several parts of it are underspecified. Performance and Social Value/Creativity is too abstract to capture the sort of variety of motivations suggested by Wright et al. (2002) for text communication. Appropriateness is underspecified with regards to what the influential Fair Play Norms or Situational Norms might be, although existing research has discussed these (Deterding, 2013). Optimality is similarly general; it does not address *how* players decide the optimal action, nor what sorts of goal they are likely to strive for. Moreover players rationality is bounded by biases and heuristics (Camerer & Loewenstein, 2004) and simple rational models are insufficient to explain player behaviour (J. H. Smith, 2006). However, the intention of this study was not to nuance every domain – many of which designers will have extensive experience in – but to contribute a high-level framework identifying and relating them in the specific context of speech elicitation.

My methodology has not identified factors relating to the *desirability* of elicited speech by other players, even though harassment and other forms of toxic behaviour are a major issue in online multiplayer games with speech communication (Kuznekoff & Rose, 2013). While such issues might be affected by the game design (Kordyaka & Kruse, 2021), they are hard to identify *from* the game design, much as the phenomenon of trash talk among chess hustlers is not clearly related to the design of the game of *Chess* (Tower, 2007).

Finally, there is nothing to suggest whether speech being a designed outcome of the game – an elicitation game – would transform the motivations identified. Yet research in motivations for citizen science (Curtis, 2015; Iacovides et al., 2013; Tinati et al., 2017) and experimental participation (Orne, 1962; Orne & Whitehouse, 2000) suggest that the real-world consequences of such gameplay is a major influence in motivation and behaviour.

Would players experience speech motivation differently if they knew their speech was being recorded and used for some real-world purpose? Until there are such games, we cannot know. There is a breadth of further research and its integration required before we will really know why players say what they do in the game. My hope is that, in the mean time, I have contributed model that can help designers right now.

3.5 Conclusion

I set out at the beginning of this chapter to understand why players might be motivated to engage with a game in a way that provides data. I did this for the case study of spoken language in entertainment games. I constructed a model that describes how a game's design influences the speech data it elicits. The two constituent sub-models reflect two routes through which players' speech is determined: either as an actuation of a control, or to satisfy a information asymmetry through communication. Central to this model is the idea that speech is elicited by moments of gameplay.

As presented, this model is too domain-specific to aid us in the design of most elicitation games unless those happen to be about the elicitation of speech data. Nor does it systematically address threats to the validity of the data that it elicited. Indeed, the role of this chapter in the thesis was to ground our understanding of motivation for individual actions in a game. Unlike existing molar theories of motivation, such as Self-Determination Theory, the type of theory developed here is geared towards moment-to-moment motivation, which I have argued is necessary to understand motivation (and validity) in elicitation games. I will adopt its perspective on motivation, as summarised above:

1. The unit of behaviour that is motivated is the actuation.
2. One's motivation for an actuation is inherently relative to the other possible actuations available at that moment in time.
3. The evaluation of possible actuations integrates multiple factors, both in-game and out-of-game.

This perspective on motivation will be reintroduced based on general theoretical grounds in the next chapter. The next chapter will then construct a theoretical model – similar in many ways to the one I have shown here – yet suitable for the design and analysis of an elicitation game for any kind of data.

Chapter 4

Modelling Data Provision: Intrinsic Elicitation

A key factor in the success of applied games for data collection is their design (Cooper, Treuille, et al., 2010; Reeves & Sherwood, 2010; von Ahn & Dabbish, 2008). Existing work has explored design strategies for *either* motivating players *or* ensuring data quality after collection (Pe-Than et al., 2012; Quinn & Bederson, 2011). Validation strategies for the latter include agreement designs, reputation systems, and automatic checks, while motivation strategies have focused individual game design elements like rewards, effectively 'gamifying' data collection games. This overall gamification+validation approach frames the key design challenge of data collection games as one of post-hoc filtering and checking a maximized volume of player-provided data. As such, it treats player motivation and data quality as *separate concerns*. Yet different forms of motivation have been shown to directly affect the kind and quality of data provided (Rogstadius et al., 2011; Wenemark et al., 2011). This marks a significant shortcoming of gamification+validation approaches and invites the search for broader *elicitation approaches* that integrate *both* motivation *and* data quality: models and design principles for motivating players to provide desired data in a desired quality from the outset.

Such an integrated approach is necessary when we seek to design applied games to collect human-subject data (such as preferences, beliefs, or performance) – games which I have labelled *elicitation games*. Within such games there is no way of validating data post-hoc against an objective or consensus ground truth. I set out in this thesis to develop such an integrated model. The necessary background for this project is now in place. In chapter 2, I established the threats to validity characteristic of the use of games, which

need to be controlled, and suggested that a moment-by-moment perspective on validity was necessary when working with 'whole' games. In chapter 3, I settled on a perspective on motivation that is suitable for the development of concrete design guidance, again a moment-by-moment approach.

Now in this chapter I introduce an integrated design approach to the development of elicitation games that I call *Intrinsic Elicitation*, akin to the principle of *intrinsic integration* in e.g. educational games or gamification (Deterding, 2015; Habgood & Ainsworth, 2011). In short, Intrinsic Elicitation captures the idea that generating desired data in a desired quality should be integrated into the mechanics of the game in such a way that it is the necessary, strategically optimal, and least effortful way for the player to pursue the game's goal. This integration is performed at the moment-by-moment level.

I will develop and defend this approach as follows: First, I discuss existing motivation and validation design strategies in more depth. I argue that data collection games *as games* introduce a systematic motivational driver *and* threat to data validity at once which existing gamification+validation work hasn't addressed. Therefore, I present a theoretical model of why players provide particular kinds of data in a game, the *Rational Game User Model*, inspired by the previous two chapters and integrating J. H. Smith's (2006) *Rational Player Model* with existing theoretical and empirical work on data collection games and survey engagement. From this model, third, I derive the design approach of Intrinsic Elicitation, comprising three heuristics for how to integrate data collection into a game's mechanics: Necessity, Centrality, and Veracity. I illustrate the utility of our approach as a predictive model and evaluative tool by analysing the data collection games *Apetopia* (Barthel, 2013), *BeFaced* (Tan et al., 2014), and *Urbanlogy* (Celino et al., 2012) through its lens.

4.1 Background

Numerous researchers have studied how to design data collection games for motivation and data quality. Several studies have probed the underlying motives of players, converging on constructs of *gaming enjoyment* (e.g. fun, competence, relaxation), *community participation* (status, recognition, social norms), and *meaning* (making a contribution to science, helping others, growing oneself), with *monetary benefits* being rarely used and reported (Curtis, 2015; Goh & Lee, 2011a, 2011b; Tinati et al., 2017). In their systematic analysis, Tinati et al. (2017) suggest that these motives can be classified following self-

determination theory (Deci & Ryan, 2000) into *intrinsic motivations* (gaming enjoyment factors like competence, the community factor relatedness, and all meaning factors) and *extrinsic motivations* (monetary benefits, achievement, status, recognition, social norms). This is particularly relevant as research on crowdsourcing and survey design suggests that intrinsic motivation leads to higher participation rates *and higher-quality data* (Rogstadius et al., 2011; van Grinsven, 2015; Wenemark et al., 2011) – comparable work on data collection games is unfortunately harder to find (N. R. Prestopnik et al., 2014; N. R. Prestopnik & Tang, 2015).

As for *motivational design strategies*, the literature has chiefly explored individual game design features such as difficulty balancing (A. Sarkar et al., 2017), visual appeal (Wang et al., 2014), graded goals (Gaston & Cooper, 2017), narrative and theming (N. R. Prestopnik & Tang, 2015), and reward systems (Gary et al., 2017; Goh et al., 2017; N. R. Prestopnik & Tang, 2015; Siu & Riedl, 2016), though with various results. In a sense, the existing literature has been less concerned with what core loops and mechanics (Deterding, 2015; Sicart, 2008, 2015) fit what data collection tasks than with 'gamifying' data collection games – adding and tweaking presumed-engaging design features. This is arguably because the majority of data collection games follows the template of the early successful and much-publicized *GWAPs* of Louis von Ahn and colleagues (von Ahn & Dabbish, 2008): players classify or transcribe presented (usually visual) data, receiving points for every (correct) input. A second such influential template is *Foldit* by Seth Cooper and colleagues (Cooper, 2015), where players generate solutions to problems where there exists a method for checking the quality of a solution, but the actual best solution is unknown; here, players score based on the number and optimality of provided solutions. There are good practical reasons for the popularity of these templates: they offer working models that are easily replicated via open access platforms like *Galaxy Zoo*; they address classification and solution discovery tasks with broad applications; and they provide straightforward means of *validating generated data* – the second main design concern of data collection games.

Commonly used *validation strategies* are agreement designs, automatic solution evaluation, and reputation systems (Cooper, Treuille, et al., 2010). In agreement designs found in *GWAPs* and most data classification games, a player's input is assessed on how much it agrees with the inputs of other players on the same stimulus or task (von Ahn & Dabbish, 2008). Poor responses are filtered out or demoted in their weighing as they are unlikely to 'agree' with the consensus of the majority and/or trusted players. This is somewhat

data-inefficient as it requires multiple people to solve the same task. In contrast, solution discovery games like *FoldIt* (Cooper, Khatib, et al., 2010), automatically evaluate the quality of each submitted solution against known and computationally formalized optimality criteria. Solutions are ranked and scored based on how close they come to the theoretical optimum. While potentially more data-efficient than agreement designs, this validation strategy obviously requires prior knowledge of solution requirements that can be computationally expressed and validated. Finally, both agreement and automatic evaluation designs often feed into reputation systems that track player performance over time to identify which players reliably provide high- or low-quality data (Cooper, Treuille, et al., 2010). These reputation scores can then be used to weigh answers in agreement designs, filter out data by low-scored players, or optimize player-task matching, e.g. serving difficult or unsolved tasks to high-scoring players first (A. Sarkar et al., 2017).

To summarize, current design research on data collection games has chiefly focused on variations of *GWAP*-style data classification games and *Foldit*-style solution discovery games, replicating their core game mechanics based on validation strategies (Cooper, 2015; Pe-Than et al., 2012; Quinn & Bederson, 2011). In this, research has treated player motivation and data quality as mostly separate design concerns addressed with separate solutions: 'gamifying' data generation with reward systems etc. to maximise data *volume*, then validating generated data with agreement, automatic evaluation, or reputation systems to maximise data *quality*. This gamification+validation approach works with the somewhat wasteful assumption that some or even significant amounts of poor-quality data are inescapable: as long as *some* players provide good data, ground truth will out and can be used to identify and reward high-quality data. More importantly, this approach necessarily requires *some* validation against a known objective or consensus ground truth.[1]

4.1.1 The Challenge of Human-Subject Data

Exactly this requirement sets *GWAP*- and *Foldit*-like games apart from games (including elicitation games) designed to collect *human-subject data* like short-term memory processes (H. R. Brown et al., 2014) or people's performance (Bellotti et al., 2013). It also limits the applicability of their underlying gamification+validation approach. Contrary to classification tasks or solution discoveries, for human-subject data, the ground truth is often

[1] See e.g. Siu, Zook and Riedl's (Siu et al., 2017) framework of mechanics for human computation games, which includes validation as a necessary component.

unknown and *unknowable* to anyone but the subject, and data validity can not be equated with players doing 'as best they can'. For subjective attitudes, values, or preferences, there is by definition no subject-external ground truth to assess them against. We often aggregate such data ("people on average give this service a 7.5 net promoter score"), but in that usually want each subject to honestly report their independent evaluation, not their 'best guess' at what an average evaluation would be. Similarly, much of human-subject research is interested in covert, non-conscious processes, tendencies, dispositions, states, or traits that reveal themselves in people's 'spontaneous' responses, e.g. the preferred walking speed as an expression of fitness or wage levels and derived valuations of time (Browning et al., 2006; Levine & Norenzayan, 1999). In these instances, the moment one communicates one answer to be 'more true' or 'more optimal', this would distort the generated data. Even where there are operationalizable scales for performance (such as IQ or money earned in a game theoretical experiment), people may be motivated to overstate (or understate) their 'true' performance ability because it is socially desirable or rewarded by the game. And again, for individual capabilities like IQ, there is no subject-external ground truth to assess how accurately the subject's current recorded performance reflects its 'true' underlying capability.

More generally, the collection of human-subject data as in elicitation games introduces specific data types, validity criteria and validity threats that gamification+validation approaches don't reliably address. Worse, where gamification+validation approaches model or even reward certain responses as 'better' or 'more true' than others, they generate particular new validity threats. In education and gamification, these threats have been discussed as *gaming the system* (Baker, 2011; Bevan & Hood, 2006; Werbach & Hunter, 2012) or *cheating* (Mccabe et al., 2001). Individuals game the system when they find ways within the rules of a system to maximise their evaluation metrics at the expense of the substantive goals intended by the system, e.g. giving short nonsense answers in a question and answer platform because every answer receives points irrespective of content. Individuals cheat when they covertly gain an advantage through means outside the rules of a system, such as secretly copying answers for a test from a colleague.

This raises the obvious question what game design approaches are better suited to human-subject data than the currently prevalent gamification+validation approach. Education and gamification research are plausible sources of alternatives since both often involve human-subject data collection and have dealt with gaming the system and cheat-

ing. One consistent argument across these two fields, going back to T. Malone (1980), is that the outcome of applied gaming – teaching particular skills, making a particular activity more engaging – should in some way be *integrated* into the game's mechanics (Deterding, 2015; Habgood & Ainsworth, 2011; K. Squire, 2006). In game-based learning, this principle is called *intrinsic integration* (Habgood & Ainsworth, 2011). It is 'intrinsic' in that (a) the learning material is part and parcel of the enjoyable, intrinsically motivating core mechanic of the game, and (b) the game's mechanics and thematic world embody and represent the learning material. Deterding (2015) suggests in direct analogy that effective gamification is intrinsically integrated by turning the target activity into the core mechanic of the game, reorganizing its 'core loop' to support intrinsic motives like competence. There is some evidence that intrinsically integrated educational games and gamified interventions outperform their alternatives (Buikstra et al., 2015; Echeverría et al., 2012).

This notion of intrinsic integration is not without parallels in data collection games. Tuite (2014) cautions that GWAPs should "match mechanics to purpose". Observing that existing validation templates are insufficient to design new GWAPs, Galli (2014) offers a range of standard game mechanics that match different GWAP task types, e.g. memorisation maps clustering. Jamieson et al. (2012) suggest identifying mechanics for human computation games by finding mechanics or real-world activities that are 'isomorphic' to the structure of the computational task. However, Jamieson et al. and others argue for this as a way to reduce extraneous effort and cognitive load and ease problem-solving, and none of them address the particular validity threats of games for collecting human-subject data.

4.1.2 The Work So Far

I will argue that we need an intrinsic integration approach that not only matches a mechanic to a particular data type, but also motivates players to provide honest data without biasing results, especially in instances where there is no subject-external ground truth to validate responses against. This intuition – that we should integrate motivation for *valid* data provision, for specific data, into the core mechanics of the game – is the core motivation behind the design of the Intrinsic Elicitation model that I will present in this chapter.

How do we go about achieving this integrated picture of motivation, validity, and game mechanics? The pattern we have observed so far in this thesis is that validity and data provision in games can be perhaps best understood at the moment-by-moment, or

molecular, level. From a consideration of threats to validity in chapter 2 we observed that the game characteristics that give rise to enjoyable gameplay can also threaten validity. Whether they do so is highly context specific: it depends on the research purpose that a game is being used for. Moreover, gameplay is highly variant, suggesting that validity can be poorly understood as a molar property of a game as a whole. Indeed, in the theory-generating work in chapter 3, I found that a key result in the developing theory was that motivation for data provision was best understood at a moment-by-moment level. A key result of that model was that players make decisions about what data to provide based on the moment-by-moment circumstances of the game. Thus, in order to develop an analogue of intrinsic integration for data elicitation, we need a clear idea of what motivates players of data collection games to take particular in-game actions (and thus provide particular data) rather than others.

I have already explored this in chapter 3 in the context of speech elicitation with games. There I developed a grounded theoretical model of motivation for the elicitation of speech in games. This was intended to give an understanding of why, in the case of a specific type of data, game players might choose to act in such a way that provides data (speech) rather than a way that does not. Two motivations were identified, being Actuation (being motivated to act in a way that changes the game's state) and Communication (motivation arising from the desire to communicate information to another player or to the game).

This work was specific to eliciting speech data and not all elicitation games are about this kind of data. Thus we need to make an abductive leap from the specific, concrete model proposed in the last chapter to more general principles that can be applied to the elicitation of data of any sort. I already made some steps towards generalisation when I sketched the resulting perspective on motivation in the light of that chapter (section 3.4.1). Here the unit of motivated behaviour is the actuation of a mechanic, one's motivation is relative to other possibilities at that moment in time, and the evaluation of possible actuations integrates multiple factors. This is in essence a rational choice model: players select the optimal actuation at a given moment by rationally evaluating the possible actions at a given moment and selecting the best candidate. A rational choice approach might therefore be particularly appropriate for building a general theoretical model of data provision in games. Helpfully, there is an existing and widely used rational choice model in the field of games, which is the *Rational Player Model* of J. H. Smith (2006). This model, introduced just below, will form the starting point from which I will develop a model of data provision to

games, which I will call the *Rational Game User Model*. After I have proposed this new model I will link it back to the model of speech motivation that was developed in chapter 3. From the Rational Game User Model I will then derive my theory of Intrinsic Elicitation.

4.2 Theory

4.2.1 The Rational Player Model

Implicitly or explicitly, game designers, researchers, and members of the public hold different mental models of players (J. H. Smith, 2006). For instance, we may view players as mostly passive objects of game stimuli (a view proposed by strong media effects research), or as highly autonomous subjects appropriating games as a mere material for their own ends, as in e.g. Sicart's (2014) humanist rhetoric of play. The model underlying much if not most game design practice is what J. H. Smith (2006) has elucidated as the *Rational Player Model*. In short, the Rational Player Model states that in-game, players are self-determined and rational actors "whose main (or only) concern is to optimize [their] chances of achieving the [game's] goals" (J. H. Smith, 2006, p. 34). Put more simply, people play to win.

This model holds substantial merit despite and because of its simplicity. Games are often distinguished from toys solely through the possession of a goal (Juul, 2005). Striving towards goals is widely seen as what distinguishes 'gaming' from 'playing' (Deterding, 2015). Across studies of player motivation and experience, players state that experiences of competence, mastery, and achievement gained from achieving game goals make playing engaging (Boyle et al., 2012). In addition, 'playing to win' is an important and actively sanctioned social norm of most gaming encounters: to make no visible effort to win during gameplay usually results in being reprimanded as a 'spoilsport' (see Deterding, 2014, pp. 174-5 for a review). Gaming is one of the few types of social situations where "setting aside all personal feelings and all impulsive inclinations" to rationally maximise one's own goal attainment is allowed and indeed expected (Goffman, 1972, p. 96). And in any domain of everyday life, goals are extensively and effectively used to motivate and direct effort (Gollwitzer & Oettingen, 2012). Within a game, goals have a similar function, directing player effort towards particular future states.

Starting with the assumption that players try to act strategically optimal to attain a game's goals is not only well-supported: it also opens the way to powerful conceptual tools

for analysing and predicting how particular game design choices will affect in-game actions, as J. H. Smith (2006) demonstrates in his formalisation of the Rational Player Model. He explicates the Rational Player Model using *game theory*, the mathematically formalized study of *strategic interaction* when two or more actors make decisions with clearly defined objectives, taking their knowledge and expectations of the *other* actors' objectives and decisions into account (Osborne & Rubinstein, 1994). As such, game theory shares basic assumptions (and mathematical inclinations) with rational choice theory in sociology and economics, namely that people are individual agents acting to rationally maximize their personal utility (Scott, 2000). Translated into gameplay, players are rational agents seeking to optimise their utility as defined by the game's goals. In this view, 'gaming the system' as defined previously is normal and indeed *expected* gaming behaviour, as is cheating: cheating is likely when the expected utility of cheating (minus the expected disutility of the chance of being caught) is bigger than the utility of available alternative actions.

Being caught leads us to the point that player's in-game choices are affected by more considerations than winning. Juul (2008) for instance articulates at least three concerns from which players can and often do assess in-game actions: the first is goal-orientation or the desire to win, matching the Rational Player Model. This is nested in a concern for gameplay as an interesting experience: players try to maximise their enjoyment, perhaps by playing in a way that is strategically sub-optimal but more novel and interesting. Even this is coached in a third wider consideration of the social implications of game actions. We often self-handicap when playing against children, for example. However, all these concerns do not speak against a game theoretic analysis of gameplay. Rather, they highlight a common narrow misconstrual of the concept of *utility*. Originating in Bentham's utilitarianism, utility expresses the tendency of an object or action to "produce benefit, advantage, pleasure, good, or happiness" (and reduce or prevent the opposite) (Broome, 1991). In later economics, this substantive conception has been replaced by an operational definition of utility as the *preferences* of individuals, as revealed in their choices (Broome, 1991). Either conception is perfectly in keeping with any and all considerations Juul (2008) (and others) have brought forward as informing in-game actions. If we value and derive pleasure from making a child happy, and more so than winning ourselves, then self-handicapping is the rational choice that maximizes our utility, as revealed in our choice to self-handicap. The utility that a rational player seeks to maximise can include the joy of winning, other joys of gameplay, and surrounding social norms at once.

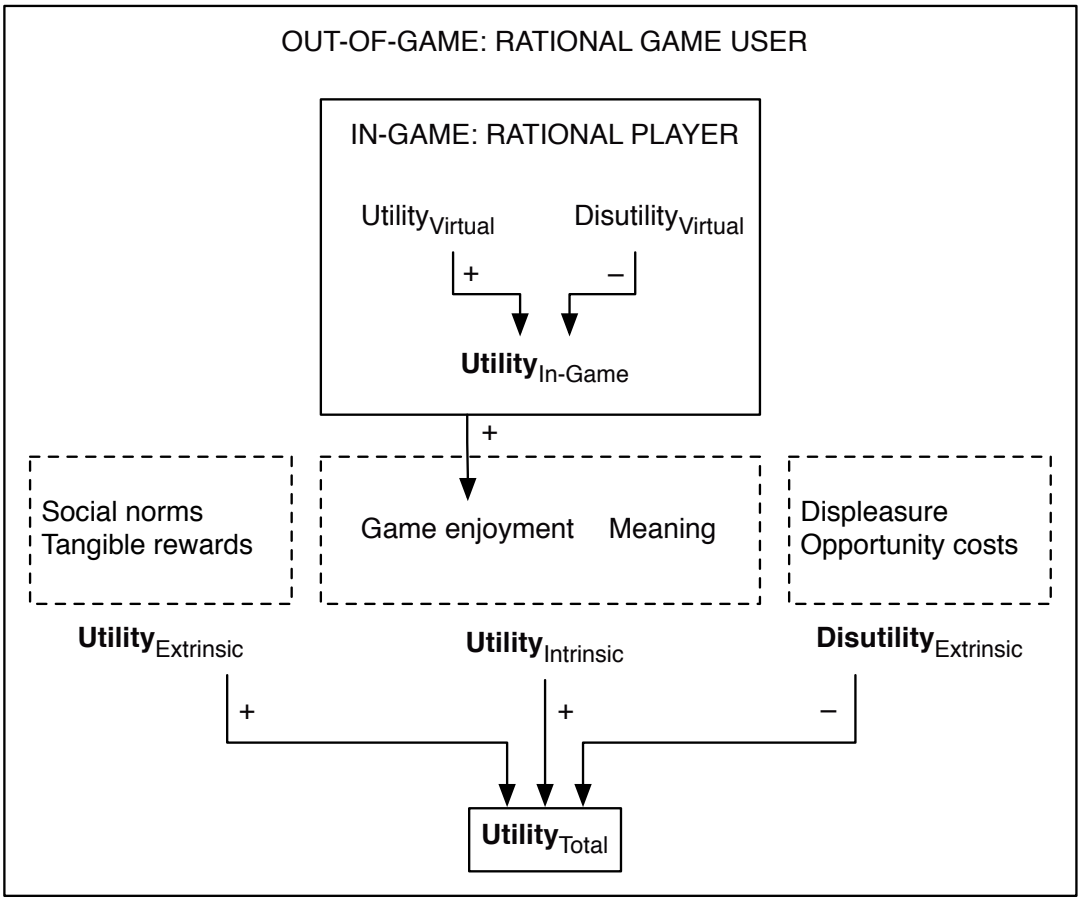


Figure 4.1: The Rational Game User Model.

4.2.2 The Rational Game User Model

For the design of data collection games, I therefore suggest a strong rational choice-style abstraction that starts with the strategic utility of in-game actions for accomplishing game goals. I see five main benefits to doing so. Firstly, this 'in-game' utility elegantly compresses many known important considerations of players outlined above, from mastery to the social norm of 'playing to win'. Second, Smith's own empirical work indicates that players do rationally play to win in-game and modulate arising social concerns through parallel out-of-game talk (J. H. Smith, 2006). Third, it is literal textbook practice and therefore easy to adopt by practitioners: game design textbooks regularly advise to use game theory to calculate the strategic utility of in-game actions and objects as part of balancing to afford interesting decisions and a perceived 'fair' chance at winning (Schell, 2009). Fourth, it allows to articulate clear, mathematically expressed hypotheses and predictions, enabling rigorous empirical testing and robust design guidance. Fifth and finally, I concur with

Healy (2017) that 'nuance' in theory is overrated: theory is more pragmatically powerful when it needs *less* starting assumptions and data to make good-enough predictions, and knowledge generation can be more usefully guided and integrated by extending simple models with variables like 'social context' only if and when the data requires it.

In-game utility encompasses two of the principle factors involved in speech motivation within the model developed in chapter 3. There Optimality is given as a factor of Actuation (the motivation to affect the game state) whereby players prioritise the actuation that optimises their performance in the game. Similarly, within the motive of Communication (the motivation to communicate information of value), one value of information was Endogenous Value, where information is valued that contributes to achieving goals in the game. Both of these express instances of in-game utility for data provision. Further, the selection of a rational-choice model also satisfies the suggestion that validity and motivation are best addressed at a moment-by-moment level from chapters 2 and 3. Rational choices can be modelled at the level of actions which lets us consider the individual validity of the data of each action.

Still, I recognise that in-game actions do not occur in a vacuum. For human-subject data collection games in particular, we need to consider what utilities existing research has identified beyond game enjoyment. Here we can draw on literatures on motivations for participating in human subject research (particularly online surveys), volunteer crowdsourcing, and citizen science, as they are structurally comparable to data collection games. For web surveys in particular, researchers have developed and tested rational choice models that predict participation decisions (Fan & Yan, 2010). Surveying major recent reviews (Baruch et al., 2016; Fan & Yan, 2010; Geoghegan et al., 2016; Rogstadius et al., 2011; van Grinsven, 2015), a picture broadly concordant with the data collection game literature emerges, notably excluding gaming enjoyment factors. First and foremost, participation is intrinsically motivated by *meaning* factors: helping others, contributing to science, growing oneself through learning. Second, participation is extrinsically motivated by perceived social norms like reciprocity and, where offered, tangible rewards like money. Third and finally, participants take into account the *disutility* of labor involved in participating, such as the opportunity costs of foregone alternative ways of spending the same time and the active displeasure of doing something boring or strenuous (D. A. Spencer, 2003). In online surveys for instance, overly long surveys or poor usability lead to low participation, high abandonment, careless responding or satisficing: doing just enough to achieve a somewhat

satisfactory answer (Fan & Yan, 2010; Leiner, 2013).

Given the general concordance of the different motivation literatures, I assume that these out-of-game factors or utilities also operate when players choose which in-game action to take (and thus data to provide) in a data collection game. This is borne out in the last chapter where I found out-of-game factors such as Appropriateness (speaking in a situationally appropriate way) and Effort Minimisation (prioritisation of the actuation taking least effort) were significant in player motivation for speech. Indeed, I concluded in line with the speech motivation literature that factors for motivation included those within the game, those at the interface to the game, and those wholly outside the game.

I summarize these in-game and out-of-game factors in what I call the Rational Game User Model (illustrated in figure 4.1): as a rational game user, players want to maximise their **total utility**. This includes out-of-game **extrinsic utilities** like *social norms* and *tangible rewards*: playing the game because doing so is incentivised and/or socially expected or sanctioned. On their own, extrinsic utilities will motivate players to play the game/provide data *just enough* to satisfy incentive criteria or social expectations with minimum effort, honesty, and care: they invite gaming the system and careless responding (Rogstadius et al., 2011; van Grinsven, 2015; Wenemark et al., 2011). They also invite dishonest responding (J. A. Johnson, 2005) if participants consider it socially desirable to over- or under-report certain traits or beliefs. This stands in stark contrast to **intrinsic utilities**, especially *meaning*. As players are motivated to help others or science and find the game itself a valid means of doing so, they will attempt to provide data that optimally serves the game's ulterior purpose, e.g. answering an open-ended question in detail and truthfully, despite the involved effort or social undesirability of the answer.

The second, game-specific intrinsic utility is *game enjoyment*. This is where the model incorporates the rational player. To maximize game enjoyment, the rational game user analyses the current game state as a rational player trying to maximize their **in-game utility** as expressed in the game's goals. To do so, they assess each currently possible action for its *virtual utility* (how much or likely the action moves them closer to goals) and *virtual disutility* (the opportunity costs of in-game resources spent on the action). While game enjoyment uniquely motivates data provision in data collection games, it also opens a second, game-specific route to the validity threat of gaming the system: if an action/data provision maximizes in-game utility more than the more honest or 'spontaneous' option, players are systematically more likely to choose the more in-game optimal and thus

enjoyable action.

Finally, the rational game user model acknowledges that data provision is effortful labor – else, there would be no need to motivate it with a game. Thus, beyond extrinsic and intrinsic utility, a rational game user will consider an action's **extrinsic disutility**. Playing a game involves effort and displaces other activities we could have done instead. One may object that because gameplay is intrinsically motivating and enjoyable, it lacks the common displeasures of labor and should be one of the most highly preferred activities. But players do regularly report negative experiences like frustration during gameplay e.g. due to poor playability and usability (Goh & Lee, 2011b), and phenomena like goldfarming demonstrate that some aspects of gameplay (like 'grinding') have a high enough disutility that people pay money to free their time for other, more preferred activities (Heeks, 2010). Even in enjoyable games, players regularly satisfice, making good-enough choices instead of investing more time and effort into calculating the absolute optimal move. This was observed in the last chapter where examples were given where players strongly minimise the effort they are willing undertake speaking to a game, despite the games being voluntarily played for enjoyment. This led to the identification of Effort Minimisation as a factor of the motivation for speech. Effort Minimisation contributes to Optimality, that players find the optimal, low-effort high-reward way to control the game with speech, which may apply also to other kinds of data-providing input mechanisms. For data collection games, this means that all else being equal, responding honestly or spontaneously should be the most effortless option, or at least as effortless as any other available choice. As displeasure and opportunity costs increase, players will be more likely to respond carelessly, satisfice, or even abandon the game.

4.2.3 Relationship with the Speech Motivation Model

The development of the Rational Game User model was informed by the speech motivation model in chapter 3. There is a significant abductive leap between the two as we go from a specific case to the general case, resulting in a significantly more abstract model. In the construction and presentation of elements of the present model, the primary justification (as opposed to the *inspiration*) was drawn from the literature, allowing it to be first published

independently[2]. As such while there are links between the models, these are conceptual as opposed to structural and different language is adopted in each.

In brief, the Rational Game User model asserts that players pick the actuation with the highest total utility, and composes total utility as the sum of intrinsic and extrinsic utilities. One of these intrinsic utilities is game enjoyment, which is composed of in-game utility and in-game disutility. The Actuation and Communication models from chapter 3 are not so neatly structured, but similarly compose motivation out of multiple influencing factors, though the nature of these differ between the models.

Actuation was introduced in chapter 3 as the motivation to affect the game state using a speech mechanic. Optimality is a factor of Actuation whereby players prioritise the actuation that optimises their performance in the game. This was my starting point. Optimality maps directly on to the concept of utility, specifically 'virtual' utility, or utility 'in-game'. The concept of Endogenous value, arising in the Communication model, is similarly about 'in-game' utility, but addressed to information instead of actions. Mapping across to the game studies literature, we see this virtual utility concept described by J. H. Smith (2006) as the Rational Player Model, and situated by Juul (2008) as a frame nested within other frames of the gameplay situation. Thus following the literature, I at once had a strongly motivated structure through which I could transform the speech motivation model.

If Optimality gives us the goal-directed action that sits at the heart of gameplay, the other factors of the speech motivation model become the layers of influences that impact actuation selection in real-world contexts, the nested frames of Juul (2008). We observed in chapter 3 several motivations that go beyond playing to win. In the Actuation model, we observed several factors that seem to limit or moderate a player's attempts to optimise their performance. Two were Effort Minimisation (the desire to perform the least effortful actuation) and Situational Norms (norms of the wider game-playing context that largely constrain what sorts of actuations are contextually appropriate). In rational choice parlance, these are merely specific disutilities: engaging in effortful action and violating social norms are displeasurable (or revealed to be dispreferred from observing player's behaviour). Thus we can generalise the model such that anything perceived as a

[2]This chapter previously appeared as David Gundry and Sebastian Deterding. 2018. Intrinsic elicitation: A model and design approach for games collecting human subject data. Proceedings of the 13th International Conference on the Foundations of Digital Games

disutility will affect the player's calculation of the optimal actuation. These were restated above as displeasure and opportunity costs, Similarly, positive utilities can be imagined: although I did not observe them in the speech motivation model, one could easily imagine that tangible reward could be used to motivate particular actuations. Fair Play Norms (voluntarily adhering to conventions of how to play) can be understood either positively as the utility of social norms (producing actuations of a socially desirable kind because it is normative to do so), or negatively as the disutility of displeasure.

The classification of motivations into intrinsic and extrinsic (Tinati et al., 2017) provides a principled way to divide these kinds of extrinsic (dis)utilites from the sort of intrinsic utility derived from game enjoyment. For maximum simplicity, the Rational Game User Model sees game enjoyment as solely determined by virtual utility. However, there are already possible alternative sources identified in the speech motivation model. Communication (the motivation to communicate information to resolve perceived information asymmetries) was modelled as being composed of three sources of information value: Endogenous Value (discussed above), Social Value, and Procedural Value. Social Value (the valuing of an information asymmetry because of what it says about ourselves or another player) can be understood as the value of information of which the communication has positive utility because of the social benefits to doing so. Procedural Value, as seemingly motivated by the desire to be skilled participants in the gameplay interaction (as distinct from being good at the game), is another positive utility. These two might be additional factors beyond virtual utility that determine game enjoyment or alternative sources of intrinsic utility not included in the Rational Game User Model. They were not included as they were not motivated from the literature above. However, the model could easily be extended with these if motivated.

4.3 Intrinsic Elicitation

From the Rational Game User Model, we can derive requirements for translating human-subject data collection into applied games, specifically their mechanics and core loops, which are commonly considered the primary formal aspects of a game (Deterding, 2015). *Game mechanics* refer to the methods by which an in-game agent effects a game state change (Sicart, 2008). They are the verbs of the game, like 'jumping', 'shooting,' or 'drawing a card'. *Loops* describe cycles of mechanic actuation, system processing, and system feedback relative to one or more game goals (Deterding, 2015; Sicart, 2014). Mechanics

can be actuated in multiple ways with differing effort: we can jump high or low, and shoot with careful or careless aim. Thus, any in-game action involves two degrees of freedom. First, a player must choose *which* mechanic to actuate. Second, they choose *how* to actuate it. This latter *how* defines the sensitivity or expressive range of mechanics as measurement instruments. Just like a 5-point Likert scale can only support operationalisations that rely on the selection between five ordinal values, and endless runner game with a single jumping mechanic delimits measurement to timed button presses in response to on-screen events. Combining this with the Rational Game User Model, I derive three systematic principles for games that enable and motivate participants to generate valid human-subject data. I summarisingly refer to these principles as the *Intrinsic Elicitation* approach to data collection game design 4.2:

1. **Necessity:** Players only engage in data provision that changes game states. *Requirement:* Embody data provision tasks as the game's interesting mechanics.

2. **Centrality:** Players select mechanics to maximise utility. *Requirement:* Make data-providing mechanics as strategically central and effortless as possible.

3. **Veracity:** Players actuate mechanics to maximise utility. *Requirement:* Where honest responses are needed, ensure that they have the highest strategic utility and lowest effort, or at least the same utility and effort as any other available option.

In summary, Intrinsic Elicitation states that data generation should be integrated into the game's mechanics such that responding honestly is the necessary, strategically central, most in-game advantageous and least effortful choice for pursuing the game's goal.

4.3.1 Necessity

A rational player will only engage with game mechanics, because these are the only means to affect the game state and thus approach the game's goals. For the rational game user, any activity not related to mechanics only increases extrinsic disutility, unless it constitutes metagaming (Sicart, 2014) like making the opponent nervous or satisfy extrinsic utilities like social norms. That is, the model acknowledges out-of-game actions, but suggests that designers can most reliably steer data provision through in-game mechanics. For data provision to occur, it must be instrumental in a mechanic. That is, it must be impossible to actuate the given mechanic without supplying the desired kind of data. Furthermore,

the game mechanics themselves must be part of enjoyable game loops – actuating the mechanic in the pursuit of game goals should be an interesting challenge or decision that elicits experiences of curiosity, competence, achievement, and the like. If acting rationally to win isn't enjoyable, players are less likely to do it, or do it well (Deterding, 2015).

Take *The ESP Game* (von Ahn & Dabbish, 2004) as an example. Two players each try to provide input that matches the other player's. Each round, the game provides both players with an image and prompts them to submit a written image label that they think the other player would use. The mechanic here is 'submit label', which is actuated by typing one or more letters and hitting the submit button. Without doing so, and the game state doesn't change. Game mechanic and data provision are one and the same, and mind-reading others is an inherently interesting challenge. This Necessity principle restates earlier suggestions that data collection games should "match mechanics to purpose" (Galli, 2014; Jamieson et al., 2012; Tuite, 2014). For example, a game for assessing fluid intelligence should involve mechanics and goals whose successful accomplishment requires fluid intelligence, like *Portal 2* puzzles (Foroughi et al., 2016). A game eliciting people's preferences in vacations should involve mechanics whose actuation expresses preferences, e.g. ranking photos of vacation places.

A corollary is that if data can be provided in different kinds that require different levels of effort, actuating the mechanic should *at minimum* require data of the target kind, not data of a kind that requires less effort. For example, if we want to collect data about people's pronunciation of words with a game where they steer a plane by speaking, the game needs to be able to recognise and require actual spoken words. The mechanic should not be actuated by e.g. volume or pitch alone, as producing humming and nonsense sounds of different volume and pitch is likely less effortful than thinking of and speaking a large variety of actual words.

4.3.2 Centrality

Often the data-providing mechanic is not the only mechanic available in the game. This may be because the game designer wants to offer meaningful choice between different courses of action, interesting variety, or because the data-providing mechanic and surrounding loop are not inherently enjoyable and therefore require additional loops and mechanics that are enjoyable. In any case, the rational player will at any step select to actuate the mechanic with the greatest current perceived in-game utility. Therefore, the

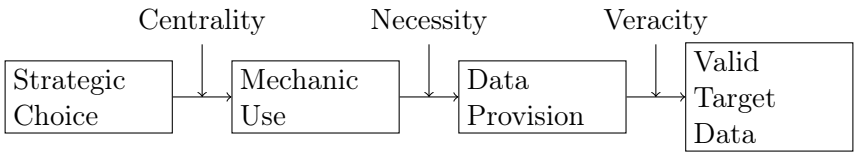


Figure 4.2: Intrinsic Elicitation

data-providing mechanic should be *central* to gameplay, either because it is the only or core mechanic the player necessarily actuates over and over (submitting words in *The ESP Game*, running in *Super Mario Bros* (Nintendo, 1987)), or because it is the strategically optimal choice in the majority of situations. Where this does not hold, the rational game user will spend the majority of their time on other mechanics, making the game inefficient.

4.3.3 Veracity

When actuating a data-providing mechanic, people choose how to actuate it: they have to pick a datum out of a set of possible datums. The rational game user will provide the datum which maximises their total utility. Where utility is constant, they will provide the datum that requires the least effort. Where effort is constant, they will provide the datum of the highest utility. This has different implications for different kinds of human-subject data. Where we are interested in assessing a participant's *aptitude* as expressed in a maximum performance (such as fitness, range of vocabulary, IQ), the design is relatively straightforward: maximum performance should maximize in-game utility. The long jump in athletics is such a simple applied game for learning the maximum jumping distance of participants. The game's mechanic (jumping) is necessarily integrated with the to be provided data (jumped distance), and as its sole mechanic, it is strategically central to the game. The required datum for actuating the mechanic is a jump of any distance. If the player were only rewarded for jumping, no matter how far, they would rationally minimise their effort and jump as short as possible – they would game the system. Therefore, the Veracity principle requires that the in-game utility of jumping further needs to continually increase and do so in excess of the required additional effort to motivate players to make an honest effort to jump as far as they can. The player could still jump dishonestly, but in that case would cheat themselves out of their own utility.

But what about revealing subjective attitudes, values, preferences, or spontaneous inclinations? Here the Veracity requirement flips into a cautionary principle. For these kinds of data, designers need to ensure that all possible data – all possible ways of actuating

the connected mechanic – are of equal overall utility, that is equally strategically worthwhile/worthless and equally effortful/effortless. In such a case, there is no marginal utility to providing one datum over another. The reasons for choosing one over the other are therefore fully sensitive to the inclinations of the individual (or non-conscious variables under study). For instance, assume a party game designed to reveal personality traits like agreeableness. Each round, players draw a card describing a social situation and three further cards describing various personality traits (agreeable, neurotic, etc.), asking the player to choose one trait to act out. The game designer wants to assess personality by observing which of the three trait card a player spontaneously chooses to act out. Assume further that the other players decide whether to give the acting player a point for their performance or not. If they are instructed to give points based on how much they *liked* the performance, acting players will be strategically biased to choose the most likable trait to act out. If certain traits are systematically more difficult to act out than others no matter one's preference, actors will be biased not to choose those. Only if choosing one trait card over the others has no such marginal utility or disutility will it be an honest signal of the players' spontaneous inclinations.

As a rational game user, the player will only maximize marginal in-game utility to the extent that doing so doesn't incur larger marginal costs in disutility or extrinsic utility, which can be connected to honesty. For example, if truthfully revealing one's sexual preference in a game of *Truth or Dare* is perceived to be highly socially undesirable, even if doing so would earn more points for one's team, the rational game user will be more likely to lie or choose a dare task. Contrariwise, the *alibi function* of games (Deterding, 2017) may enable more honest responding: as players can claim plausible deniability ("I didn't *want* to do it, I *had* to do it to win"), this lowers the perceived disutility of acting in accord with one's spontaneous inclinations, even if they are thought to be socially undesirable.

4.4 Evaluation

The most immediate use of Intrinsic Elicitation is as an heuristic evaluation tool to assess game designs for data elicitation. To demonstrate this, I will discuss three games, chosen because they attempt to elicit data in diverse ways using different reward mechanisms. The first two games, *Apetopia* and *Befaced*, are elicitation games that were first introduced in chapter 1 and are now considered in depth here. These are interesting because they have similarities to intersubjective consensus and solution-based games respectively. The third

game is *Urbanopoloy* (Celino et al., 2012), an applied game for data collection. *Urbanopoloy* is not an elicitation game, but gives an opportunity to test the wider applicability of the Intrinsic Elicitation model.

4.4.1 Apetopia

Apetopia (Barthel, 2013), first briefly introduced in Chapter 1, is a game that collects data about how individuals perceive the relative similarity of different colours. This is used to construct a model of colour similarity in human perception. While a single consensus model is desired, this game arguably achieves this by collecting individuals' human-subject data (i.e. about their individual perceptions, not their expectation of the consensus). This is itself not unusual, as it is common to aggregate the data of multiple participants to make conclusions about the population from which they were drawn. However, this game is interesting because it includes an agreement mechanic for rewarding or punishing players for the data they provide while on the other hand this mechanism is sufficiently obscure that casual players would likely never know that it exists. Thus in *design* it is similar to an intersubjective consensus game, like *The ESP Game* (von Ahn & Dabbish, 2004), while in *experience* and perhaps thus in the data it elicits it might arguably be considered an elicitation game, collecting human-subject data about the individual player.

In *Apetopia* you speed through a gritty urban environment moving left or right to pass through coloured gates, dodge bombs and heaps of rubbish, and to collect coins to increase your speed (Figure 4.3). The goal of the game is to travel as far as possible. The game ends if you lose all of your health, which happens when you collide with obstacles. The graphics are rendered in a gritty black-and-white pencil drawn style. The only colours (except that of coins) are the colour of the sky and the colour of two gates that periodically appear in your path, forcing a choice between them.

The game can be understood as presenting a series of trials within a seamless infinite runner. Each trial presents the player with a choice between two colours in the guise of choosing between a pair of coloured gates. A reference colour is presented as the colour of the sky and the colour of a flag shown between the gates. The player can choose a gate by moving left or right using the direction keys before they collide with the gates. As the player moves forward at a constant rate, the player has a limited time to decide. Between the two gates there is a wall. If players collide with the wall the trial ends without them choosing a colour, but they lose some health. At the beginning of each trial the sky colour

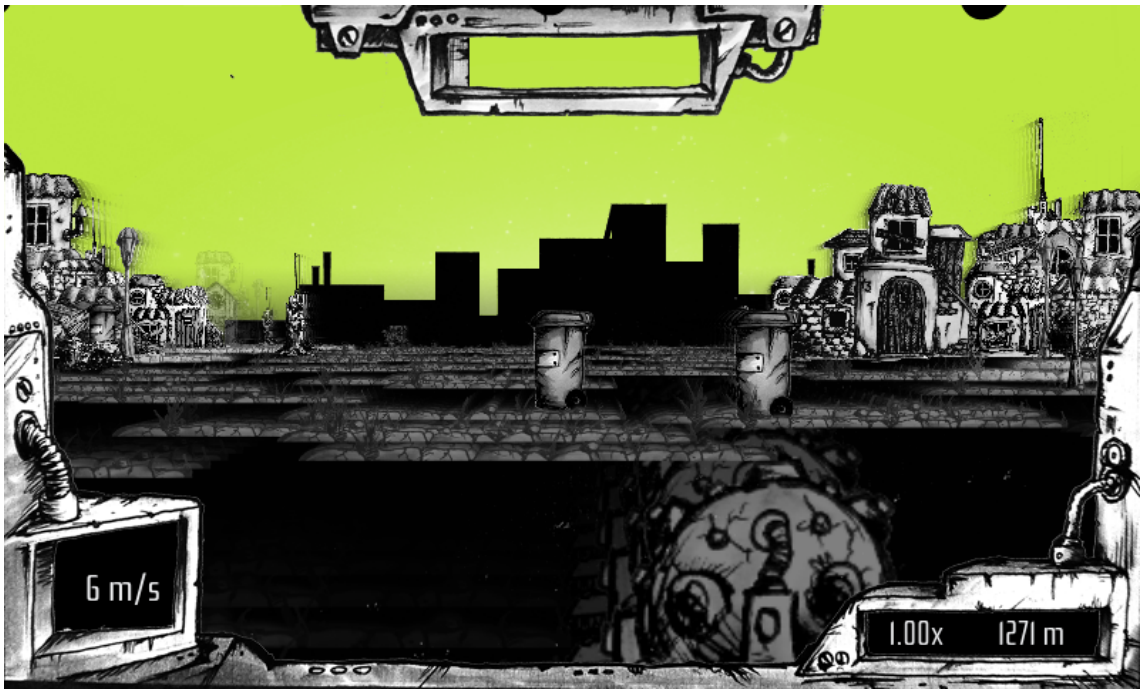


Figure 4.3: During play of *Apetopia*, players must dodge bombs (foreground) and other obstacles (centre midground). The players health bar is displayed at the top of the screen. At the bottom of the screen is displayed their current speed (left), the current multiplier (right) (it is unclear what this does) and distance travelled (far right).

is grey. The player must pass several obstacles and has the opportunity to collect coins, while the sky fades to a colour for the trial. After this, no more obstacles or coins appear so by the time the two gates appear there are no obstacles to avoid.

The instructions, presented at the beginning of the game (Figure 4.4, direct you to choose the coloured gate that most closely matches the colour of the sky. The instructions do not describe what happens when moving through a gate. The instructions merely state that "some gates are magic, they can boost or repel you" (Visual Computing, HTW Berlin, n.d.), without explaining why this is. Another reviewer understandably makes the mistake that the game does not reward choosing one gate over another (Schrier, 2016, p.42). However a reward does exist whenever there is a sufficiently strong consensus from previous presentations of the colour to other payers (Barthel, 2013). If a player agrees with such a consensus they are boosted forwards so for a short time so they move much faster. When they disagree with the consensus, they are punished by their distance score being decreased meters and with the loss of health. As these outcomes happen only when consensus of players is sufficiently strong, it happens only occasionally. It seems highly unlikely that a casual player would determine the mechanism involved.

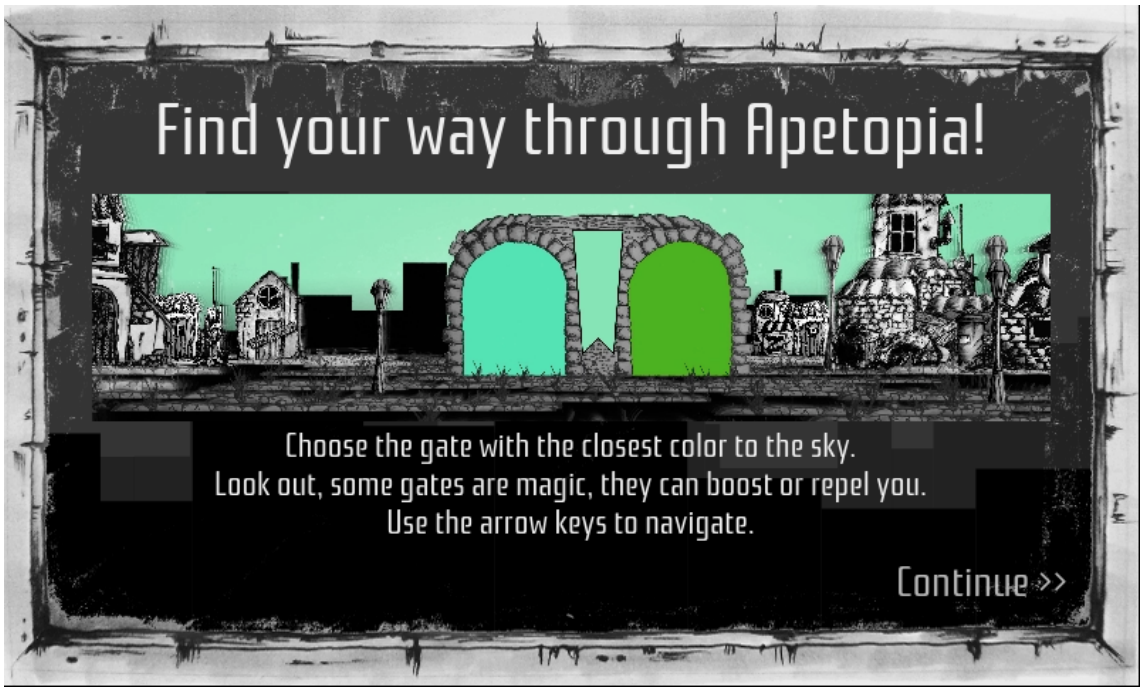


Figure 4.4: In *Apetopia* the player uses the arrow keys to dodge obstacles and to choose between differently coloured gates (a). The instructions at the start of the game tell the player to "choose the gate with the closest colour to the sky", and warns that some gates "can boost or repell you" (b).

4.4.1.1 Necessity

The Necessity principle requires data provision to be the unavoidable result of actuating one of the game's mechanics. The data provision mechanic is choosing a gate, which is actuated indirectly via the mechanic of moving left and right. Because the player is always moving forward, the movement mechanic leads, most of the time, to a gate being chosen. However, picking a gate is not in fact necessary. It is possible to collide with the centre wall that joins the two gates. If the player does this, neither gate is picked and the game continues as normal. Colliding with this wall does cause the player to lose health but on the other hand the gates unpredictably impose a penalty (occurring the player disagrees with the consensus). Is a player likely to choose to avoid providing data via this route? Such an uncertainty only exists due to a violation of Necessity. However, we can help to resolve it by modelling a rational player.

Colliding with a wall imposes a penalty, yet passing through gates sometimes imposes a penalty (which a player likely would not know the cause of). To determine how likely it is that a player would, in order to avoid the larger penalty of picking the wrong gate, take the lesser penalty of colliding with the centre wall, we can consider the virtual utility of

Gate	**Speed** (+)	**Multiplier** (+)	**Health** (+)	**Distance** (+)	**Mine rate** (-)
Agrees	0	+0.05	+0.05	Boost forward	+0.2
No Consensus	0	+0.02	+0.02	0	+0.2
Disagrees	×0.75	×0.75	−25	−200	+0.2
Centre Wall	×0.75	×0.75	−25	0	0

Table 4.1: Rewards and penalties in *Apetopia* by whether or not the gate selected agrees with the consensus, or whether the player collides with the centre wall. Each column is labelled with whether it is in the rational players interest to maximise (+) or minimise (-) that property. These values were extracted from the Visual Computing, HTW Berlin (n.d.) source code.

doing so. The costs of each gate option are given in Table 4.1, where it can be seen that choosing the incorrect gate is indeed objectively worse than colliding with wall. However, we must consider the *expected* utility, given the uncertainty whether a choice of gate would be correct or incorrect. At the rates I observed, there was approximately a 15% chance of picking an incorrect gate at random. Thus, as health is recovered from coins at a very slow rate (1 per coin) and only a handful will be picked up between gates, the gate-avoiding strategy would not be adopted by a rational player. Thus, despite the violation of Necessity, we would still expect the rational player to provide data rather.

When playing *Apetopia*, Schrier (2016) found that the game stopped presenting gates, yet the rest of the game continued. Here it was possible to continue using the movement mechanic without providing any data. I suspect this was due to a connection issue with the server, as the next trial to show is requested following the completion of the last trial. This technical issue also led to a failure of Necessity, as the movement mechanic no longer could encode a player's choice. While such technical failures might seem tangential to the game design, they are also a common occurrence with delivering games online. As such there are choices to be made about how to handle such a situation. For example, the game might generate new trials internally and save up data on the assumption that the connection will be later reestablished, maintaining Necessity.

How would the principle of Necessity suggest the game be redesigned? First and obvious changes would be to remove the option of colliding with a wall instead of choosing a gate, and ensuring the server was always able to send new trials to display. However, more interesting changes might be suggested also. Currently data provision is only indirectly necessary from the use of the movement mechanic. This has been shown to lead to two cases in which data is not provided (either wall collision, or when no trials are downloaded).

Necessity would suggest that *all* actuations of the mechanic should provide data. To achieve this with a similar movement mechanic, one option would be for the colours of the gates to be always shown on the sides of the screen. The sky could then continuously change colour, and the player would need to move in the direction of the best match moment-by-moment. While this would cause issues with obstacles (leading to violation of Veracity as players are biased to dodge obstacles instead of reveal their true colour perception), it would make data provision a fully necessary, integrated part of the movement mechanic.

4.4.1.2 Centrality

Much of the game is centered around dodging obstacles and collecting coins and not with moving through gates. The gates appear approximately every 7 seconds, and it is only when going through the gates that the player provides data. The data providing mechanic is reasonably frequent, but is not the central mechanic in the game: neither in the strategy (collecting coins is more effective), nor in use (you collect several coins and dodge several obstacles between each set of gates). Even if one data point per 7 seconds is considered sufficient, the gate mechanic still feels unnecessary and adjunct to the rest of the game experience, harming player enjoyment.

What would the centrality principle suggest for redesigning the game? By making the choose gate mechanic more frequent, it would increase the rate with which data is collected. For instance, instead of dodging obstacles and collecting coins, the core activity of the game could be made to be navigating through gates. This would, however, show up the main failing of the game: that it is not clear what the gates do and thus choosing between them is unsatisfying.

4.4.1.3 Veracity

According to Veracity, providing the desired kind of data should have the highest overall utility, or at least not a lower overall utility than any other choice; in *Apetopia* this means selecting the gate that corresponds to ones honest perception of the closest colour. The presence of the agreement mechanic means that the reward for honestly choosing a gate is unreliable. Most of the time an honest judgement gets no greater reward than a dishonest choice, as both gates give the same reward. Sometimes honesty gives a greater reward, but only if the player agrees with the consensus. However, I also found occasions were my honest attempt was punished for not matching the consensus. Still, if it is assumed that

players' honest judgements are statistically more likely to match than to disagree with the consensus, overall the honest response strategy gives the highest utility.

There is a significant exception to this however. This is based on the assumption that the player chooses between honest, dishonest, and random answering strategies. The player may adopt a different strategy, depending on their mental model of the reward mechanism. Once I had determined that reward was based on agreement, I found myself explicitly considering what I thought the consensus would be, rather than my own judgement. In terms of the Rational Game User Model, this became a factor biasing my selection of game actions. Responses that I believed to match the consensus had an increased (expected) virtual utility, due to the higher probability of consensus. This becomes a problem for Veracity if it is ever the case that my honest perception could disagree with this consensus. Plausibly this could occur if a player finds, for example, that the consensus prioritises hue over saturation. Their own judgement that two colours are similar because they are similarly saturated, despite a small difference in hue might be overruled by recognising that the consensus will likely judge the difference in hue more severely than the difference in saturation. While this is only speculation, it illustrates the problem with using an agreement mechanic for driving rewards in an elicitation game.

An alternative to partially address this would be to drive the reward mechanism based on consistency within the player's own choices. The same pairs of colours could be presented multiple times throughout the game. After the first presentation, to get the reward the player would need to choose the same option. While the player might at first merely need to remember what they chose last time, this would soon become impractical once the player has seen enough colours. So long as there are no reliable external cues to base a decision on (e.g. relative positions of the colours), the rational strategy for the player would be to decide based on the internal cue of their own perceptual judgement, as this would be the most reliable signal to the consistent answer without remembering. Of course, this would not stop the player choosing to always e.g. invert their answer so they provide the opposite of what is desired. Nor is it clear how long a player would need to play for (or perhaps in how many separate sessions a player would need to play) to minimise memory effects.

4.4.1.4 Summary Evaluation

Apetopia is an interesting attempt to make an elicitation game that drives feedback (occasionally) using an agreement mechanic. The principle of Necessity highlighted that the player did not in fact need to provide data. There is no obvious reason why this should be. With a redesign of the gate image to remove the inner wall, it could be made fully necessary. Adopting the principle of Centrality showed that providing data could be made more central to the game by increasing the frequency of gates, though this would highlight a core failing of the game that moving through gates is confusing and unsatisfying. Indeed, the fact that other game mechanics were put into the game may be precisely because this core mechanic was not satisfying by itself. The Intrinsic Elicitation approach would suggest that redesigning to find a satisfying core mechanic be considered. Finally, Veracity found that the agreement mechanism had the potential of biasing the motivations of players and thus the data provided if they became (consciously or unconsciously) aware of it.

4.4.2 BeFaced

BeFaced (Tan et al., 2014) is a camera-controlled tile-matching elicitation game for the iPad that is designed for collecting images of the players' faces and/or tracked feature point data (Tan et al., 2013) while they are making a range of facial expressions. It was first introduced in chapter 1. It is interesting as it drives feedback mechanisms by comparing player input against a model.

BeFaced uses tile-matching game mechanics similar to *Bejeweled* (PopCap Games, 2001): players swipe on tiles to swap their positions and make lines of three or more. However, while in *Bejeweled* the tiles are automatically cleared, in *BeFaced*, the player has three seconds to make a facial expression matching the icon on the tiles. This is captured in real time from the iPad's camera. A classifier checks whether the players facial expression matches the type of tile to clear. If the players' facial expression matches the target expression with sufficient probability, the tiles are removed (Figure 4.5).

The classifier constitutes a model of the correct input, similar to a 'ground truth' model in a solution-based game such as *FoldIt* (Cooper, Khatib, et al., 2010). This is able to match facial expressions that satisfy certain constraints. However, the data that is collected is about the individual – a photograph, or feature point data of their face – in addition to being a particular 'solution' to the classifier's model.

The game is not attempting to elicit spontaneous expressions of e.g. disgust or surprise

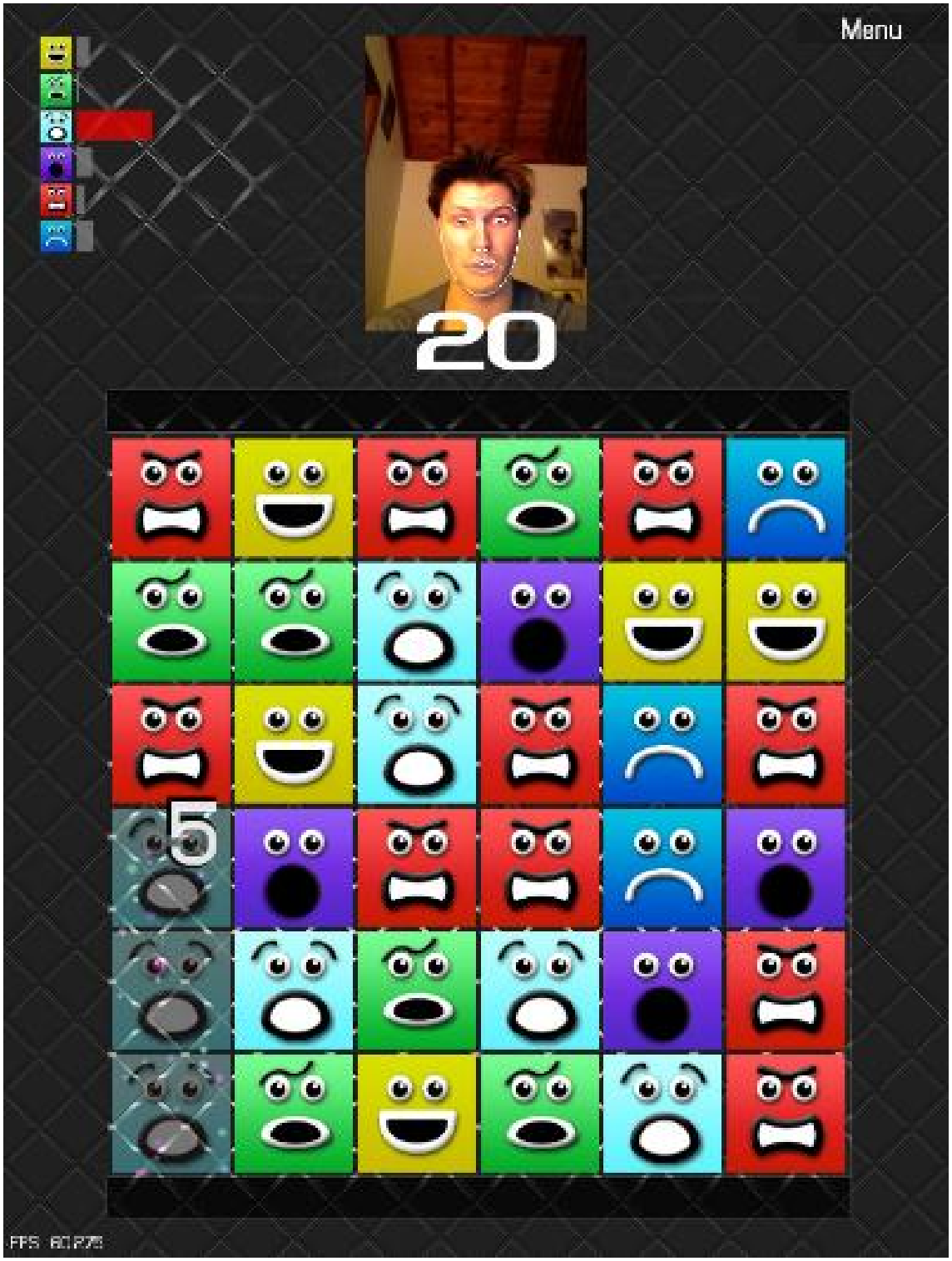


Figure 4.5: In *BeFaced* players match groups of tiles and then need to clear them by making a facial expression that matches the tiles to clear. Image from Tan et al. (2013)

(however the player would naturally produce them). Instead it aims to elicit expressions that closely match defined visual cues. Indeed, players are trained by the game to produce the desired expression: the algorithm becomes more likely to match an expression if a player has previously failed to match that expression and gradually raises this difficulty again over time (Tan et al., 2013).

4.4.2.1 Necessity

The tile-clearing mechanic satisfies Necessity in that it requires the player to make an expression to activate. While the player may choose not to make a facial expression, this is effectively choosing not to actuate the tile-clearing mechanic and the player then does not get the benefit of clearing the tiles. It may be that other images can satisfy the classifier, such as a printed image of a face held in front of the camera. If so, an actuation of the mechanic would not necessarily encode the desired data (the *player's* face), and thus Necessity would not be satisfied.

As part of training the player to produce the correct expressions the classifier becomes easier to satisfy if the player repeatedly fails to make a match. If it was allowed to become too lenient and permit non-facial expressions to trigger the mechanic, then the mechanic would no longer satisfy Necessity. A player unwilling to make facial expressions might then allow the first several tests to fail until the classifier accepts their resting face, for example.

4.4.2.2 Centrality

The tile-clearing mechanic is central to the game. While not the only mechanic, it must be used frequently: after a player has matched a group of tiles, they must make a facial expression to clear it. Clearing tiles is the only way of gaining points in the game. The alternative is for the matched tiles to be reset without being cleared – losing progress towards the goal. Using the tile-clearing mechanic therefore has the highest virtual utility when it is available. Of course, the displeasure of actuating the mechanic may still cause a player to stop playing. If a player does not choose to stop playing, we would expect data provision to be approximately as frequent as the tile-matching mechanics deliver matched groups.

I note that, while the tile-clearing mechanic is strategically central to the game there is no practical reason why this additional step is necessary and it provides the player with no meaningful choices. As such, while data collection may have been integrated into the

game, it has not been integrated into the enjoyment of the game.

4.4.2.3 Veracity

Veracity is satisfied when providing the desired kind of data gives the highest overall utility, or no less overall utility than any other option. Whether *BeFaced* satisfies Veracity depends on the quality of the classifier. Assuming the classifier is able to accurately identify facial expressions matching the desired certain type, then providing desired data (the correct facial expression) will always have higher utility than providing any other data. The alternative to the player who does not want to do so (due to the effort or displeasure involved, for example) is to stop playing.

However, Veracity foregrounds a number of concerns where less than a perfect match between classifier and desired data might be significant for data quality. First, given the effort involved in making a particular facial expression, we might expect players to converge on the lowest-effort expression that satisfies the classifier. For example, if mouth shape is the primary feature used to determine whether the player is smiling, we might expect smile-expressions to lack movement of other features: 'false smiles'. As such, the images collected may not appear genuine attempts to make a given expression. Second, the actuations (facial expressions) would not necessarily be those corresponding to the individual's natural expression of a given emotion. A player's 'angry face' expression need look nothing like how they look when actually angry. This is the intended behaviour, as evidenced in the use by the classifier of 'dynamic difficulty adjustment' – when a player fails to make the correct face, it becomes more lenient the next time to help 'train' the players how to make the correct expression. This shows that the authors expect that the expression required will often not be the natural one. These two caveats aside, the game otherwise satisfies Veracity.

Designing a game to collect photographs and feature point data of *natural*, spontaneous expressions would be more challenging due to the need to avoid such convergence to the model. Replacing the model with other players and multiplayer mechanics would not necessarily help if a player can most easily communicate 'sadness' by an exaggerated expression. One possibility would be for the game to get the players to act out times from their own lives when they have felt particular emotions. In such a game, rewards could be given for the authenticity or 'naturalness' of one's acting. So long as this reward is orthogonal to the particular way expressions are made, Veracity would be satisfied.

4.4.2.4 Summary Evaluation

Because *BeFaced* seeks to collect photographs and feature point data of players making *particular* expressions, it is able to use a classifier to detect and reward these ensuring Veracity. The design of the game makes the provision of such facial expressions necessary and central to the game. Particular areas where a mismatch between classifier and desired data would be expected to result in a decrease in data quality are where player's can make 'low effort' expressions, and when the performative expression is unlike a player's natural expression.

4.4.3 Urbanopoly

Urbanopoly is a competitive multiplayer mobile game for gathering location-based data about urban landmarks, clothed in the fiction of players being rich landlords. The goal is to have the most virtual cash by buying and trading real venues on the game map that generate different daily cash bonuses every day the player logs into the game. When players open the game, they can see and click on venues in their geographic vicinity. When a venue is free, they can purchase it. If the venue already owned, the player spins a "wheel of fortune" resulting in a venue-related task such as making an advertisement (entering venue information and shooting a photo of it), answering venue-related quiz questions, or rating photos. All these activities receive virtual cash rewards. The wheel can also trigger a chance to take the venue from the owner or paying rent to the owner. Once multiple entries are collected on a venue, a weighted majority agreement algorithm is used to identify consensus truth and trigger a cash penalty for non-consensus answers. Thus, *Urbanopoly* has nine mechanics or verbs: log in, move to venue, buy, spin wheel, answer quiz, judge photos, take venue, pay rent (figure 4.6).

4.4.3.1 Necessity

The Necessity principle requires data provision to be inherent to the mechanics changing the game state. Four of *Urbanopoly*'s nine mechanics entail data provision: purchasing venues requires buyers to enter their name and category from a predefined list; advertise (enter information and photo); answer quizzes (pick the correct datum from a list of previous player answers); and rate photos on a scale of 1-5. In each case, the game state doesn't change until players enter data. Yet the data provision is sometimes weakly integrated into the mechanics: there is no diegetic or practical reason why buying a venue should entail

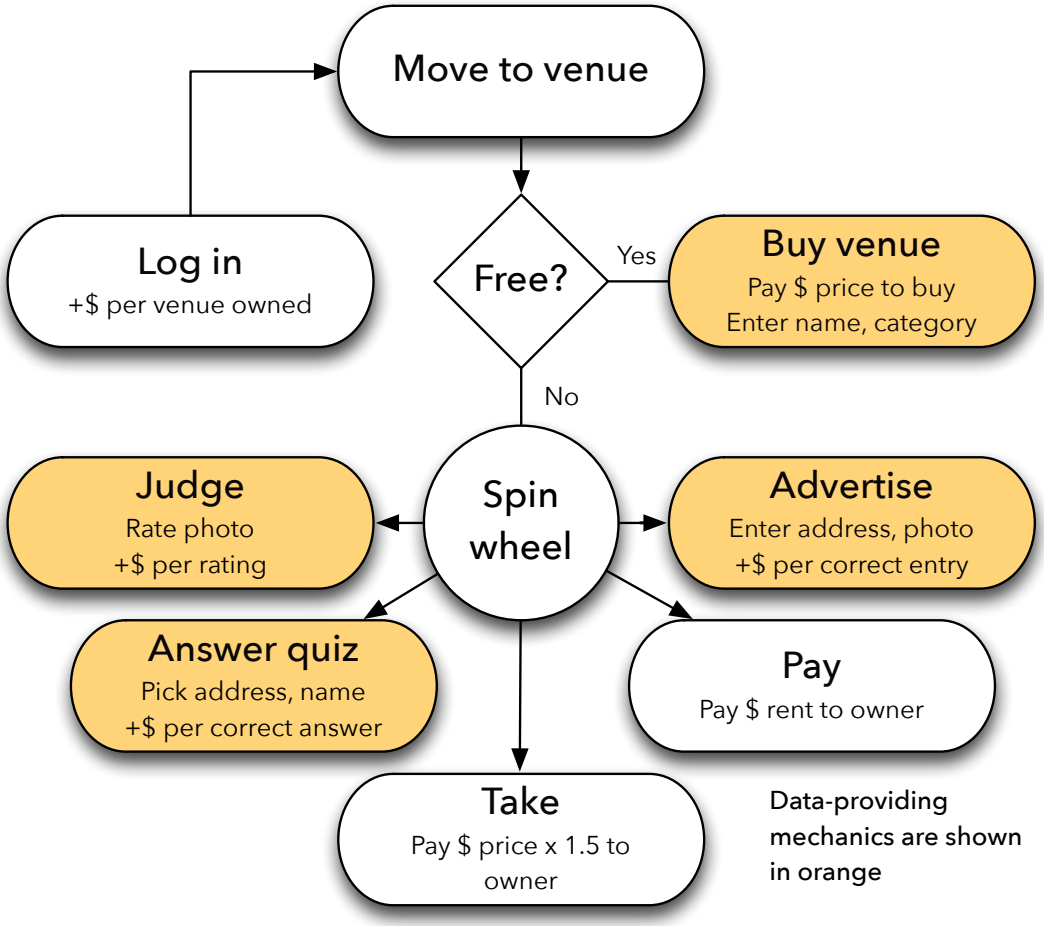


Figure 4.6: The core game loop of *Urbanopoly*.

entering its name and category – the interesting choice is which venue to purchase based on price and likely earnings. Thus, entering data during purchase does not partake in or contribute to the intrinsic utility of game enjoyment. Contrast this with answering quiz questions, where guessing the answer and providing verification data are one and the same. Rating a photo, finally, is data provision as a game mechanic, but barely an interesting core loop: there is no challenge or interesting decision to expressing one's preference, as any rating is equally rewarded. Finally, the daily log in and rent payment allow players to advance towards their game goal (earning cash) without entering any data.

4.4.3.2 Centrality

The presence of such "pure gaming features" (Celino & Cerizza, 2012) isn't problematic as long as data-providing mechanics are strategically central to winning the game, which

makes it optimal to spend the majority of playtime actuating them (Centrality principle). Translated into *Urbanopoly*: Are the data-providing mechanics the best time investment for virtual cash earned? Given the fact that players need to physically move to venues to actuate them, and that verifying or entering data may require checking house numbers, street names, and the like, each game action comes with non-trivial time and effort. Here, the game shares one major downside with its inspiration, *Monopoly*: once a player owns a range of properties, checking in daily and receiving rent payments from other players become a steady cash source compared to which the effort of data-providing actions seems hardly worth it. The less venues a player owns, the more rational it is to invest time into selecting venues and spinning the wheel, as only this allows to acquire cash through data-provision tasks (which are more likely to show on the wheel than the pay rent option) and take venues from their owner. Here, the Centrality principle can guide the balancing of virtual cash rewards per mechanic: these should be arranged so that the data provision most desired by the system designer delivers the best cost/benefit ratio.

4.4.3.3 Veracity

The Veracity principle states that providing the desired kind of data should have the highest utility and the least effort. Here, *Urbanopoly* faces the challenge that it cannot distinguish honest from dishonest entries at the time at which initial data is provided, as it relies on multiple entries on the same item to derive a consensus truth. This means that (in the short term), it is rational for players to game the system and enter as many data points as fast as possible regardless of their accuracy, as every entry is rewarded equally. However, players are warned that their 'karma' will penalise them if their entries are later found to be incorrect. This phrasing calls on meaningfulness and social norms, but it also refers to the fact that players whose entries lie outside the consensus will receive a cash penalty later. This is arguably a good first step to ensure Veracity, but relies on players understanding said delayed penalty. However, it misses out on optimising for broad coverage as a data quality. Players receive the same amount of reward for doing the same data-providing activities repeatedly for the same places. And since venue density varies geographically, a rational player will move to and provide data on the same close-by, densely packed venues again and again.

4.4.3.4 Summary Evaluation

Analysing *Urbanopoly* through the lens of the three principles of Intrinsic Elicitation immediately foregrounded several design shortcomings. Evaluating for Necessity showed that entering data is weakly integrated into the buy venue mechanic, likely causing player frustration not enjoyment. As for Centrality, the more venues a player owns, the less central data-providing mechanics become to them, suggesting to rebalance the cash payouts per mechanic. The biggest takeaway came from the Veracity principle. While introducing a cash penalty for non-consensus answers should prevent careless responses, the current game design makes it rational to revisit the same cluster of close-by venues constantly, rather than providing new data on venues not yet covered.

4.4.4 Discussion

Here I have analysed three games using Intrinsic Elicitation: *Apetopia*, *BeFaced* and *Urbanopoloy*. Using Intrinsic Elicitaiton highlighted for each potential design concerns and helped to suggest some potential solutions.

In all three games, the data proving mechanics largely satisfy Necessity, but these mechanics do not always contribute to the intrinsic enjoyment of playing the game. In *Apetopia* the choice between coloured gates interrupts the apparently core gameplay of collecting coins and dodging bombs. In *BeFaced* the tile-clearing mechanic is merely a hoop for the player to jump through, offering no meaningful choices to the player. In *Urbanopoly* entering data about a property to purchase is weakly integrated into the mechanics, in comparison to answering quiz questions, where the data provision is integrated into the mechanics.

BeFaced and *Apetopia* make their data-providing mechanic strategically central by not providing any other meaningful option. In *Apetopia* the player is forced to make a choice (or run into a wall, which was seen to not be preferable) by the relentless forward movement. This forward movement is a core to the challenge of the game. In contrast, in *BeFaced* the player is required to use the mechanic to progress in the game or score points though there is no practical reason for this beyond the data collection itself. *Urbanopoloy* on the other hand has several mechanics and the player can to some extent avoid data provision by simply engaging with other parts of the game.

Finally, Veracity helped to highlight potential issues in the validity of the data of each game. In *Apetopia* it revealed the potential for agreement mechanics to bias individual's

data towards the consensus while agreement mechanics in *Urbanopoly* suggested that the potential of delayed penalties might not reliably bias the data towards consensus enough. In *BeFaced* I used it to suggest areas where a classifier might potentially not work as intended and allow players to minimise effort to the expense of the desired data.

4.5 General Discussion

Current applied games for data collection are dominated by two basic templates – *GWAP*-style classification games and *Foldit*-style solution discovery games. For both, the literature provides a general design approach of gamification+validation: data volume is motivated with particular game design elements, data quality is then separately ensured with validation strategies. While popular, I noted that this approach falls short when it comes to human-subject data, as these often involve no subject-external ground truth to validate against, which the gamification+validation approach requires. Furthermore, by making certain in-game responses more enjoyable or strategically opportune, gamification strategies may threaten validity by biasing or overshadowing 'spontaneous' expressions of preferences, attitudes, or beliefs.

In response, I articulated the need for broader elicitation approaches that *integrate* motivation and data quality at once. Inspired by the model of speech motivation developed in chapter 3 and based on literature on data collection games, crowdsourcing, survey response, and citizen science, I developed a rational choice model explaining why players choose to provide certain data in games, the *Rational Game User Model*. Building on and incorporating J. H. Smith's (2006) Rational Player Model, it predicts that players choose in-game actions providing data that maximise their overall utility, comprising three factors: extrinsic utility such as social norms and monetary incentives; extrinsic disutility such as opportunity costs and displeasure of the effort of providing the data; and intrinsic utility, comprising meaning and game enjoyment. Game enjoyment in turn is maximised by the player trying to maximise their in-game utility, i.e. choosing the course of action that maximises their odds to win.

From this model, I then derived three principles for designing human-subject data collection games I summarisingly call the *Intrinsic Elicitation* approach. A good data collection game integrates data generation into its enjoyable mechanics and core loops such that it is (a) necessary to actuate the mechanics, (b) strategically central to gameplay, and (c) providing spontaneous or honest data has the highest utility and lowest effort, or at

least equal utility and effort compared to all other available options.

Finally, I illustrated the value of this approach by using it as a heuristic evaluation tool for two existing elicitation games *Apetopia* and *BeFaced*, as well as the existing data collection game *Urbanopoly*.

4.5.1 Limitations

The simplicity of the model is both strength and weakness. It allows us to generate clear predictions to test, falsify, or refine the model. For instance, it predicts that players will abandon the game when the perceived overall disutility is greater than the perceived extrinsic and intrinsic utility. Similarly, if the perceived marginal loss of meaning due to a strategically optimal but dishonest in-game action is greater than the perceived marginal gain in game enjoyment, we can predict that players would refrain from choosing it. In terms of limitations, I fully expect that in the course, several assumptions will need to be complicated. First, behavioural economics demonstrates that people's rationality is bounded by biases and heuristics (Camerer & Loewenstein, 2004). Second, we know that players gain game enjoyment from more than just playing optimally: curiosity, surprise, engrossment, relatedness are other important sources (Boyle et al., 2012). For instance, a key aspect of good game design is giving players meaningful choices. Dominant strategies (where there is only one rational move) are not a concern for the rational player, but rational game users may feel like they lack autonomy or are not making any meaningful choices in this situation, reducing their overall game enjoyment. Against the Centrality principle, game designers may therefore need to introduce alternative mechanics that are sometimes the preferable choice so as to make choosing the data-providing mechanic a non-trivial choice. Third, I suspect that there may be systematic causal effects between individual factors that I didn't specify – e.g., self-determination theory would predict that adding tangible rewards may undermine intrinsic game enjoyment (Deci & Ryan, 2000). Designers may wish to use their own more nuanced knowledge of player behaviour to violate some of the principles.

Beyond these conceptual limitations, it may not be possible for every kind of data to be integrated into a game in such a way that satisfies all aspects of Intrinsic Elicitation. For example, it might be technically impossible for the game to distinguish target from non-target data – in a language elicitation game, natural language processing algorithms may not be fast or sophisticated enough, for instance. Following the model, however, it is

possible to suggest implications for such limit cases. In the example, we would expect noisy data as players produce both target and non-target forms. If some forms are more effortful to produce than others, we can expect that the collected data will be biased against those forms, and we may want to use statistical techniques to identify and counteract this bias.

4.6 Conclusion

This chapter presented a framework for the design and analysis of elicitation games – games to collect human-subject data – for which I argued that existing gamification+validation approaches were not suitable. This included a model of the player – the Rational Game User Model – and a design approach that I called Intrinsic Elicitation.

However, the question remains whether, even adopting this design approach, enough data of sufficient quality would be collected for practical use. We might expect that the many potential sources of variance in games would add potentially significant noise. In particular, if we were to design a game to elicit an individual's spontaneous inclinations, can these in fact be detected given the noise of the data? Empirically testing the framework (or a game developed using it) against a practice-as-usual control is the best way to evaluate its effectiveness. Data from a practice-as-usual control can be considered the gold standard. If the quality of data collected by the game is lower than that of the experiment, the next question to ask would be what additional factors need to be incorporated into the Rational Game User Model. Finally, it is also necessary to check that such games are indeed enjoyable when designed following the Intrinsic Elicitation. Does integrating data provision into the core mechanics and loops of the game, as suggested by this design approach, lead to enjoyable games?

In order to do this, in the next chapter I will introduce a novel elicitation game – designed using Intrinsic Elicitation – which I will use through the rest of the thesis. There I report the first two of four empirical studies using this game. In these studies I test this game against a practice as usual control while observing effects on data accuracy and enjoyment.

Chapter 5

Trading Accuracy for Enjoyment

Having developed a framework for the design and analysis of elicitation games in the previous chapter, here I begin to explore it empirically. So far in this thesis, we have dealt with validity and (moment-by-moment) motivation. Now a shift in language is required from *validity* to *data quality*. As validity relates to the overall argument that is made in support of a conclusion (Messick, 1995) it is awkward to speak of operationalising validity as a measureable quantity. Data quality refers to "fitness for use" (Tayi & Ballou, 1998), which in human participant research centrally involves validity. Data quality has variously been operationalised in the literature. Here I will operationalise it as accuracy. Of course, the accuracy of elicited data is a part of its validity – finding factors that typically increase or decrease accuracy in game data will no doubt affect validity claims and judgements. The purpose of the change is to highlight that hencefourth we will be dealing with a quality of *data*, rather than a quality of an *argument*. Other common aspects of data quality raised in human subject research are satisficing and careless versus careful responding, study completion versus dropout, or missing/dropping response items; these can be framed as behavioural engagement or validity threats (Hawkins et al., 2013).

The first empirical question to answer about the Intrinsic Elicitation model is to what extent is is sufficient for collecting accurate data from its players while maintaining enjoyment. After all, the use of applied games for data collection in general is grounded in the dual premise that games (a) are more enjoyable than standard methods while (b) producing data of similar if not better quality (Hawkins et al., 2013). Enjoyment encompasses different conceptualisations of positive user experience (flow, intrinsic motivation, etc.) that are believed to drive behavioural engagement (Reeve, 2014), e.g. playing more games, providing more and more complete data, exerting more effort and care in respond-

ing. If enjoyment is improved only at significant cost to data quality, this points either to a limitation of the model or a trade-off inherent in the use of elicitation games.

Empirical research directly testing these premises (game design brings (a) higher enjoyment and (b) equal or better data quality) has been largely limited to *gamified* online surveys and experiments – studies that use presumed-motivating design elements from games, rather than creating full-fledged data collection games (Deterding et al., 2011). This research shows mixed results. While some found that adding game design elements can increase both data quality and quantity (Van Berkel et al., 2017), or collect data of equivalent quality and quantity, but with more enjoyment (Friehs et al., 2020; Hawkins et al., 2013; Levy et al., 2016), recent reviews of gamified surveys (Keusch & Zhang, 2017) and assessments (Lumsden et al., 2016) suggest that gamification tends to improve the user experience (i.e. enjoyment), but not necessarily impact behaviours such as satisficing, omitting items, or abandoning surveys, and with those, data quantity and quality.

However, similar empirical work on elicitation games or other full-fledged data collection games has been missing. In my literature review, I found one case study (Crowston & Prestopnik, 2013) and one comparison between a gamified and full game variant of the same citizen science data classification task (N. Prestopnik et al., 2017), both suggesting that the full-fledged game produces similar engagement but lower-accuracy data. But either study features no real experimental control.

This lack of high-quality evidence on the data validity of data collection games matters, as discussed in chapters 2 and 4, games *as games* generate new, systemic validity threats, especially for human-subject data: the social norm and empirically observed reality of entertainment gaming is that players ought to voluntarily participate for the sake of (mutual) enjoyment, and to this end, relegate game-external consequences and concerns (Deterding, 2013) and make more or less rational, strategically optimal moves (J. H. Smith, 2006) – modulated by other norms like maintaining good relations with the other players (Juul, 2008). This suggests possible trade-offs between engagement and data quality: if players are playing a game *as a game*, they should not distract themselves with game-external concerns like answering honestly and carefully; and if they are answering honestly and carefully, this may diminish the enjoyment of getting fully engrossed in the game. Put differently, the most strategically optimal or fun in-game action need not be the most honest and considered response out-of-game. To give a practical example for latent subjective properties: imagine we design a charade-style social guessing game to elicit people's pre-

ferred ice cream flavours. If people really 'get into the game,' they may claim that they like chocolate even though they actually prefer woodruff – because chocolate is easier to mime and guess, or because the previous person already mimed woodruff and repeating them would be boring. We would not expect similar effects from a gamified survey that was still framed and approached *as a survey*, but e.g. clothed into a nautic theme, juicy feedback, or additional game mechanics that don't connect to data collection (Levy et al., 2016).

If data collection games were indeed prone to suffer from lower data validity, especially for human-subject data, this would add an important caveat to their current popularity, and suggest that research and practice should look into design strategies for improving data quality and validity. In response, I decided to test *how the enjoyment and data quality of a human-subject data collection games compares to a practice-as-usual control.*

To this end, I designed *Adjective Game*, a browser game to elicit adjective order – the order in which people use adjectives to describe a phenomenon, like "big black cat". Established experimental methods for eliciting people's adjective order offer a good practice-as-usual control. Further, while individual grammatical intuitions about adjective order are a latent subjective property, they are highly predictable for adult native speakers, which provides a rare opportunity to compare data from game and control conditions to an approximate 'ground truth'. This allows us to perform the present research into accuracy of player-provided data while using a game that can claim ecological validity as an elicitation game, as justified below. The design of *Adjective Game* followed the Intrinsic Elicitation design approach (chapter 4).

I conducted two preregistered studies ($n = 96$ and $n = 136$) that compared this data collection game with an equivalent linguistic experimental setup.[1] Each study operationalised accuracy differently, mirroring two different paradigms in linguistics. In both cases, the game proved significantly more enjoyable than the control, but also produced significantly less accurate data.

The rest of this chapter is structured as follows. Section 5.1 will introduce the linguistic case study, adjective order and picture description tasks. Section 5.2 presents the design of the game, how it follows the Intrinsic Elicitation model, and the picture description control task. Sections 5.3 and 5.4 report Studies 1 and 2, followed by a general discussion in section 5.5. Section 5.6 concludes the chapter.

[1] All materials, code, data, and preregistrations for both studies can be found at https://osf.io/jac6s/

5.1 Background

5.1.1 Adjective Order

The order of words we use expresses shared rules of grammar. For example, in reference to a sea-going vessel that is large and scarlet, an English speaker would say it is a "big red boat", but not (in a neutral context) a "red big boat". This is an example of adjective order. People's intuitive judgments about adjective order are strong and reliable (Bache & Davidsen-Nielsen, 2010; Cinque, 2010; Sproat & Shih, 1991). Different paradigms in linguistics offer different explanations for this fact: a *generativist* would take it as evidence of shared innate universal rules (Chomsky et al., 2019; Cinque, 2002). In contrast, a *constructivist* would argue that people's individual grammatical intuitions are indeed individual, but became coordinated with those of other speakers in the course of socialisation.

Either way, linguists studying a language like English (and native speakers speaking it) can predict other speakers' intuitions about adjective order with high accuracy. This makes adjective order an ideal case study for my purpose, because it gives us a rare approximate 'ground truth' for a latent subjective property (individual adjective order intuitions) against which I can compare data elicited by both a game and a standard data elicitation task. Mirroring the two linguistic paradigms mentioned above, I will do so using alternatively an assumed generativist universal grammar (Study 1) and a person's separately elicited individual grammaticality judgments (Study 2).

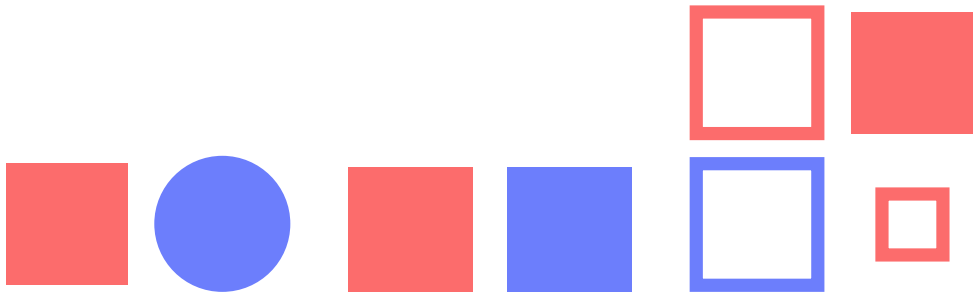


Figure 5.1: Simplified examples to illustrate contrasts in a picture description task

5.1.2 Picture Description Tasks

One common linguistic method for getting participants to produce language data revealing adjective order is the picture description task (Ambridge & Rowland, 2013; Eisenbeiss, 2010). Here a picture (or video, object, etc.) is shown to participants and they are asked

to describe it. This can happen in various degrees of prestructuring. If we present a shape and just ask "What is this?", we have a relatively unstructured elicited production task. By adding structure, we can make it more likely that certain forms of language (like sequences of multiple adjectives) are used. One effective way to do this is to use contrasts (Eisenbeiss, 2009). To identify one shape in contrast to another, e.g. Figure 5.1b, a participant might simply say, for example, "square". However, to distinguish between two squares that differ only in colour, as in Figure 5.1a, the participant might say "red square", or "blue square". In this way, a context can be constructed such that a phrase of arbitrary length might be elicited, such as Figure 5.1c, which might elicit "big empty red square". Such a string of adjectives followed by a noun is called a modified noun phrase. In my case study, I will use this language elicitation paradigm as my control condition, since it is an easily implemented and replicated paradigm that is well-established and thus ecologically valid in linguistics.

5.2 Materials

5.2.1 The Data Collection Game

Having set out my case study – a linguistic picture description task to elicit adjective order – and design approach – Intrinsic Elicitation –, I will now describe how I designed my data collection game and experimental control.

To maximise participant reach, I built *Adjective Game* as an HTML5/JavaScript game, playable in-browser both on desktop PCs and smartphones. Following Intrinsic Elicitation, I started designing the game by identifying a core mechanic that would necessarily generate modified noun phrases while staying as close as possible to the picture description task. This resulted in a casual puzzle game similar to tile-matching games such as *Two Dots* (Playdots Inc., 2014) or *Candy Crush Saga* (King, 2012), albeit with a novel data-eliciting input mechanic (for screenshots, see appendix A.1).

5.2.1.1 Gameplay

The game presents a series of levels of increasing difficulty, each consisting of a grid filled with blocks that have various shapes, colours, sizes, and filling. The goal of each level is to clear all blocks in a given number of moves. To clear blocks, players enter a string of exactly three words that must contain exactly one noun. Players enter strings by tapping/clicking

the labelled buttons at the bottom screen, which show the permitted words for the level – these are nouns describing possible shapes (circle, square, triangle), and adjectives for the possible colours (green, blue, red), sizes (small, large), and filling (empty, filled).

The string entered so far is displayed on the screen. Once a player has constructed a three word string, all blocks currently visible on screen that match this string are cleared simultaneously. It makes not difference in which order words are entered, e.g. players could input "triangle empty small" to clear small empty triangles. Only strings that don't contain a noun or don't identify any blocks visible on the screen are rejected by the game. The player can undo partial inputs. Entering a valid string expends one of a limited number of moves.

When blocks are cleared, the remaining visible blocks fall down to fill spaces in the screen grid, and new blocks fall in from above to refill the grid. This continues until the total number of blocks for a level is hit. Clearing groups of orthogonally adjacent blocks is worth more points than clearing the same number of isolated ones. If a cleared group contains three blocks or more, the player earns a bonus move. The bigger the size of the cleared group, the higher the score bonus and number of bonus moves.

It is possible to perform better or worse at clearing a level (the score achieved upon clearing a level is translated into a three-star rating), and most levels are impossible to clear without earning bonus moves. This invites and requires strategic planning ahead to identify sequences of groups of blocks that can be efficiently created and cleared to maximise score and bonus moves. On each turn, the player thus has to balance simply maximising the number of blocks cleared, manipulating the board to create and clear groups, and tracking how many moves they have left.

5.2.1.2 Realising Intrinsic Elicitation

How did *Adjective Game* realise the three design principles of Intrinsic Elicitation? In terms of *necessity*, the input mechanic involves selecting a sequential order of two adjectives and one noun – the building blocks of a modified noun phrase. Players cannot enter their own words, nor can the mechanic be triggered by any other input. This makes producing a modified noun phrase a necessary part of the game's mechanics.

Inputting three-word adjective and noun strings is also *central* as there are no other mechanics available to the players, making it the one with the highest virtual utility by default. Had players been allowed to clear blocks with just one or two-word strings ("square",

"green circle"), these would on average have cleared far more blocks and thus would have had a higher virtual utility and centrality.

Veracity required the most consideration and iteration by far. To do so, I worked systematically through the different virtual, intrinsic, and external utilities proposed by the Intrinsic Elicitation model (meaning, social norms, opportunity costs, etc.).

First off, I chose not to prescribe, hard-code, display, or reward a particular word order (e.g. that the noun comes last). Had I done so, this would have rendered the mechanic inexpressive of the players' own grammatical intuitions of adjective order. This would be comparable to running a multiple-choice questionnaire about voting preferences and telling participants which party to select.

Second, I quickly realised that I couldn't change the board state or give any dynamic feedback in response to any 'partial' input of just the first or second word, such as highlighting presently selected words. This would have created an informational virtual utility in starting with particular inputs to probe and explore the game state for potential combinations.

Third, while I decided that each valid input must have a noun, I ensured that in every level, no single type of adjective (colour, shape, filling) is always required. This not only adds an interesting asymmetry to the game, but also makes it more likely that in the course of play, players will produce a richer range of different adjective orders, as different orders will be more strategically optimal with changing board states.

Finally, I spent significant time and effort to find ways to control actuation effort as a potential disutility. I particularly tried to avoid any order effects of button arrangements that make certain word sequences easier or harder to input, as eye, mouse, and/or finger have to travel different distances from one word to the next. I thus chose to randomise the order in which buttons appeared in every level. This left some differences in actuation effort, but ones that would be randomised out in the final data. (In Study 1, the order of buttons is entirely random. In Study 2, buttons are grouped into columns by type (noun, colour, size, etc.), but the order of columns and the order within columns is random. I also considered whether and how to allow players to reset or undo partial or incorrect inputs. Allowing partial clearing of an input string back to front, as with a 'backspace' key, might have biased players to retain words entered first. The game therefore only allows you to clear the entire input at once.

5.2.2 The Experimental Control Task

I designed the experimental control task to (a) be as close and ecologically valid to the standard linguistic contrastive picture description task (Eisenbeiss, 2009) while also (b) minimally deviating from the data collection game in all non-game aspects of the interface and interaction.

To this end, I reused the HTML5/JavaScript framework of the game, retaining styling, layout, and core interaction, while removing all immediate 'gamy' interface features (e.g. juicy feedback like exploding stars), core structural game features (goals, levels, progress feedback), and emergent game dynamics and aesthetics (core challenge, increasing difficulty, uncertainty). I also reworded all text speaking of "game" and "play" with equivalent phrases like "experiment" or "interact". Differences between game and control are summarised in Table 5.1.

The control task presents screens that each require a single input. Each screen shows four blocks, one of which is indicated by a double-lined red box (always the top left). The three remaining blocks all differ from the first in only one aspect (colour, shape, etc.) to ensure a three-word description is appropriate and needed to contrastingly describe the target. Users still input the three-word string by clicking/tapping on word buttons, as in the game. Only an input selecting the target block is accepted; incorrect inputs trigger a prompt to the user to select the highlighted block. When the target is selected, the next task starts automatically, without the blocks moving or being cleared.

The control task thus matches a standard contrastive picture description task for eliciting adjective orders. It does so with the same core interaction (and actuation method) as the game. However, as the player does not need to choose the shape, and each trial is separate from the others, all strategic considerations are removed. Similarly, uncertainty is reduced as no unseen new shapes fall down to replace those removed. While new shapes appear each trial, the arrangement is always similar and does not express any properties of strategic interest. Challenge is much reduced as the visual search of shapes is much reduced, as is the strategic challenge of selecting the optimal description to use.

The control condition has a single tutorial screen which shows the participant a single shape (a filled red circle) and presents three words in a random order: 'filled', 'red', and 'circle'. It gives brief instructions to the participant what to do. After selecting all three words in any order, the tutorial is complete.

Feature	**Game Condition**	**Control Condition**
Tutorial	3-level tutorial introducing 1) single word input, 2) block falling mechanic 3) strategic choice in description with three-word inputs	1 level tutorial introducing three-word input for a single shape
Tutorial Instructions	"Matching blocks disappear / Select a word below", "You've got to clear each level in a limited number of moves. / Clear 3 adjacent blocks for a bonus move", "Only blocks that match every word are cleared"	(ex. 1) "Choose from the words below to describe the shape / Think carefully and use the order that feels most grammatically correct to you" (ex. 2) "Choose from the words below to describe the shape"
Instructions Per Trial	(ex. 1) None (ex. 2) Help button ("Choose a 3 word phrase to clear blocks", "Only blocks matching every word are cleared", "Clear groups of 3 to get an extra move (group of 4 = 2 extra moves, etc.)")	(ex. 1) "Describe the highlighted shape in the order that feels most correct to you" (ex. 2): "Choose from the words below to describe only the highlighted shape."
Gameplay		
Mechanic	Enter 3 words to identify one or more blocks	Enter 3 words to identify a single block
Game-specific Goal	Clear screen of blocks	None
Failure Condition	Run out of moves	None
Strategy	Identifying larger groups of blocks. Clearing blocks to create groups of blocks next turn. Gaining bonus moves.	None
Scoring	Current score displayed. Score increases for clearing blocks. Larger groups cleared increases the score more	None
Challenge	Visual search of blocks and buttons. Planning/predicting multiple turns	Blocks always in same arrangement. Visual search of buttons.
Uncertainty	Initial arrangement of level. New blocks from above per move	Arrangement of next trial per move
Interface		
Input	(ex. 1) Randomised grid of buttons (ex. 2) Randomised columns of buttons	As game
Graphics	Simple coloured shapes.	As game
Input feedback	Move successful. Illegal move.	As game
Game-board Feedback	Particle effects on blocks cleared. Level progress indicator.	None
Interstitials	Level start and end dialogues with numerical and three-star scores	None

Table 5.1: Comparison between game and control conditions

5.2.3 Ecological Validity of Adjective Game

I have described *Adjective Game* as an elicitation game. As such it must satisfy the three requirements that I set out in chapter 1: it must be 1) designed for enjoyment, 2) designed to collect meaningful data about players as individuals, and 3) be able to record this data and transmit it to researchers. The first and third of these are straightforward: it *was* designed for enjoyment, and it *did* transmit its data to me (via an online database). There are two steps to justifying the second requirement, that *Adjective Game* is designed to collect meaningful data about its players as individuals. First, I will argue that the data is meaningful. Second, I will argue that that the data is about the players as individuals.

First, adjective order is meaningful. It describes how an individual chooses to (spontaneously) order adjectives, which is largely determined by language, and provides a lens into the syntactic structure of (their) language. Adjective order has been well researched in linguistics. Admittedly, this means it is largely understood for many languages and thus novel data in this area is likely to be of less value to researchers. This is acceptable as the intention here is the metamethodological study of applied games, not contributing to linguistic research.

Second, the adjective orders collected are data about the players as individuals. While the data of an individual speaker can be used to inform the study of a wider language (e.g. data from an English speaker can tell us something about the English language), we are not measuring the language directly, but the individual's behaviour. The desired data for such studies is the individuals' natural and spontaneous inclinations to order adjectives in particular ways. This allows us to in principle test individual and contextual differences in adjective order that differ from the linguistic consensus. One could imagine such a game being used to, for example, assess an individual's learning of a second language (though its appropriateness for this has of course not been established).

Satisfying these requirements, we can reasonably describe *Adjective Game* as an elicitation game. However, to confuse things slightly, the studies reported here adopt a secondary, metamethodological use of this game – not to elicit data about language *per se*, but about the effectiveness of the game itself. The reader may recall that in chapter 1, I argued that templates that use intersubjective consensus or compare participants input to a ground truth are not suitable for use in elicitation games. Yet, in the study designs below, I compare participant data against an idealised ground truth model of English grammar. It is important to note however that this kind of analysis is *in principle* not applicable to

the substantive data collection goal of the game: it could not be used if one's goal was actually to discover an individual's grammatical inclinations. Similarly, the comparison of a participant's data against a ground truth cannot be used to drive any feedback mechanisms within the game (without sabotaging its ecological validity as an elicitation game) because rewarding accurate responding would bias participants to the idealised grammar and threaten the validity of the collected data as a measurement of the individual's own intuitions. We can *only* compare individuals' supplied grammatical intuitions to a ground truth when when we are already consciously assuming what those grammatical inclinations should be for the specific purpose of assessing the accuracy of the game as a measurement instrument.

5.3 Study 1

The first study compared *Adjective Game* and control task in terms of participants' self-reported enjoyment and accuracy (operationalising data validity). Based on prior work and the Rational Game User Model, I hypothesised that:

H1 Players experience more enjoyment in the game condition than the control.

H2 Accuracy is lower in the game condition than the control.

H3 Accuracy in the game condition will be higher than expected by random chance.

A preregistration can be found at https://osf.io/hab82/, along with a repository containing all study materials, code, and data at https://osf.io/u2nze/. The study received ethical approval from the departmental ethics board at the University of York.

5.3.1 Method

5.3.1.1 Materials

I used the *Adjective Game* and Control Task described above as materials.

5.3.1.2 Dependent Variables

In this first study, I operationalised accuracy in adjective order following a generativist linguistic paradigm, according to which there is a single shared true English grammar that native speakers have access to. Thus, I counted participant inputs as accurate when they

conformed with English grammar and inaccurate otherwise, proposing a correct adjective order for the task of *size, filling, colour, noun*. For each participant I calculated accuracy as the number of accurate inputs as a proportion of the number of recorded inputs. It is this proportion that was used in comparing accuracy between conditions. An example set of resulting judgements is provided in (1), with ungrammatical forms prefaced by an asterisk. I welcome readers to compare their own judgements. When doing so, consider each phrase in a neutral context and without special intonation.

(1)
a. big red circle
b. big empty circle
c. big filled circle
d. *red empty circle
e. *red filled circle
f. *red big circle
g. filled red circle
h. *filled big circle
i. empty red circle
j. *empty big circle

I operationalised enjoyment using the Interest/Enjoyment subscale of the Intrinsic Motivation Inventory[2], a well-established 5-point Likert scale that is frequently used in games HCI to assess player enjoyment (Mekler et al., 2014).

5.3.1.3 Sample

Powered to detect a difference in enjoyment with an effect size of $d = .50$, using an alpha of 0.05 and power of 0.8, a power analysis suggested a minimum sample of 50 participants per condition. I recruited 100 adults with the first language of English via Prolific, using the demographic filters provided by that platform. The study offered £1.00 for completing a 10 minute task (£6 per hour) entitled "A study where you describe shapes". After exclusions, 96 participants completed the study, 47 in the game, 49 in the control condition. 2 participants were excluded and their data deleted because they reported an age of under 18 (as their reported ages were -129 and -131, this may have been an input error). 2 participants were excluded because they reported their first language as other than English during the study. Both of these exclusions were in line with the experiment's preregistration. The above sample size excludes the following: data was not correctly recorded for 2 participants, suggesting they did not complete the study. 2 data records were not associated with a Prolific ID so these were deleted in line with my ethics application. Finally, due to a

[2]https://selfdeterminationtheory.org/intrinsic-motivation-inventory/

mistake configuring the study, I over-sampled participants. As, following anonymisation, it would not be possible to identify the 'extra' participants (to support re-analysis in the case that significance was extremely close), data from the over-sampled participants was deleted during anonymisation. Anonymisation was performed immediately upon study completion.

5.3.1.4 Demographics

For reasons of data minimisation for anonymous publication[3], limited demographic information was collected about participants. This included their age, gender, and gaming experience. Age and gender are widely used as summary demographics. While I do not expect there to be significant effect of age or gender on participants' gameplay, a heavy skew to these demographics would invite further consideration of the fairness of the sampling process and the generalisability of results. Similarly, as gaming experience may impact gameplay, this variable was collected to ensure the sample is reasonably familiar with games and thus plausibly representative of the general population. Were the sample highly familiar with games (or not familiar at all) it would limit the generalisability of the results as such players might, for example, adopt different strategies in the game leading to differences in accuracy. However, no particular relationships were hypothesised between demographics and dependant variables.

5.3.1.5 Procedure

Participants completed a short demographics questionnaire which included their age, gender, whether their first language was English, and gaming experience. They were then randomly assigned to either a game or a control condition. In each case, the participants continued until they had supplied 20 complete inputs, excluding inputs made in the tutorial section of each condition. At the end of the experiment (once players had made 20 successful inputs), participants completed the Interest/Enjoyment subscale of the Intrinsic Motivation Inventory.

[3]For a brief discussion of anonymisation see the section 1.8

5.3.1.6 Analysis

While the preregistered analysis used t-tests, following reviewer requests, I report non-parametric equivalents below. This change to a more conservative statistical test fitting the non-parametric data patterns made no difference to the direction of the results; the original t-test results can be found at https://osf.io/u2nze/.

Though most hypotheses are directional, two-tailed tests were preregistered whenever I felt that the alternative directed hypothesis would be a meaningful result worthy of report and further investigation, even if such a result would be unexpected. In general, if results in an unexpected direction would be interesting not just as a refutation of the original directed hypothesis but in their own right, it seems to me to me that they should be capable – or perhaps *permitted* – to achieve significance. While I recognise this is not the norm, I believe that such a principled approach to choosing between one- and two-tailed tests is desirable, where possible, as it reduces the degrees of freedom available to the experimenter. One-tailed tests would be used when a result in an unexpected direction could not be meaningfully interpreted. To take a concrete example from this study, if it were found that the applied game *decreased* enjoyment, the overwhelming interpretation would be that I had simply failed to adequately operationalise my materials (i.e. the elicitation game) and as such would tell us nothing of real interest. The use of two-tailed rather than one tailed tests does not affect the interpretation of positive results as the type 1 error rate does not change.

Analysis was conducted in Python 3.10.2 (Van Rossum & Drake, 2009) using SciPy 1.8 (Virtanen et al., 2020), Pandas 1.4.1 (pandas development team, 2020), Numpy 1.22.2 (Harris et al., 2020) and raincloud plots (Allen et al., 2021). Power analyses were performed in R 4.1.2 (R Core Team, 2021) using the pwr package (Champely, 2020).

5.3.2 Results

5.3.2.1 Demographics

Out of 96 participants, 60 reported their gender as female, 36 as male. The median age was 27, with ages spanning from 18 to 58. One participant who had presumably mistakenly entered their age as 130 was also included in the analysis. Most participants reported playing digital games frequently, with 65 (68%) playing at least several times a week. Only 19 (20%) participants reported playing once a month or less frequently. The demographics

do not suggest any particular threats to the generalisability of the results.

5.3.2.2 Enjoyment

I used a one-tailed Mann-Whitney U test to see if enjoyment is greater in the game than control. Enjoyment was significantly greater in the game ($M = 3.70$, $SD = .83$) than control condition ($M = 3.09$, $SD = .93$), $U = 1590.5$, $p < .001$, $d = .70$, see figure 5.2.

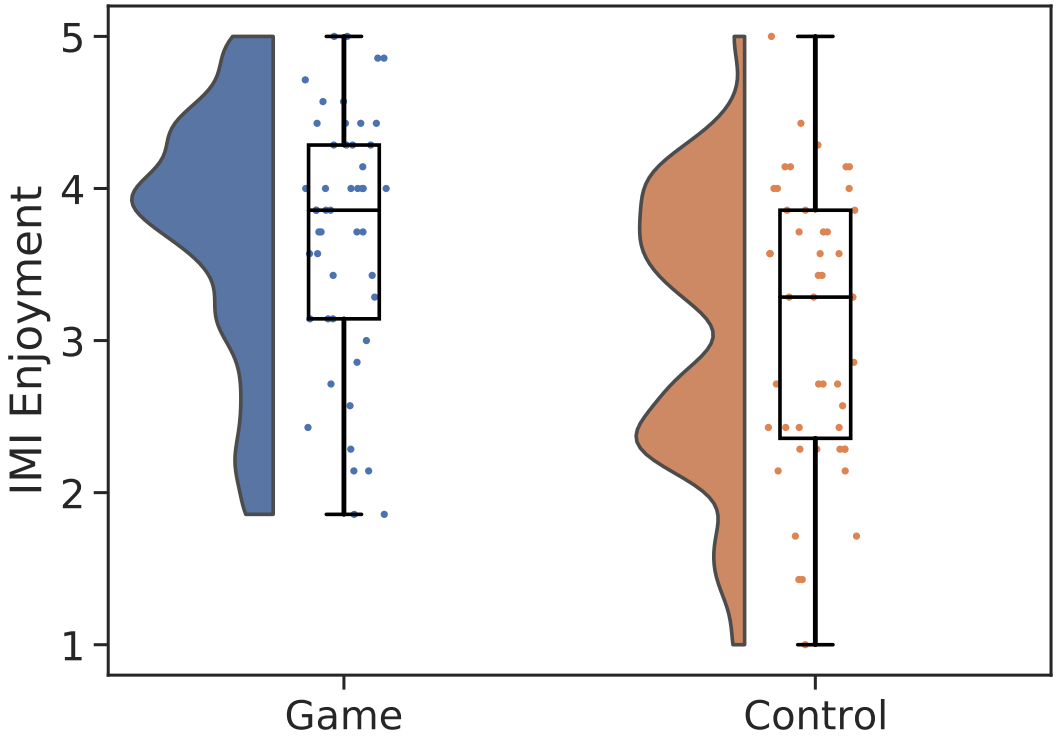


Figure 5.2: Enjoyment as measured by the IMI Interest/Enjoyment subscale is higher for the game than control.

5.3.2.3 Accuracy

A two-tailed Mann-Whitney U test found that the game elicited significantly less accurate inputs ($M = .45$, $SD = .35$) than the control ($M = .67$, $SD = .30$); $U = 734.5$, $p = .002$, $d = -.68$, see figure 5.3.

To determine whether accuracy in the game condition was higher than expected by random chance (H3), I used a two-tailed Wilcoxon test to compare data quality for the game condition against a theoretical random player mean. The proportion of word orderings that would be correct based on completely random answering can be calculated as the proportion of the correct word orders for any given input out of the total possible word

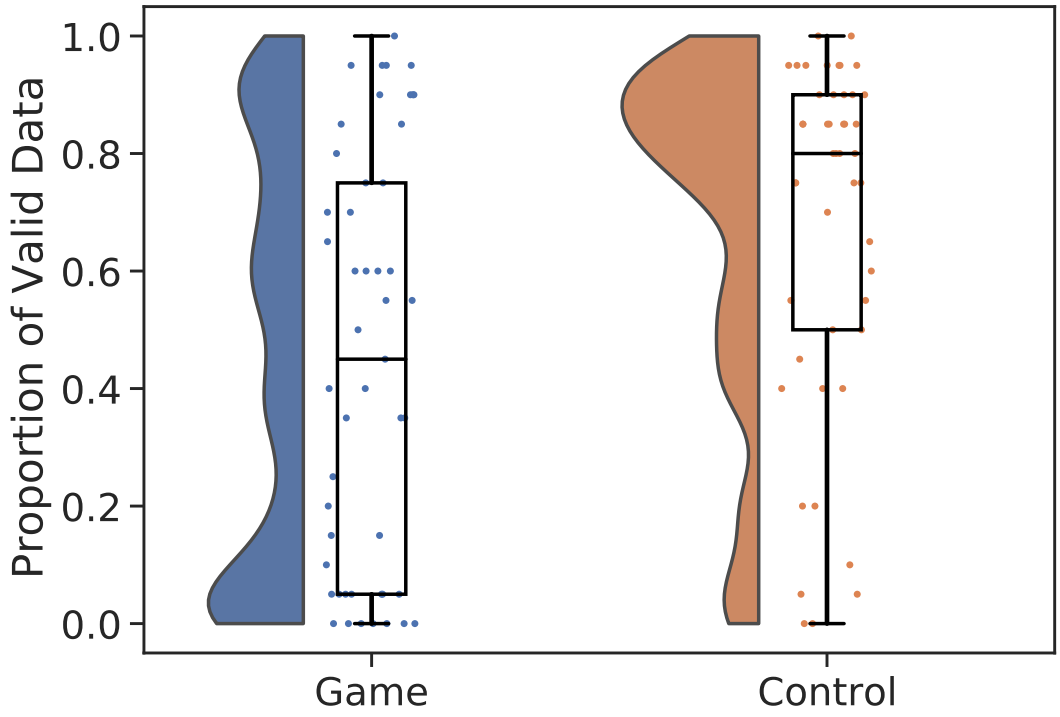


Figure 5.3: Accuracy as share of standard English grammar-conforming word-orders in total word-orders collected is higher in the control than game condition

orders. For this, we are looking at the proportion of grammatical inputs out of "potentially mechanic actuating inputs", a subset of recorded inputs. This is necessary for comparison with our theoretical player. The game can only ever be triggered by inputs of the correct form. Firstly, a hard requirement is that they contain a noun. Secondly, because the adjectives are mutually exclusive, only a single adjective of a given category can be used in a single description. Therefore, our theoretical player who gives functional inputs, but behaves purely randomly where it does not have a mechanical impact would always follow this constraint. Note that this considers inputs possible with the game and not the inputs likely to be necessary in the first 20 moves. This is subtly different from the preregistration which overlooked that adjectives of a category in the game were mutually exclusive. This change makes no difference to the direction of the results. Taking all this into account, there is a single correct word order for an input and six possible permutations of 3 words, making $\frac{1}{6}$ or 16.67%.

There was a significant difference in the scores for the game condition ($M = .48$, $SD = .37$) compared to the theoretical expected value; $w = 154.5$, $p < .001$, $d = .84$. More grammatical inputs were elicited than would be expected from our theoretical random player.

5.3.3 Discussion

In line with my hypotheses, the game showed higher enjoyment (H1), but lower accuracy (H2), though higher accuracy than expected from a random player (H3). The effect size for accuracy differences between game and control was surprisingly large ($d = -.68$). I therefore reviewed the study to find potential confounds and alternative explanations for this large effect size that I could control for in a conceptual replication. First, while native speakers of the same language are widely considered to be consistent in adjective order, this claim may not be convincing to all readers. Some may take a constructivist stance that participants' individual intuitions of adjective order did not align with the ideal grammar used. Hence, I decided to compare adjective order in the second study with a separate elicitation of participants' own grammatical intuitions. Second, I found that time played differed significantly between the game ($M = 348.72$, $SD = 96.72$) and control ($M = 270.17$, $SD = 209.84$) conditions; $U = 1834$, $p < .001$, $d = .48$. Overall, participants in the game condition had spent longer playing. This might have had an effect on reported enjoyment, as engaging longer with the interface might have made the task less novel and more boring. I therefore decided to delimit the second study by time frame, not number of inputs. Third, in the control condition, participants were expressly instructed to provide words in a 'correct' order: "Describe the highlighted shape in the order that feels most correct to you". A similar instruction was missing in the game condition. This may have increased the observed difference in accuracy. Hence I decided to replace this instruction with a more neutral one in Study 2.

5.4 Study 2

Study 2 was a conceptual replication of Study 1 with several changes to test the robustness of my results: using a constructivist operationalisation of accuracy, holding usage time constant, and removing a prompt that could have induced demand effects. A preregistration can be found at https://osf.io/sg3uk/, along with a repository containing all materials, code, and data at https://osf.io/4g9fh/. As in my first study, I hypothesized that:

H1 Players experience more enjoyment in the game condition than the control.

H2 Accuracy is lower in the game condition than the control.

Additionally, given my observation that participants in the game condition took longer

to play, I hypothesized that:

H3 Participants in the game condition will take more time per game input than in the control.

5.4.1 Method

5.4.1.1 Materials

I used *Adjective Game* and the control task described above, with the following updates: in the control condition, to reduce possible demand characteristics, I replaced the instruction "Describe the highlighted shape in the order that feels most correct to you" with "Choose from the words below to describe the shape." In the game condition, I fixed various bugs. Importantly, I removed a menu button at the top-right of the screen that was accidentally included in the first study, which opened a sliding menu that gave the option to return to the main menu, repeat the tutorial, restart the level, and showed a line of debug information revealing the condition, labelled either 'Game' or 'Tool'. It is unlikely that this would have affected participants' responses, as it did not reveal the nature of the other condition to the participant. I also altered the levels and order of levels to improve the game's learning curve and game balance, improved particle effects, and added a help button that opened a modal dialogue with brief instructions about how to play. Finally, I changed how the order of word buttons was randomized: instead of randomly positioning buttons on a grid, buttons were positioned in columns by type (colour, size, shape, etc.), and the order of columns and of buttons in each column was randomized per level. This solution was slightly more usable for players while still retaining randomisation.

5.4.1.2 Sample

Power analyses was performed for each of my hypotheses based on the effect sizes observed in Study 1[4]. The largest of these suggested a sample size of 140 was needed to detect an effect of $d = .48$ (for H3) with a statistical power of 0.8 with an alpha of 0.05. I recruited 185 adults with the first language of English via Prolific. Sampling stopped with 67 in the game condition rather than 68 as was preregistered as the study ran out of money to continue to the unexpectedly high number of exclusions. The study offered £1.20 for completing a 12 minute task (£6 per hour) entitled "A study where you describe shapes".

[4]An exploratory test corresponding to H3 was reported in the discussion of Study 1

After excluding 4 incorrectly submitted records, 9 participants who reported their first language as other than English, and 1 participant who withdrew their submission, 171 participants were included in the published data set. Of these, a further 32 were excluded from statistical tests: 7 were excluded because they had submitted fewer than 16 moves, and 25 were excluded because they reported a bug that, in their judgement, may have influenced how they played the game. Their records are still included in the participant demographics. All exclusions followed the process specified in the preregistration. Thus, a total of 139 participants were included in the statistical analysis. Of these, 66 were in the game condition, and 73 in the control condition.

For each participant, their final 16 valid inputs were used to determine accuracy. This value was selected to ensure as broad a range of players as possible was included in the analysis. The value 16 corresponds to two standard deviations below the mean for inputs (extrapolated to 8 minutes) per user in Study 1, meaning we would expect approximately 95% of participants to be included.

5.4.1.3 Procedure

The procedure was the same as reported for Study 1 with the following differences. Rather than each participant providing 20 inputs to the game (excluding the tutorial), once the participants had finished the tutorial, the participants played the game or engaged with the control task until 8 minutes had passed, regardless of how many inputs they provided.

5.4.1.4 Dependent Variables

As before, enjoyment was operationalised using the Interest/Enjoyment subscale of the Intrinsic Motivation Inventory.

To operationalise accuracy, instead of comparing participants' word strings to ideal English grammar, I compared each participants' strings to that participant's own grammaticality judgments. To this end, after the completion of the Intrinsic Motivation Inventory at the end of the study, participants were presented with a list of modified noun phrases and asked to judge each as either grammatical or ungrammatical. The phrases corresponded to the different ways the types of adjectives (colour, size, etc.) could be ordered in the game/control task. Only noun-final phrases were included, as English has a strong requirement for noun-finality in these contexts. I elicited a total of six judgements from each participant on the phrases given in list (2).

(2) a. red big square

b. big red square

c. big filled square

d. filled red square

e. red filled square

f. filled big square

Accuracy was determined for each participant as the proportion of their recorded game/control inputs for which they had a positive corresponding grammaticality judgement. An input and a judgement correspond if both phrases are similarly ordered with regards to the adjectives they contain. For example, if a participant entered 'small blue circle', this would be compared against their grammaticality judgement for (2b), as both phrases are similarly ordered for colour and size adjectives. If (2b) was judged grammatical, this input would be considered accurate, and inaccurate otherwise. Inputs that were not noun-final were judged inaccurate. For each participant, accuracy was calculated as the number of inputs determined to be accurate in this way as a proportion of recorded inputs. It is this proportion that was used in comparing accuracy between conditions.

5.4.1.5 Analysis

As with Study 1, where there were parametric tests in the preregistration I have reported non-parametric equivalents. This change makes no difference to the direction of the results. As with Study 1, two-tailed tests are selected for directed hypotheses whenever a result in the unexpected direction would be worthy of further investigation. Analysis software was the same as Study 1.

5.4.2 Results

5.4.2.1 Demographics

Out of the 171 participants included in the initial data set, 106 reported their gender as female, 62 as male, and 1 as other. They ranged in age from 18 to 70. The median age was 31. Of the 139 participants included in the statistical tests, 57% reported playing at least several times a week or more, where 29% played once a month or less. The demographics did not indicate any threat to the generalisability of the results.

5.4.2.2 Enjoyment

A one-tailed Mann-Whitney U test found that enjoyment was significantly higher in the game ($M = 3.77$, $SD = 1.07$) than control ($M = 2.99$, $SD = 1.06$) condition; $U = 3387$, $p < .001$, $d = .73$, see figure 5.4.

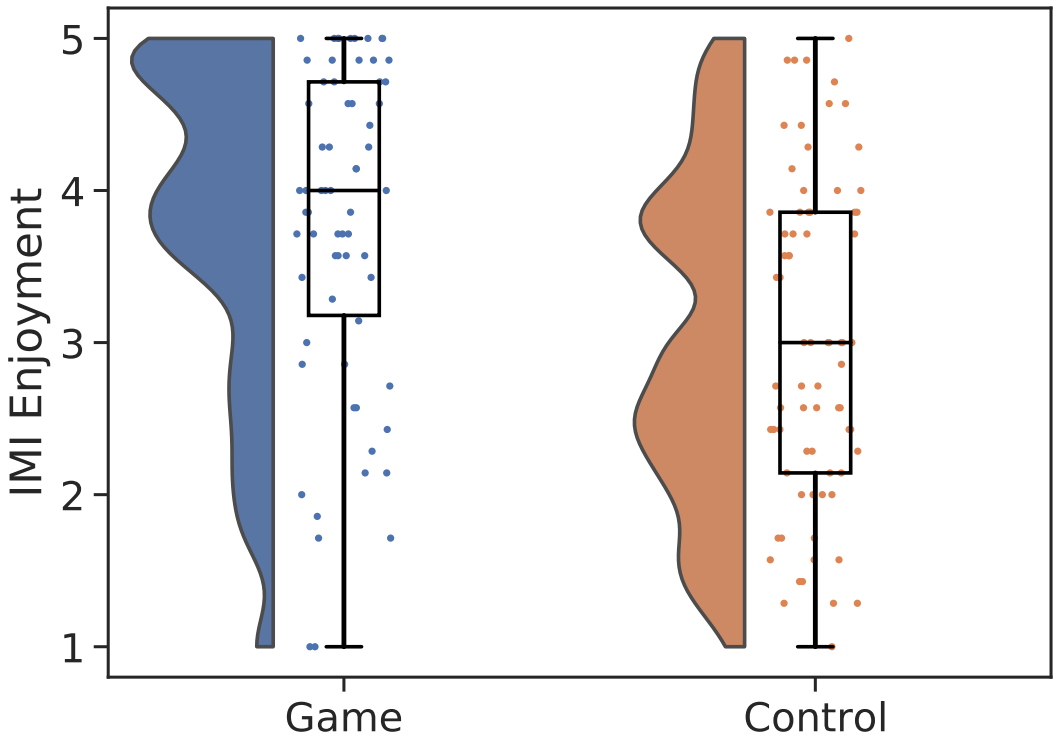


Figure 5.4: Enjoyment as measured by the IMI Interest/Enjoyment subscale is higher in the game than control condition.

5.4.2.3 Accuracy

A two-tailed Mann-Whitney U test was used to compare accuracy in the game and control conditions. Accuracy was calculated as the proportion of the last 16 inputs whose order matched the grammaticality judgement separately elicited for that participant. Accuracy was significantly lower in the game ($M = .32$, $SD = .26$) than control condition ($M = .43$, $SD = .29$); $U = 1872$, $p = .02$, $d = -.40$, see figure 5.5.

5.4.2.4 Time per input

A two-tailed Mann-Whitney U test showed that time per input was significantly higher in the game condition ($M = 15.37$, $SD = 5.07$) compared to the task condition ($M = 9.87$, $SD = 3.30$); $U = 4067$, $p < .001$, $d = 1.30$.

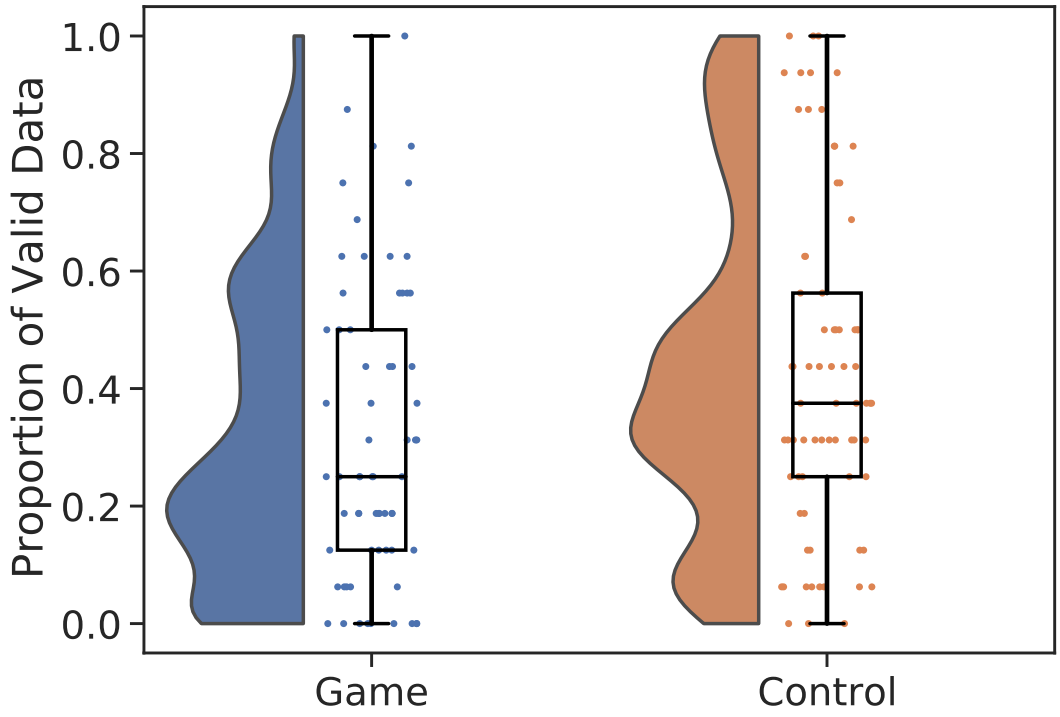


Figure 5.5: Accuracy as proportion of word orders that correspond to the participant's own grammaticality judgments is higher for the control than game condition

5.4.3 Discussion

I again found that enjoyment was higher and accuracy was lower in the game condition than the control, in line with my hypotheses (H1, H2). Also in line with H3, time per input was greater in the game condition compared to the control. This makes the game less efficient as a data collection method than the control in inputs-per-minute. As the interfaces for inputting data were largely identical, this extra time cannot come from inputting the data itself. The difference probably lies in increased cognitive effort from increased visual search and planning involved in strategising during gameplay. However, while the effect size might appear very large ($d = 1.30$), this only amounted to about 6 seconds per input in real terms. For the amount of data required for the experiments reported here, this would work out at a difference of around two minutes between conditions. Such numbers suggest time efficiency is a lesser concern.

5.5 General Discussion

At the start of this chapter I highlighted that data collection games *as games* may present new systemic response biases that threaten validity and data quality – namely that partici-

pants choose responses that are more in-game strategically optimal (virtual utility) or more fun (intrinsic utility) than a more careful and honest alternative response. My results give the worry reason: across two studies, the game condition proved more enjoyable, but also produced less accurate data. Data collection games, at least for the kind of human-subject data that were elicited in this case study, present a trade-off between enjoyment and data quality. That said, since the test game was carefully designed to minimise game-germane response biases, my results also suggest other, presently unaccounted factors at work not accounted for in the Intrinsic Elicitation approach. I will work through the ramifications of these findings regarding accuracy and enjoyment. Evaluation of the Intrinsic Elicitation approach will be left until the discussion in chapter 7.

5.5.1 Accuracy

In two studies, accuracy as a form of data validity was significantly lower in the data collection game than the comparable experiment. This was the case no matter whether accuracy was measured relative to English grammar or to the participants' own judgments. This stands in contrast to the majority of current research on *gamified* surveys and experiments, which finds that gamification at least does not negatively impact data quality (Keusch & Zhang, 2017; Lumsden et al., 2016). I cannot say whether whether this reduced accuracy holds across all kinds of (latent) human subject properties. For instance, eliciting competencies may be less prone to inaccurate responding, where the game can simply encourage and reward people to do their possible best.

Thus, I encourage future research to try to replicate my findings for different elicited properties to establish their generalisability. Methodologically, I urge that such work use careful practice-as-usual controls, akin to gold-standard randomised controlled trials in medical research. Good controls have been largely amiss in past research. While studies without controls (such as Cooper, Khatib, et al., 2010; Crowston and Prestopnik, 2013; Iacovides et al., 2013; von Ahn and Dabbish, 2008) are necessary first steps, they can also easily provide false comfort that applied games are 'quite' enjoyable and produce 'lots' of data with 'above-random' quality. But this doesn't answer the hard, practically important question whether the extra work of turning a survey or experiment into a game pays off with better enjoyment, engagement, and better-or-equal data quality.

While I expected *some* differences in accuracy, following prior evidence that data collection games are less accurate than gamified equivalents (N. Prestopnik et al., 2017) and

the suggestion that games as complex stimuli will necessarily increase variance (chapter 2), the substantial effect sizes surprised me, not the least since I followed the Intrinsic Elicitation approach to give each possible input the same utility as much as possible. This suggests that there were relevant factors affecting player input choice outside of this model.

What, then, differed between game and experimental task that may not similarly manifest in gamified surveys? Perhaps whether or not the game is played *as a game*. Framing a task as a game, in contrast to an experiment, may affect participants' inclination towards careful answering. When Orne (1981) reflected on the nature and origin of *demand characteristics* – the social cues a research participant uses to make sense of what kind of situation they find themselves in and what behaviour therefore is expected of them –, he expressly linked this to situational frames as understood in sociology and recent game studies (Deterding, 2013). He suggested that participants would recognise that their role in the frame of an experimental study was to be a good participant. In contrast, in the frame of gaming, participants might take on the role of players and act accordingly, thus disregarding implicit or imputed normative expectations to respond carefully entailed in the good participant role. This would fit findings by Lieberoth (2015) that merely framing an activity as a game changes people's experience and behaviour. While I took take great care to label the overall study and each condition as neutrally as possible and avoid any direct instructions to respond grammatically 'correct' in either condition, especially in Study 2, participants in the control condition might still have imputed from the overall format of the task that they are engaged in a linguistic experiment where 'correct' word order is expected, while participants in the game condition did no such thing. After all, the framing of "game" or "experimental task" was not just afforded by explicit verbal labeling, but by the whole structure and characteristics of the task itself. The game not only presented superficial characteristics of a game (like a gamified survey), but played like a game. I therefore suggest future research looking to directly induce and assess different framings and see whether this affects behaviours like careful responding and with it, data quality.

Another possible explanation for the observed lower accuracy in the game condition is that the particular mental operations involved in assembling puzzle-solving moves in the game differ from those involved in assembling natural language structures. That is, game participants approached the task as a puzzle whereas control condition participants approached it analogously to spoken language production. Relatedly, the cognitive load

involved in strategising optimal moves in the game might have led participants to adopt tactics like 'offloading' a first likely choice as a first input. One way of probing this explanation would be to further separate the puzzle-solving part of the game from data entry, e.g. ask participants to first choose a combination of words and then say them aloud.

Another alternative partial explanation may be that the operationalisation of accuracy used and the direct instruction to provide grammatically correct inputs in the control condition of Study 1 inflated differences. Both may indeed have had an effect – after all, the effect size for accuracy shrank from $d = -.68$ to $d = -.40$ when these two aspects of the study design were changed in Study 2. This invites future research into the effects of directly asking participants to provide accurate data (akin to common survey design strategies asking participants to respond honestly (Vésteinsdóttir et al., 2019)), and future conceptual replications using different operationalisations of accuracy and data validity.

Inspecting the data, I also observed that accuracy seemed to markedly decrease overall between studies, for both conditions. To test this, I looked at the first 20 inputs in both conditions for Study 1 and 2, evaluating accuracy in comparison to the idealised grammar. An exploratory two-tailed Mann-Whitney test shows that accuracy indeed is significantly higher in Study 1 ($M = .58$, $SD = .35$) than Study 2 ($M = .37$, $SD = .27$) value; $U = 8799$, $p < .001$, $d = .70$.

One possible explanation for this overall decrease is simple regression to the mean. Here, direct replications can help. Another possible explanation are history effects: the first study was run before the COVID-19 pandemic, the second study in the middle of it. A recent analysis of Amazon Mechanical Turk studies identified a surge of new participants, an increase in diversity of participants, a reduction in participant reflectiveness, and an increase in failed attention checks during the pandemic (Arechar & Rand, 2021). Similar history effects may also apply between the studies using Prolific reported here. Finally, Study 2 was time-limited rather than input-limited, which may have induced a sense of time pressure in some participants, resulting in less careful responding. Though again, neither time pressure nor the impact of COVID-19 in Study 2 would explain the differences in accuracy observed in Study 1.

5.5.1.1 Operationalising Accuracy

The two studies reported in this chapter operationalised accuracy in two different ways: through comparison to an idealised English grammar in Study 1 and to a participant's

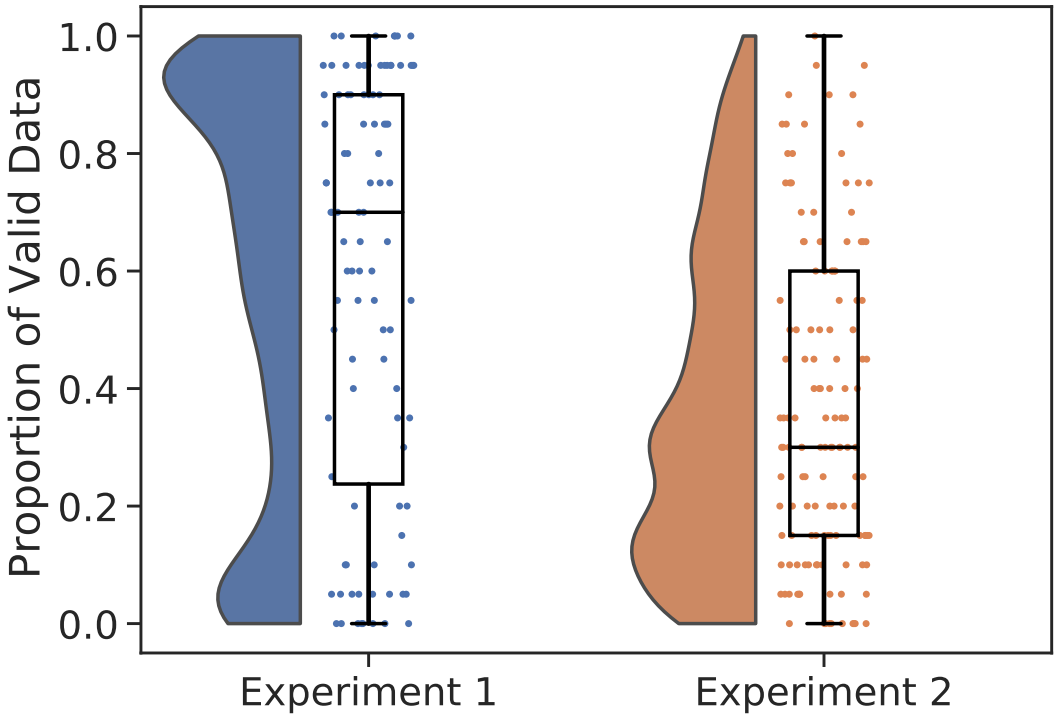


Figure 5.6: Combining both conditions, overall accuracy between studies decrease, measured as agreement with ideal English grammar

separately-elicited grammaticality judgements in Study 2. Each operationalisation has a different claim to validity rooted in different linguistic paradigms: either we assume the correct English grammar is known, or we assume the participants report their grammaticality judgements correctly. The former approach is methodologically simpler and requires less participant time. This raises the question if the choice of operationalisation makes a difference.

First, the two studies report consistent findings consistent results despite operationalisations. An exploratory reanalysis of accuracy (hypothesis 2) in study 2 using the idealised grammar operationalisation reports a slightly stronger effect: accuracy is lower in the game condition; $U = 1715.5$, $p < .01$, $d = -.54$. (The original analysis found an effect size of $d = -.40$)

Second, we can calculate the correlation between the two operationalisations. An exploratory Spearman's rank correlation was computed using the data in Study 2, where each operationalisation of accuracy was calculated over a participants' last 16 inputs. This is visualised in figure 5.7. A strong positive correlation was found between the two measures of accuracy $r_s(137) = .76$, $p < .001$.

The two operationalisations give similar results and are strongly correlated with each

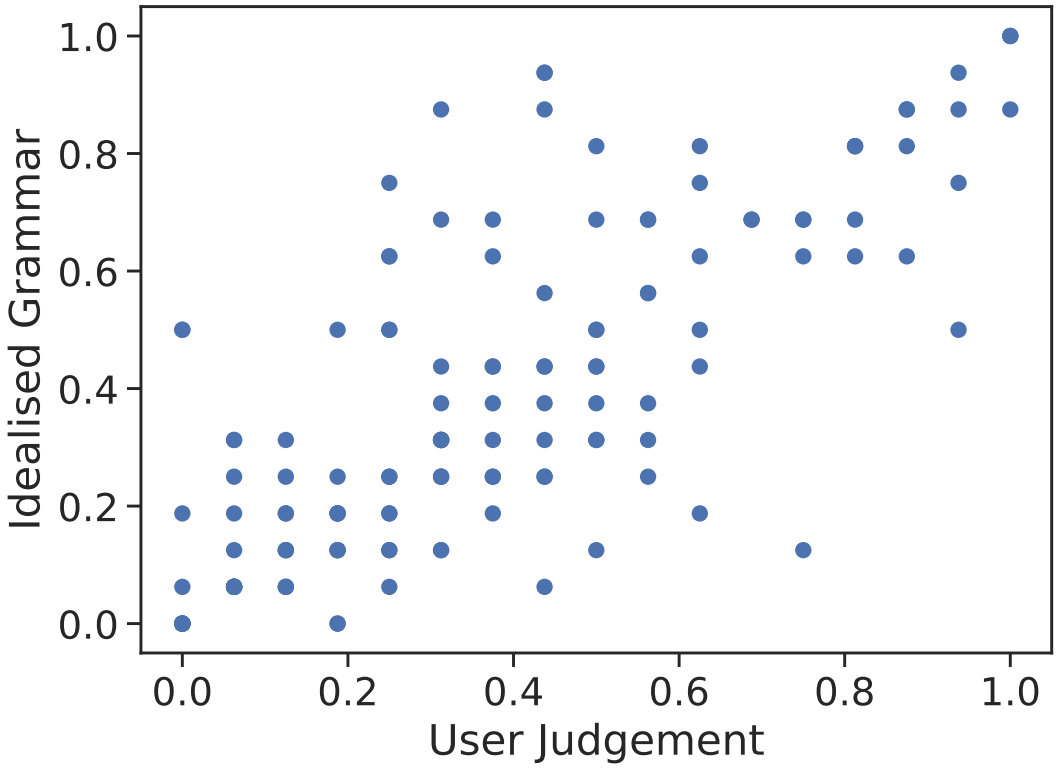


Figure 5.7: The two operationalisations of accuracy show a strong positive correlation. In each, accuracy is calculated as a proportion of the participant's last 16 inputs, leading to the apparent discritisation in this plot.

other. Adjective order was originally selected as a datum in part because it would be possible to operationalise accuracy without needing to separately elicit ground-truth data from participants. As the choice of operationalisation does not seem to make a significant difference, this seems to be justified. The remaining experiments in this thesis will operationalise accuracy using the idealised grammar operationalisation.

5.5.2 Enjoyment

By and large, prior work found that data collection games are enjoyable (N. Prestopnik et al., 2017), and that gamified data collection (i.e. the incorporation of only game design *elements*) features a more positive user experience than comparable controls (Hawkins et al., 2013; Keusch & Zhang, 2017; Lumsden et al., 2016). My findings align with this: players did experience more enjoyment in the game than control, with effect sizes in line with a recent meta-analysis on the engagement/experience effects of gamification in cognitive assessments (Vermeir et al., 2020). This supports prior claims that data collection games can make participation in data provision more enjoyable. I hasten to add that this

does not answer whether such higher enjoyment corresponds or leads to higher behavioural engagement – prior work on gamified data collection suggests it may not (Hawkins et al., 2013; Keusch & Zhang, 2017; Lumsden et al., 2016). Future work is needed that would track correlations between experiential enjoyment and behavioural engagement in data collection *games* (not gamification), especially in non-paid, volunteer contexts.

I suggested above that framing might be responsible for how participants engaged with the game compared to the control. This might have affected enjoyment as well: if people perceived their activity to be voluntarily play rather than paid labour, this in and of itself might have e.g. satisfied people's need for autonomy experiences, generating higher enjoyment (Deterding, 2016a). This opens the broader question whether the same qualities that make data collection games less accurate also make them more enjoyable (e.g. framing), or if the two are separable. The fact that *gamified* data collection seems to improve enjoyment without a loss of data quality suggests the latter. Contrary to my original intuitions, one possible upshot of this is that *gamified* data collection may prove to be a better option for practitioners than data collection *games*. If future research is able to identify and dissociate design aspects that drive one but not the other, this would hold great practical value, as it would help designers avoid the observed trade-off between accuracy and enjoyment.

5.6 Conclusion

Games are not an easy path to engaging participants in providing valid data. They are more challenging and effortful to design than practice-as-usual experiments or surveys. Thus, we need to justify their additional cost with commensurate benefits. The standard rationale is that data collection games are more enjoyable, but provide equal if not better data quality. Yet to my knowledge, this rationale had never been put to a rigorous test, comparing a data collection game to an equivalent control. In two studies using linguistic data elicitation as case material, I found that an elicitation game was indeed substantially more enjoyable than a practice-as-usual study design. Yet I also observed a significant trade-off, in that this game also provided less accurate data. Since I expressly followed the Intrinsic Elicitation design approach dedicated to minimising validity threats and eliminated other likely confounds in a conceptual replication study, there is reason to believe that this trade-off is real. However, I have no ready explanation for its existence, apart from the possibility that framing a task as a game versus as an experiment may induce

different demand characteristics for careful responding.

In this chapter I have tested a game designed following the Intrinsic Elicitation approach and found that, while the game proved to be enjoyable and elicited the desired data from participants, the data was not of the quality of an equivalent experiment. What factors were missing in my use of the Rational Game User Model? In my design I was careful to control virtual utility and actuation effort, however I suggested in the discussion that social norms in the form of demand characteristics might have been responsible for the decrease in accuracy observed. Thus this is a factor that deserves further consideration.

In the next chapter I run two more experimental studies using *Adjective Game* to see if demand characteristics for careful responding exist when using data collection games, and whether their effect is great enough to explain the trade-off observed in the two studies reported here. This will also show if social norms, understood as demand characteristics are a significant factor to consider when using Intrinsic Elicitation and the Rational Game User Model.

Chapter 6

Manufacturing Demand Effects

The experiments in the previous chapter tested an elicitation game for collecting adjective order data – *Adjective Game* – against a standard experimental task. A trade-off was observed – while the game increased self-reported enjoyment, it decreased accuracy of the data elicited. The reason for this trade-off was not immediately obvious. In contrast, related work on *gamified* experimental tasks found that game elements increase (Van Berkel et al., 2017) or at least maintain (Friehs et al., 2020; Hawkins et al., 2013) data quality. This raises the question: what is the cause of the decrease in accuracy observed in the last chapter, and why were the results observed different to those from previous work on gamification?

Interpreting my results, I suggested that *demand effects* might be in part responsible for the decrease in accuracy observed. Demand effects are biases caused by role expectations arising from the social framing of an experimental situation (Orne, 1962). One such role expectation in experiments is the 'good subject', who (consciously or unconsciously) sees their role to confirm the experimenter's hypothesis (Weber & Cook, 1972). It is plausible that such demand effects might affect gameplay that is performed in an experimental situation. Indeed, framing effects were already identified in chapter 2 as among the possible threats to validity characteristic with the use of games in quantitative data collection. Further, in chapter 4 social norms were discussed as a potential source of extrinsic (dis)utility within the Rational Game User Model, biasing the data elicited by a game. I also discussed that social norms demanding attention for game-external considerations in data collection games may diminish enjoyment, as they don't allow players to unselfconsciously and fully pursue the intrinsic utility of game enjoyment derived from playing well.

As such it is plausible that framing, through demand effects, could be responsible

for (some of) the difference between the game and experimental control reported in the previous chapter, and between my results and those of gamification studies: Participants of a superficially-gamified survey or experiment may detect the social frame 'experiment' and with it, that their expected social role is to answer diligently and accurately, thus despite the presence of game features, data quality would be maintained. In contrast, when a participant is encouraged to frame the task 'merely' as a game, without obvious experimental analogue (as in the previous chapter), such demand effects might be absent and thus elicited data might be of lower quality. Improvements in data quality, to the extent they have been seen in gamification studies, may be in part conditional on players identifying contextual social cues signaling what kind of data they are expected to provide. Understood within the Rational Game User Model (chapter 3), it may be that social norms (in the form of demand characteristics) is a significant source of extrinsic utility to providing data in a particular way. In other words, the fact that a game is played *as a game* – because of the way it is framed – might harm data quality, but also increase enjoyment.

One way that a system can be framed as either game or non-game is through metacommunicative framing (Deterding, 2013, pp. 187–190). Merely changing the verbal labelling and visual appearance of an activity as a game has been shown to increase self-reported interest and enjoyment (Lieberoth, 2015). If metacommunicative framing can affect enjoyment, it is plausible that it might also affect how a game is played and thus have an effect on data quality. If demand effects can improve data quality, we might wish to manufacture them, though only if the effect on enjoyment did not counteract the benefit of using a game in the first place. A benefit of labelling of the kind performed by Lieberoth (2015) is that it can be performed easily and cheaply. However, this was originally tested by adding a metacommunicative game framing to a non-game task, whereas here I will – somewhat perversely – be adding a metacommunicative experiment framing to a game task.

In summary, in this chapter, I test whether demand effects, understood following Orne (1962) as role expectations of the experimental situation, affect the data quality and enjoyment in an applied game for data collection. I ran two experiments to test the effect of two simple manipulations on enjoyment and data quality in an experiment game. The first, inspired by success in 'framifying' tasks (Lieberoth, 2015) to increase enjoyment, manipulated the metacommunicative frame within which the game was presented, without affecting other aspects of the game such as gameplay or visual appearance. The second experiment was a maximal positive control (Hilgard, 2021) to gauge the greatest possible

effect of demand effects on data quality that might affect studies in the literature short of using conditional financial incentives, by repeating explicit on-screen instructions on every level in the game.

6.1 Background

Framing has already been briefly discussed in terms of its potential impact on threats to validity in section 2.3 of this thesis. There demand effects are mentioned but only in the context of particular threats to validity that might arise with the use of games. Before continuing, I will reintroduce demand effects as arising from role expectations of social frames in more detail.

6.1.1 Frames and Role Expectations

Frames are socially acquired situational norms which structure our behaviour and experience (Deterding, 2013). They apply to contexts of so-called response-present interaction: situations where two (or more) individuals are able to mutually observe and respond to one another (Goffman, 1983) And, as Goffman (1983, p. 2) remarks that "presumably the telephone and the mails provide reduced versions of the primordial real thing", we can reasonably also extend this to online experimental participation. In order to make sense of these situations, people draw on a shared collection of frames, such as "going to the doctor". Within these, we take on normatively understood roles: doctor and patient, lecturer and student, or experimenter and experimental participant. The frame is jointly maintained by the co-attention of participants on those things that belong to the situation. For instance, when playing a game together we might verbally agree that we will play a game (metacommunicatively framing the coming interaction as gameplay which we participate in voluntarily). We might observe the material arrangement of the situation and in doing so pay attention to the state of the game and ignore things 'outside' the game. Finally, we engage in behaviours appropriate to the gaming situation: moving the game pieces in a manner delimited by the rules, not looking at each others' cards, and so on. While doing so, we allow ourselves to become engrossed, playing for its own sake and without considering consequence; after all, this is just part of the role expectation of being a player (Deterding, 2013).

In contrast, the frame of the social science experiment has markedly different situational

norms and roles. Firstly there are roles of experimenter and subject, which structure how individuals interact within the frame. A subject expects to be given instructions by the experimenter that must be obeyed, that boring or repetitive tasks may be performed with high levels of attention, that the ideal outcome of the experiment is the proof of the experimenter's hypothesis, and that the task, however arbitrary it may appear is carefully directed towards some lofty purpose. In short, it is involuntary, externally controlled, and consequential, the opposite of the game frame.

Frames also affect our experience via our expectations. As such, we might expect the framing of an activity as a game to contribute to it being more enjoyable (as observed by Lieberoth, 2015), similar to how player expectations of gameplay have been shown to lead to a 'placebo' effect (Denisova & Cairns, 2015). In contrast, a game that is framed as an experiment might be experienced as less enjoyable than were it framed as a game, for multiple reasons:

1. **Expectation setting**: players expectations might directly lead to lower enjoyment.

2. **Self-fulfilling prophecy**: players expectations might lead them to be less active in involving themselves in the game, leading to lower enjoyment

3. **Dysphoric Tension**: the misalignment between the material arrangement of the game and the normative demands of the experimental situation (it looks like a game, but I'm told it's an experiment) might give rise to a 'dysphoric tension' associated with boredom and disengagement, particularly if I expect my gameplay to be observed (Deterding, 2019). For instance, although I want to become unselfconsiously engrossed in the game, I feel the need to reflexively monitor and control my involvement to remain aware of the ongoing social demands upon me as an experimental participant.

4. **Performance Cost**: when a player's selection of an actuation – understood within the Rational Game User Model (chapter 4) – is influenced by the extrinsic utility of social norms, this actuation may be sub-optimal with regard to virtual utility, thus harming enjoyment.

Thus games, being played as games, might be more enjoyable than games that we are really aware are experiments.

6.1.2 Demand Effects

Informed by this sociological view of frames and framing Orne (1962) suggested that the normative social expectations held by experimental participants may in some cases give rise to so-called demand effects. In particular it was suggested that participants might be led to act as a 'good experimental subject'. Based on a discussion of the assumed purpose of experiments (to prove hypotheses), and the interests of the experimenter (to publish successful research), Orne proposed that participants are likely to consciously or unconsciously identify an experimental hypothesis and act in a way to confirm it. Thus demand effects pose a threat to external validity comparable to the placebo effect in a pharmacological trial (Orne, 1969). Participants identify the experimental hypothesis through the *demand characteristics* of the experimental situation, which Orne (1981, p. 153) defines as "the sum total of cues available to the subject before the experiment, the instructions during the experiment, the covert communications during the experiment, and the nature of the procedure itself that communicate the experimental purposes and the desired behavior." Standard methodological responses to this threat include double-blinding, obfuscating the hypotheses under test, disguising the nature of the experimental manipulation, and employing deception (Rosnow & Rosenthal, 1997).

However, it is far from clear that participants will necessarily be 'good subjects' and thus display the corresponding demand effects. Other subject roles have been proposed. For example, participants may instead be be faithful (scrupulously honest even if they suspect the hypothesis), negativistic (the 'screw you effect'), or apprehensive (Weber & Cook, 1972). Following Orne's (1962) original formulation of demand effects as role expectations all experiments will contain demand effects – whether or not of the 'good subject' variety. The issue is separating these from the variables of interest. It may be that typically demand effects do not significantly confound experiments (De Quidt et al., 2019); further, it is particularly hard to experimentally test and especially falsify the existence of demand effects (Weber & Cook, 1972), meaning that in many cases they might be more of an experimental bugbear than a genuine threat to validity (Berkowitz & Donnerstein, 1982; Berkowitz & Troccoli, 1986).

How significant are demand effects? One online study using economic 'games' bounded the size of demand effects at between 0.25 and 1 standard deviations, depending on the nature of the task (De Quidt et al., 2018). The more explicit the incitation of demand effects, the more effective it was. More attentive participants conform more to strong

demand effects. In contrast, a study of several standard political science survey experiments run online failed to find evidence for demand effects (Mummolo & Peterson, 2019). Even when a financial incentive was given to give responses in line with the experimenter's hypothesis, participants in most experiments did not display demand effects. This suggests that participants were unable to exhibit demand effects, perhaps due to inattention, being unable to remember the instructions, or being unaware of which response was the demanded one. Furthermore, using non-naïve subjects has been shown to reduce effect sizes, suggesting they are not more likely to match the experimental hypothesis (Chandler et al., 2015).

Online experiments, like the ones reported here, might be particularly susceptible to demand effects (Berinsky et al., 2012) as participants are likely to be experienced with experiments, and may be more motivated due to the design of the labour market to please the researcher (Goodman et al., 2013). In particular, online participants want to be paid and endorsed via reputation systems that online participant pools often provide (Peer et al., 2017). On the other hand, online experiments minimise direct interactions with the experimenter, which is reported to be among the most likely causes of demand effects (Rosnow & Rosenthal, 1997).

Whether demand effects can be manufactured by an online study is far from clear. If they can be elicited, demand effects have the potential to 'improve' the data that is provided (and thus in most cases threaten validity). In principle, if such subjects can be encouraged to view a particular type of data as most demanded by their situational role, it should increase the instances of that data. However, as previous studies have suggested, such effects are difficult to reliably control (Orne, 1969). Thus we do not know from the social scientific literature whether or not we are likely to see demand effects in applied games.

In the context of my project, we can draw two hypotheses with regards to the effects of manufactured demand effects in applied games. First, accuracy of data will increase when demand characteristics are added to the game in a context where accuracy is socially demanded by the experimenter, due to role expectations invoking the desire to be a 'good subject' and provide helpful data. Second, enjoyment will decrease when demand characteristics are added to the game, due to the experimental framing conflicting with the typical autonomous, inconsequential expectations of gameplay. To explore a possible mechanism, I will also consider the hypothesis that participants perceive an activity's

frame less as play when experimental demand characteristics are added. I will now test these hypotheses in a pair of experiments.

6.2 Study 3

I ran an experiment to investigate the effect of metacommunicative framing on data quality and enjoyment. For this I used *Adjective Game*, introduced in the previous chapter, and attempted to metacommunicately frame this game task an as experiment. This study tested two hypotheses:

H1 Accuracy will be higher when the activity is metacommunicately framed as an experiment than when it is metacommunicately framed as a game

H2 Participants will report lower enjoyment when the activity is metacommunicately framed as an experiment than when it is metacommunicately framed as a game

I will also perform a manipulation check to see whether I have successfully manipulated perceived framing. To do this, we will test whether participants report the task as experiment (rather than a game) more when the task is framed as an experiment.

Participants were divided into two conditions: experiment-framed, where I presented the game in a way to encourage players to view it as an experiment; and play-framed, where I attempted to minimise the experimental framing as much as possible. My manipulation was restricted to how the game was introduced and the wording used to describe the game. Both conditions included all game features and the actual task being performed was identical.

6.2.1 Method

6.2.1.1 Materials

The elicitation game *Adjective Game* was used. The version of the game was similar to that used in the Study 2, with minor improvements and bug-fixes, and changes to effect framing, as described below. The most significant improvement to the game was to improve the quality and reliability of automatic level generation for those participants who reached the end of the handmade levels. The modifications reported here were performed with the intention of metacommunicately framing one condition as a game and the other as an experimental task, while keeping gameplay identical. Furthermore, I did not want it to

appear 'too obvious' (Orne, 1962) and alert the participants to the true (as opposed to the participants' own inferred) hypotheses. Screenshots of the manipulation can be found in appendix A.3.

A modified loading screen was added to the game. On this screen, which followed immediately after the consent form, participants in the experiment-framed condition were introduced to the task and its purpose. This highlighted that it was an experiment and emphasised the contribution of participants. The purpose of the task (to collect word-order preference data) was clearly stated. It stopped short of giving any explicit instructions as to how the participants should respond. It continued to refer to the task using non-game language and maintained a formal, experimental visual style: a white screen with black text. There was a forced wait of 10 seconds on this screen to prevent participants skipping through. In contrast, in the game-framed condition the loading screen was integrated with the visual style of the game. The screen was coloured and animated. The instruction to "play however you like" was presented.

Unlike Studies 1 and 2 reported in the previous chapter, due to changed ethics requirements, a consent form was integrated with the game (in addition to participants providing informed consent to begin the study on Prolific). The visual presentation of this consent form was adapted to each condition while the content remained identical to that presented on Prolific. In the experiment-framed condition it was black serif text on a plain white background. In the game-framed condition it was coloured and decorated and had a gamepad emoji. This information stated that the research involved applied games and specified that the study was to investigate "how the design and presentation of a game influences the way it is played."

Finally, in the experiment framed condition, references to play, level, etc. were changed throughout the rest of the game to neutral terms. However, no game elements (score, three-star ratings, etc.) were removed or modified. Also the visual presentation remained the same in both conditions, including animations and particle effects.

6.2.1.2 Sample

A sample size of 168 was selected to identify an effect of $d = .50$ with a statistical power of 0.8. This estimation was based on a t-test and then adjusted on the worst-case Asymptotic Relative Efficiency between t-test and Wilcoxon test of 0.864 (Riffenburgh, 2011), finding a sample size of 148. This value was increased by 20 to allow an exclusion rate of 12%.

Item	Experiment	Game
Pre-Game		
Consent form style	Black and white	Colourful
Post-consent form interrupt	Not used	"We want you to play a puzzle game. You can play however you want."
'Loading' text	"An Interactive Web App for Data Collection You will now see an experimental task designed to identify your grammatical preferences, i.e. what word orders feel natural to you or that you would tend to use. In this experiment you will interact with this app for 8 minutes while we record the inputs you make. Thank you for participating in this study. Your contribution will help our research to find ways to collect scientific data with interactive apps."	"Play however you like for 8 minutes. We will tell you when to stop. Have fun!"
'Loading' style	No animation, 10 second wait	Animation of gameboard filling up
In-Game		
Tutorial level 1 end	"Good job, you cleared the task"	"Good job, you cleared the level"
Tutorial end	"You've got it. Let's continue"	'You've got it. Let's play!"
Level intro	"Task n"	"Level n"
Level complete	"Task completed"	"Level Cleared!"
Game End		
Interrupt	"Thank you for participating in the experimental task. Please click on 'next' to continue to the questionnaire."	"Thank you for playing. Please click on 'next' to continue to the questionnaire."

Table 6.1: Summary of differences between conditions in Study 3

This rate was somewhat lower than the exclusion rate of Study 3 based on bug reports and insufficient moves, which was 19%. This was motivated by further development on the game with the aim of reducing bugs. In total 168 participants were recruited via Prolific. After exclusions 107 participants were included in hypothesis tests, 57 in the high framing condition and 50 in the low framing condition. 5 participants were excluded because they reported their first language as other than English, 13 for reporting that a bug affected their performance of the task, and 36 because they had not submitted 16 moves. All exclusions were in line with the experiment's preregistration. The study offered £1.00 for completing a 10 minute task entitled "An interactive web app for data collection".

6.2.1.3 Procedure

Participants were recruited from Prolific, where they were shown a description of the study that described the task in neutral terms as "an interactive web app for data collection". Participants were randomly assigned to one of two conditions as soon as they loaded the game. The same participant information was presented again so that consent could be captured within the study.

Participants played the game for 8 minutes, which included the tutorial. The tutorial inputs were excluded from data collection. After this time, the player was interrupted. They were directed to the a series of questionnaires. First, a play framing questionnaire was presented, followed by an enjoyment questionnaire, and then a demographics questionnaire. Finally they were asked if they had experienced any bugs that affected how they played and were thanked for their time.

6.2.1.4 Dependent Variables

Accuracy *Adjective Game* elicits a series of inputs from the player in the form of moves in the game. Each input is formed of three words. Accuracy is operationalised as the proportion of a player's last 16 inputs that correspond to standard English word order. Correspondence to standard English word order is defined as in Study 1 (chapter 5).

Enjoyment Consistent with Studies 1 and 2, enjoyment is measured using the Intrinsic Motivation Inventory Interest/Enjoyment Subscale[1] delivered at the end of the experiment.

[1] https://selfdeterminationtheory.org/intrinsic-motivation-inventory/

Play vs. Experiment Framing As manipulation check for the effects on the framing of the experimental situation, I used a bipolar self-report scale of the degree to which a task is perceived either as play or as an experiment. This is an adapted version of the 'Direct Play' questions of the Play Experience Scale (Pavlas et al., 2012). The original questions contrasted 'play' with 'work'. As this might be perceived as incongruous in an experimental context I replaced references to 'work' or 'working' with 'doing an experiment'. The four questions used are below. As in the Play Experience Scale, each is asked on a Likert scale of 1-6 between 'strongly disagree' to 'strongly agree'. Item 4 is reversed:

1. When I was using the app, it felt like I was playing rather than doing an experiment
2. I would characterize my experience with the app as playing
3. I was playing a game rather than doing an experiment
4. Using the app felt like doing an experiment

6.2.1.5 Demographics

Demographics of age, gender, and gaming experience were collected as in the studies described in chapter 5. A heavy skew to these demographics would invite consideration of the generalisability of the results.

6.2.1.6 Analysis

While non-parametric tests were reported for enjoyment in Studies 1 and 2, Studies 3 and 4 report a parametric test, matching the preregistration. Similarly, play vs. experiment framing is analysed with a parametric test. As no effect was detected in either case, the relatively more powerful, parametric test is reported.

As in the studies reported in chapter 5, two-tailed tests were selected despite there being a directed hypothesis whenever a result in the opposite direction would be a meaningful result worthy of further investigation. This does not increase the type 1 error rate. While it arguably gives a higher-than-necessary type 2 error rate, this would not have changed the interpretation any of the null results reported in this chapter.

Analysis was conducted in Python 3.10.2 (Van Rossum & Drake, 2009) using SciPy 1.8 (Virtanen et al., 2020), Pandas 1.4.1 (pandas development team, 2020), Numpy 1.22.2 (Harris et al., 2020) and raincloud plots (Allen et al., 2021). Power analyses were performed in R 4.1.2 (R Core Team, 2021) using the pwr package (Champely, 2020).

6.2.2 Results

6.2.2.1 Demographics

Out of 107 participants, 96 reported their gender as female, 58 as male, and 2 as other. The average age of the participants was 30. A little over half play digital games frequently, with 58 (54%) playing at least several times a week. About a third of participants (35, 33%) have low frequency of gameplay, playing once a month or less frequently. These numbers are do not suggest any particular threats to the generalisability of the results.

6.2.2.2 Framing

A two-tailed independent-samples t-test was conducted to compare reported perception of play framing in the experiment-framed and play-framed conditions. There was a no significant difference in the scores for experiment-framed ($M = 3.59$, $SD = .59$) and play-framed ($M = 3.71$, $SD = .47$) conditions; $t = -1.17$, $p = .24$, $d = -.23$.

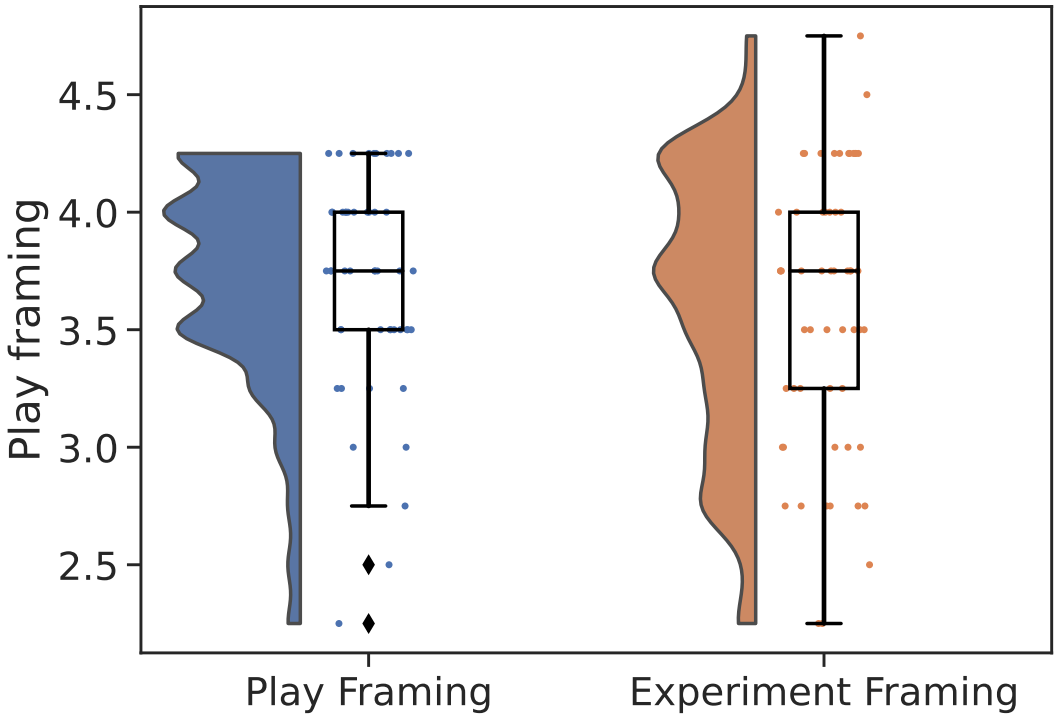


Figure 6.1: No significant difference in play framing between conditions. Play Framing is measured on a 6-point Likert scale.

6.2.2.3 Accuracy

A two-tailed Mann-Whitney U test was conducted to compare Accuracy in the experiment-framed and play-framed conditions. Accuracy was calculated as the proportion of inputs whose order corresponded to standard English word order out of the last 16 inputs. There was no significant difference between the experiment-framed ($M = .46$, $SD = .32$) and the play-framed ($M = .47$, $SD = .29$) conditions; $U = 1386$, $p = .81$, $d = -.03$.

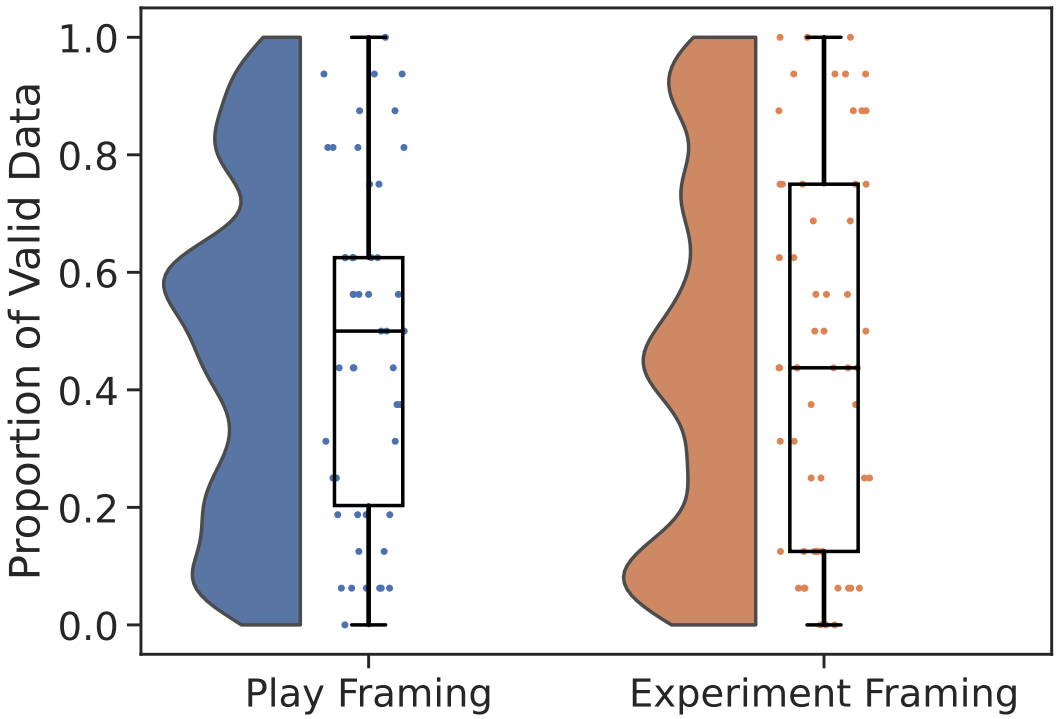


Figure 6.2: Accuracy as the number of standard English word-order phrases collected as a proportion of the last 16 inputs, shows no significant difference between conditions.

6.2.2.4 Enjoyment

A two-tailed independent-samples t-test was conducted to compare enjoyment in the experiment-framed and play-framed conditions. There was a no significant difference in the scores for experiment-framed ($M = 3.33$, $SD = .33$) and play-framed ($M = 3.29$, $SD = 0.28$) conditions; $t = .66$, $p = .51$, $d = .13$.

6.2.3 Discussion

Any effect of metacommunicative framing on accuracy, enjoyment, and play framing was either absent or too small to detect. This is in contrast to the findings of Lieberoth

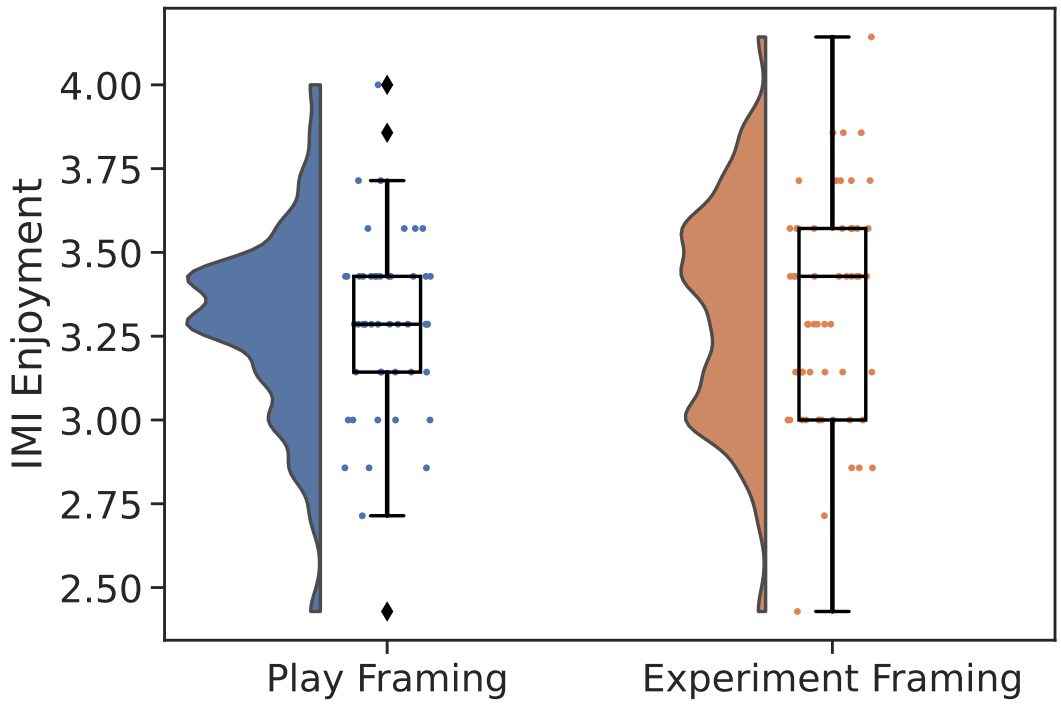


Figure 6.3: No significant difference in enjoyment between conditions. Enjoyment is measured on a 5-point Likert scale.

(2015), as I will discuss further in the general discussion. There are at least three plausible explanations for my results.

First, the metacommunicative framing manipulation may not have been strong enough to induce demand effects. Demand effects are hard to control (Orne, 1969). As the manipulation was largely restricted to how the experiment was originally introduced, it may have simply been overlooked or – over the course of 8 minutes of cognitively involving gameplay – forgotten. If so, it may be that a stronger manipulation, where possible, could be used to manufacture demand effects. In particular, increased prominence of the framing content throughout the game may be more successful. Alternatively, verbal framing may be insufficient, perhaps because it was overpowered by the structure of the task itself – the material arrangement of the situation – which may itself have strongly signalled a gameplay frame. Furthermore, the very fact that participants encountered this game as a study on Prolific may have imparted such a strong framing effect that it overshadowed my manipulation.

Second, participants may not have possessed or adopted the 'good subject' role for this experiment and thus not have been sensitive to the demand characteristics that were manipulated. Participants on the Prolific platform are likely to be experienced at under-

taking experiments, which has been shown on MTurk to decrease effect sizes (Chandler et al., 2015) potentially because they are less sensitive to demand effects. As it was all online, there was no face-to-face contact, believed to more strongly induce demand effects (Rosnow & Rosenthal, 1997). The use of a different participant recruitment method may give different results: applied game studies that rely on voluntary participation, for instance, may be more susceptible.

Finally, demand effects, as understood broadly here as role expectations, may not exist. Previous results showing conformity to the experimenter demand might not be a result of role expectations. This motivates a maximal manipulation to attempt to induce them on both accuracy and enjoyment.

While there was no effect of the experimental manipulation on play framing, there was a large amount of variance in the perception of play framing. This raises two possibilities. First, some participants may have been more affected by the frame than others, with differential effects on accuracy. Second, the measurement of play framing might not have been sensitive to differences in participant's experience of play framing.

First, to explore the relationship between perceived play framing and accuracy, an exploratory Spearman's rank correlation was computed. There was no significant correlation between play framing and accuracy ($r_s(105) = .18$, $p = .06$). While this is approaching significance, the direction of the result: higher play framing associated with greater accuracy, is not straightforward to interpret with respect to my hypotheses, particularly as this is an exploratory test. Furthermore, the causality is uncertain: accurate participants might self-reflect their accuracy as evidence that they were playing rather than performing a task. Alternatively, a third factor, e.g. playfulness, might cause participants both to be (in)accurate and frame their activity more or less as play. As such these results give no information as to whether or not play framing and accuracy are related.

Second, to explore the possibility that the measurement of play framing was noisy and not sensitive to differences in play framing, an exploratory Spearman's rank correlation was computed between play framing and enjoyment. Intuitively we would expect there to be a correlation, as play is usually experienced as enjoyable. A correlation was found ($r_s(105) = .49$, $p < .001$). This makes sense, as we would expect that participants who perceived their activity as play would also enjoy it more. While this was an exploratory test, it adds confidence to our operationalisation of play framing, despite it not observing the hypothesised effect.

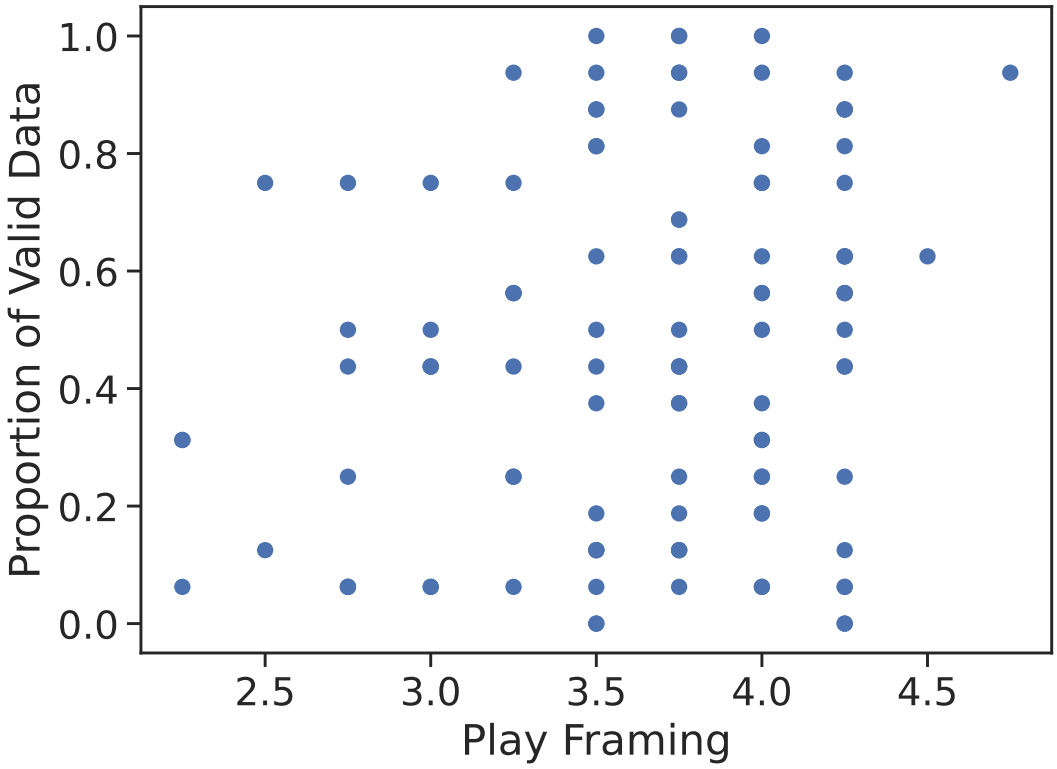


Figure 6.4: No significant correlation between play framing and accuracy.

Overall, these results suggest participants in online game experiments to be resistant the sort of manipulation of framing performed here. Whatever the reason, we can conclude that metacommunicative framing was not responsible for any of the the differences in accuracy observed Studies 1 and 2. Further, similar attempts to improve data quality using demand effects induced by metacommunicative framing are unlikely to be successful. This raises the question of whether even strong manipulations would be able to induce demand effects.

6.3 Study 4

To test whether strong manipulations are able to induce demand effects, I ran an experiment seeking to estimate the maximum effect size a manipulation of demand effects could hope to have in the absence of financially incentivising data quality – a 'maximal positive control' (Hilgard, 2021). This will suggest how significant demand effects might be to data quality and enjoyment outcomes in applied games and thus whether manufacturing demand effects is a worthwhile design strategy.

To do this, I ensured the instruction was explicit. In the previous experiment the manipulation was separate to, and preceded, the task. In contrast, Lieberoth's (2015)

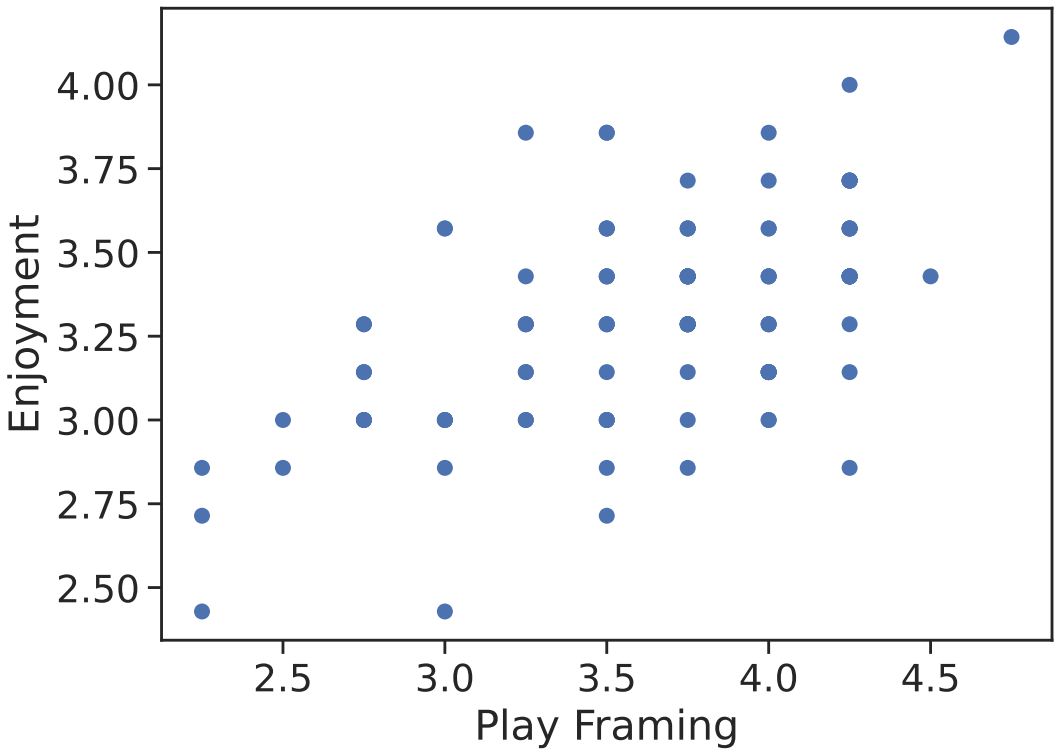


Figure 6.5: A significant positive correlation was found between play framing and enjoyment.

successful manipulation extended throughout the task. Thus I presented the instruction on every level in the game so that it was always visible when the player was making an input to maximise the effect.

If the instruction induces a role expectation ("I am supposed to act in this way because I am being told to by the experiment") we would expect to observe a demand effect on accuracy. Higher accuracy should be observed in the with-instruction condition as following the instruction should lead to more accurate data. Furthermore, if a role expectation is induced, we would expect this to lead to the situation to be perceived through an experimental frame and this will lead to a demand effect on enjoyment. Thus I will test two hypotheses:

H1 A greater proportion of the data will be accurate when an instruction is given to provide accurate data

H2 Participants will enjoy the game less when an instruction is given to provide accurate data

As in the previous study, I will measure perceived play vs. experiment framing as a

manipulation check.

6.3.1 Method

6.3.1.1 Materials

Adjective Game was used again with the integrated consent form as in Study 3. The modified loading screen included for that experiment was removed, and no presentation or labelling differences were made within the game – this was consistent with the play-framed condition of Study 3. Gameplay was identical with the game version in Study 3. The only difference was that an instruction dialogue box was added to the game. The conditions differed only in the wording of this instruction.

The instruction given to participants in the with-instruction condition was: "While you are playing this game, enter words in the order that feels most grammatically correct to you." The instruction in the without-instruction condition was "Play this game however you like.".

The visual presentation was identical between conditions. In each case, there was a slight pause after the instruction had appeared before a confirmatory checkbox labelled "I understand" appeared. The button to begin the experiment only appeared once the checkbox was checked. The game began when this button was clicked.

During the game in the with-instruction condition a written instruction was presented. This appeared on screen at the beginning of every level of the game, but did not require any interaction. The instruction was "Enter words in the order that feels most grammatically correct". To draw attention to it, it faded in at the start of the level. This animation was subtle and in keeping with the visual style of the game.

6.3.1.2 Sample

Following a power analysis, a sample size of 168 was selected to identify an effect size of .5 with a statistical power of 0.8. This estimation was based on a t-test and then adjusted based on the worst-case Asymptotic Relative Efficiency between t-test and Wilcoxon test of 0.864 (Riffenburgh, 2011), finding a sample of 148. This value was increased by 20 to allow an exclusion rate of 12%. While the exclusion rate in the previous experiment was higher than this at 32%, I significantly decreased the required number of moves from 16 to 10 hoping to counteract this. In total, 168 participants were recruited via Prolific following a preregistered sampling protocol. After exclusions 126 participants were included

in hypothesis tests, 67 in the high framing condition and 59 in the low framing condition. 4 participant records were missing in the database, suggesting the participants did not complete the experiment. 12 participants were excluded because they reported their first language as other than English. 21 were excluded due to submitting fewer than 10 moves. 4 were excluded due to reporting bugs. All exclusions were in line with the experiment's preregistration. The study offered £1.00 for completing a 10 minute task entitled "An interactive web app for data collection".

6.3.1.3 Procedure

Participants were recruited with the same study information and consent form as in Study 3. Following the consent form, participants were shown an instruction that differed between conditions with a confirmatory checkbox and button to begin the game. After this, participants played the game for 8 minutes, which included the tutorial. After this participants were asked to complete the play vs. experiment framing scale, followed by the IMI Interest/Enjoyment subscale, and finally the demographics questionnaire. Finally, they were asked to report any bugs they had experienced and thanked for their time.

6.3.1.4 Dependent Variables

Accuracy was operationalised as in Study 3, except the number of inputs considered was decreased to 10 in an attempt to reduce the exclusion rate. Other dependant variables are operationalised as in Study 3.

6.3.1.5 Analysis

As with Study 3, two-tailed tests are selected for directed hypotheses whenever a result in the unexpected direction would be worthy of further investigation. Analysis was performed using the same software as in Study 3.

6.3.2 Results

Out of 164 participants, 86 reported their gender as female, 61 as male, and 4 as other. The average age of the participants was 31 and they varied in age between 18 and 78. A participant who presumably mistakenly entered their age as 130 was also included. Of the 126 participants who, following exclusions, were included in hypothesis tests, a little over

half play digital games frequently, with 71 (56%) playing at least several times a week. About a third of participants 39 (31%) report playing once a month or less frequently.

6.3.2.1 Framing

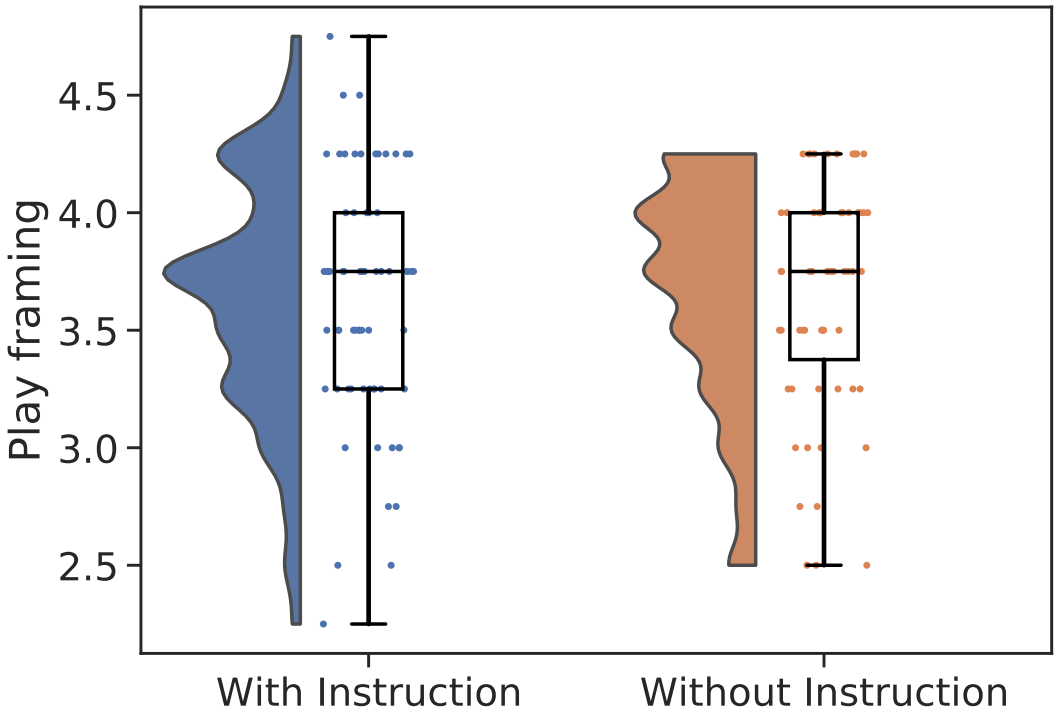


Figure 6.6: No significant difference in play framing per condition. Play framing is measured on a 6-point Likert scale.

A two-tailed independent-samples t-test was conducted to compare judgement of task framing in the with-instruction and without-instruction conditions. There was no significant difference in the distribution of framing judgement between the with-instruction (M=3.6306, SD=.5156) and without-instruction ($M = 3.65$, $SD = .49$) conditions; $t = -.24$, $p = .81$, $d = -.04$.

6.3.2.2 Accuracy

A two-tailed Mann-Whitney U test was conducted to compare Accuracy in the with-instruction and without-instruction conditions. Accuracy was calculated as the proportion of inputs whose order corresponded to standard English word order. The distribution of Accuracy differs significantly between the with-instruction ($M = .55$, $SD = .33$) and without-instruction ($M = .45$, $SD = .31$) conditions; $U = 2396$, $p = .04$, $d = .33$. Accuracy was higher in the with-instruction condition.

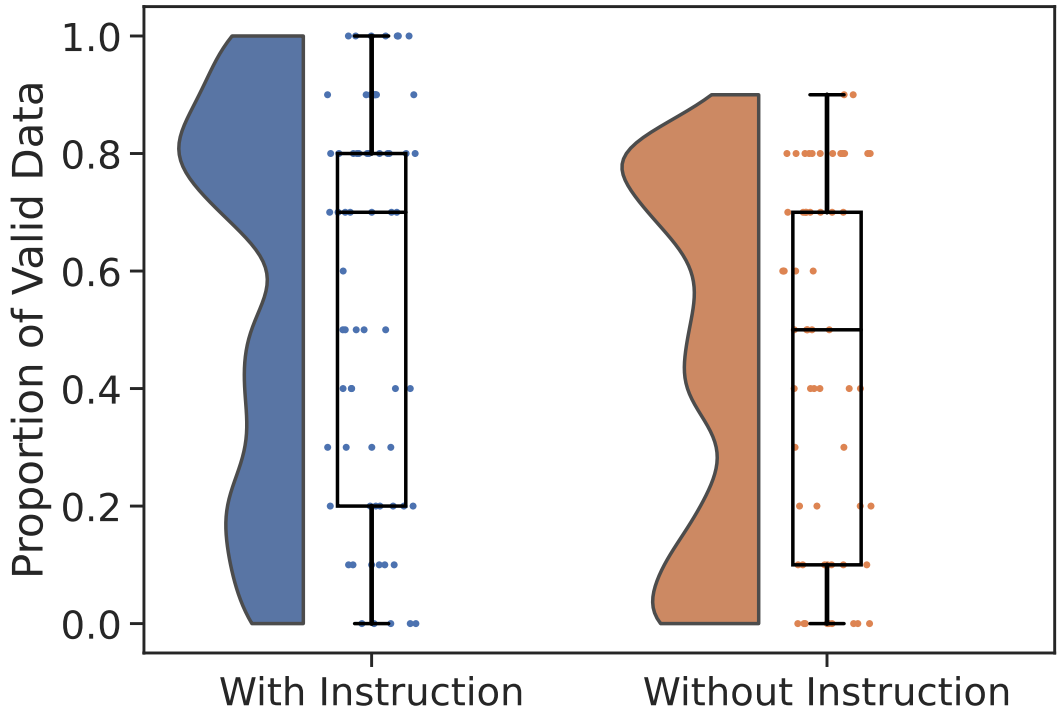


Figure 6.7: Accuracy as the number of standard English word-order phrases collected as a proportion of number of inputs is higher in the with-instruction condition

6.3.2.3 Enjoyment

A two-tailed independent-samples t-test was conducted to compare enjoyment in the with-instruction and without-instruction conditions. No significant difference was found in the scores for with-instruction ($M = 3.31$, $SD = 0.29$) and without-instruction ($M = 3.38$, $SD = .37$) conditions; $t = -1.26$, $p = .21$, $d = -.23$.

6.3.2.4 Exploratory Correlations

For comparison with Study 3, an exploratory Spearman's rank correlations was performed between play framing and accuracy and between play framing and enjoyment. No correlation was found between play framing and accuracy ($r_s = .02$, $p = .79$). A positive correlation was found between play framing and enjoyment ($r_s = .49$, $p < .001$).

6.3.3 Discussion

Giving participants explicit instructions to give a particular desired kind of data improves accuracy ($d = .33$). This is within the (wide) range of effect sizes identified by De Quidt et al. (2018), and without negatively impacting enjoyment.

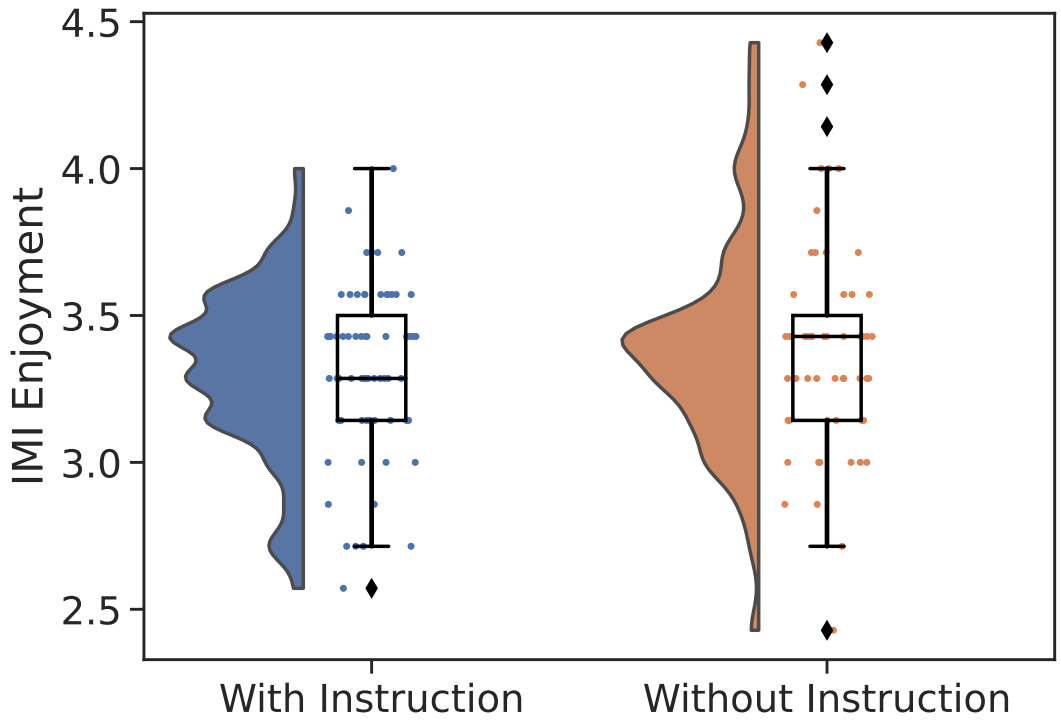


Figure 6.8: No significant difference in enjoyment per condition. Enjoyment is measured on a 5-point Likert scale.

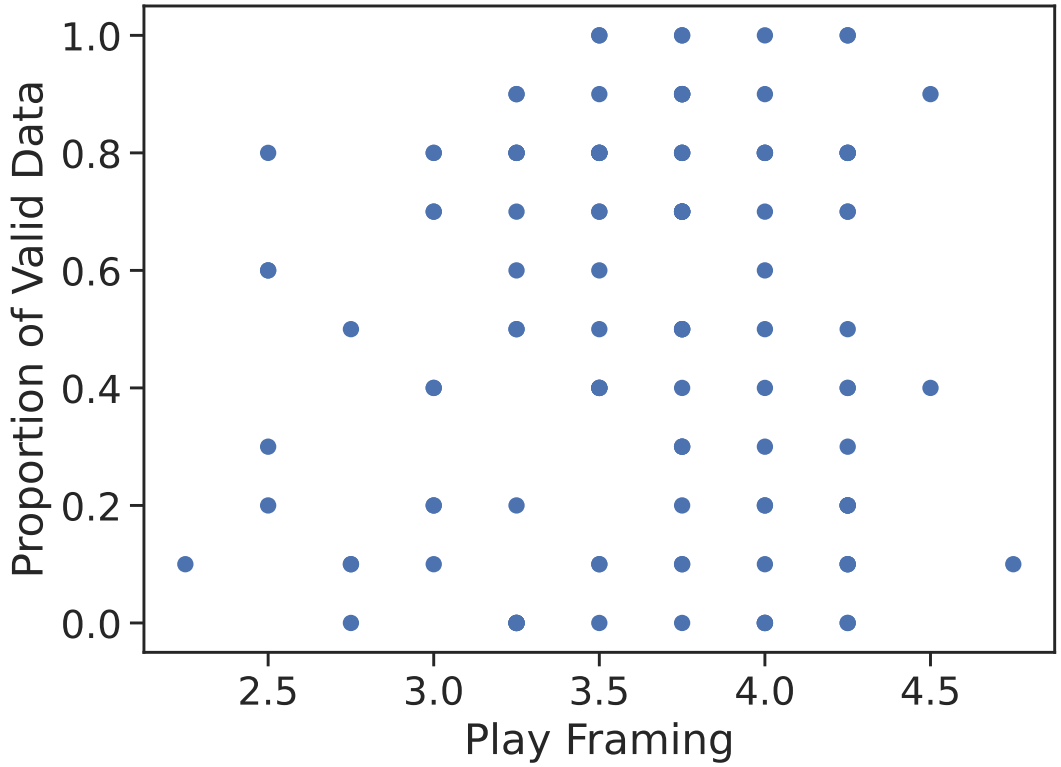


Figure 6.9: No significant correlation between play framing and accuracy.

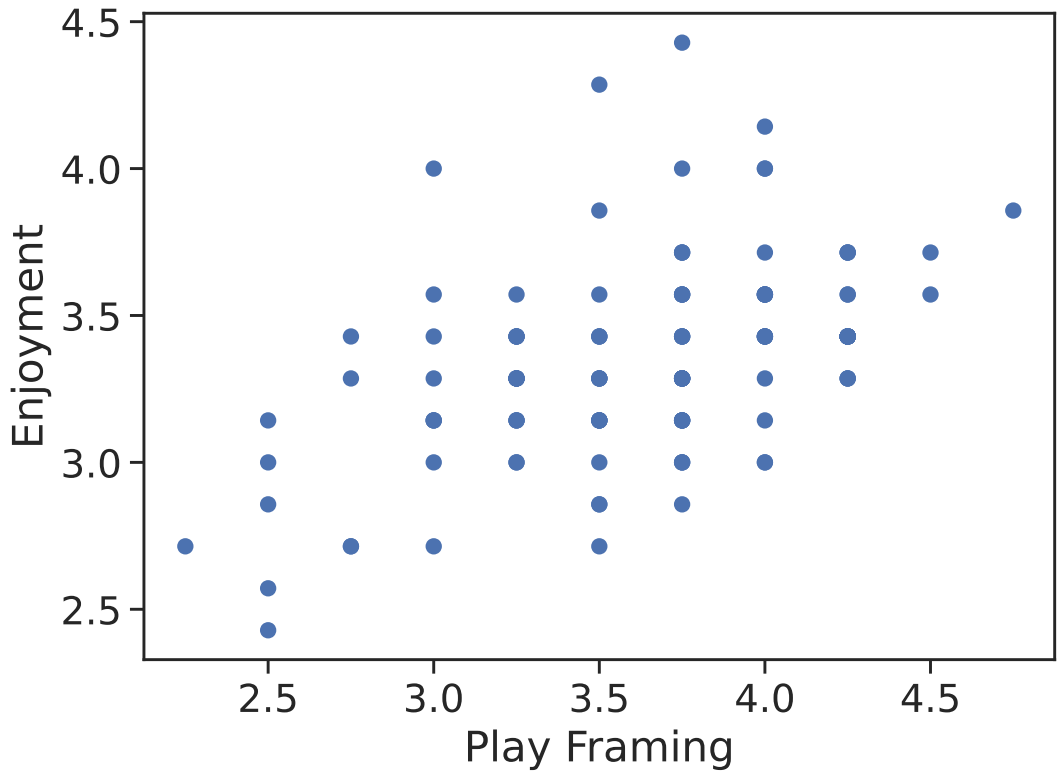


Figure 6.10: A significant positive correlation was found between play framing and enjoyment.

Demand effects are one possible explanation for this data pattern. Under this account, participants, having being given the experimental instruction, identified what was desired of them within the experiment and, satisfying their situational social role, followed that instruction. Instruction following is part of the role expectation of being a participant within an experiment.

These results can also be explained without recourse to demand effects. It is possible that, rather than interpreting the instruction within the experimental frame as a normative demand of the experimental situation, participants interpreted it as game information, which would be congruent with a game frame. It is normal for games to give instructions to help players learn to play. Such instructions are typically guiding players towards a satisfying game experience. Thus, participants might have based their instruction-satisfying behaviour on norms and utility expectations of gameplay, rather than the norms of experiments. That I saw no difference in play framing between the with- and without-instruction conditions may suggest participants did not see the task as an experiment and that this explanation is more likely, but such evidence is weak as neither experiment successfully reported any effect on the play vs. experiment framing measure. Furthermore, the mean

scores measured on this scale (Study 1: 3.59, 3.71; Study 2: 3.63, 3.65) were were in the middle of the 6-point scale (3.5) suggesting in aggregate players did not perceive it more as play than an experiment. On the other hand, the instruction was first presented before gameplay, which would make clear it was not a part of the game.

Finally, I did not detect an effect on enjoyment, though I cannot disconfirm the existence of a small effect due to the statistical power of the experiment. I also did not detect an effect on perceived play framing, but this tells us little. The most likely explanation is that giving an instruction does not affect whether the participant interpreted their behaviour as play – and as I did not intentionally manipulate framing this is unsurprising. I again observed a positive correlation between play framing and enjoyment, which makes strong intuitive sense, suggesting that the play framing scale is measuring something meaningful, although this result was exploratory. To demonstrate that there is no effect of instruction on either enjoyment or play vs. experiment framing would require a more highly powered study than was possible here.

6.4 General Discussion

I found evidence that demand effects – broadly construed as situational role-conforming behaviour – might have an effect on the accuracy of data from an applied game. A relatively maximal manipulation (short of conditional financial rewards) found an effect size of $d = .33$. This contributes to our understanding of data quality in game(ified) experiments and surveys. In particular, we should pay attention to presentation differences in studies to identify possible demand effects, particularly those related to instructions to participants, as these may be partially responsible for effects on data quality. As gamification studies are typically presented as a more or less transparent veneer over an existing task, such as a survey, some degree of desirable demand effect may be common. In contrast, applied games studies typically present a game as a game, and may therefore not be influenced by such demand effects. This is far from a problem with the studies themselves: demand effects do not necessarily threaten validity in these cases. Indeed, one benefit of gamified approaches to data collection may be that they do not necessarily interfere with the underlying social norms of experiment or survey responding that may maintain data quality. However, it makes comparison of different methodologies more difficult, especially if the aim is to understand the role of particular game features.

I was unable to detect any effects of my metacommunicative framing manipulation in

Study 3. One potential reason for this is because of the existing and significant framing effects of the Prolific platform. Having likely participated in many studies before, Prolific participants are trained and socialised as to a set of norms expected of them. They will be aware of Prolific's reputation system, payment system, and the fact that Prolific serves as a platform for running experiments. The norms and roles of participating on Prolific will thus themselves constitute a frame. This wider frame can 'key' – or transform the frame of – an inner activity, such as participating in a study. Following instructions is not only normative within experiments, but it gains a specific meaning in Prolific experiments due to the ability of experimenters to withhold payment. The wider interaction with the Prolific platform which brackets the experiment might pose a very strong frame that is hard to shift within an 'inner' frame.

Another possible reason is because the activity itself absorbs participants in play, overriding any effects of metacommunication. Involvement an activity may make participants forget wider framing for a time as attention is focused on the experience rather than the medium (Rettie, 2004), i.e. the *play* of the game rather than the game as a task presented on ones computer monitor. As such, the metacommunicative framing effects presented in Study 3 might have been forgotten. However, participants could not have been so involved so as to filter out the instruction in Study 4, which was visible on screen during each level and did have an observed effect.

I did not detect an effect of demand effects on enjoyment. This is in contrast to the very large effect ($\eta^2 = 0.197$) observed by Lieberoth (2015) when 'framifying' a non-game task[2]. This experiment was run in person, with framing established partially by a physical game board visible to participants throughout the whole task. The direction may also be relevant: it might be easier to give a non-game task the 'benefit of the doubt' than to ignore the gaminess of a game. However, I wonder if – ironically enough – demand effects might be partially responsible for Lieberoth's findings. We might expect that shallow framing of the kind used would be most easy to see through; if asked to evaluate a superficially gamified task described as a "business consulting tool" (Lieberoth, 2015, p. 5) players might feel like they are expected to enjoy it more. Furthermore, the study design was a group randomised trial (where individuals are assigned to a group and groups are assigned to conditions) (Murray, 1998) yet the analysis performed did not control for the intraclass correlation within groups. This would potentially increase the observed effect size if participants

[2]In the text, Lieberoth (2015) describes this effect size as small to medium

positively affected each others' enjoyment scores. Some effect of my manipulations on enjoyment cannot be ruled out, however. As the effect sizes ultimately observed on demand effects were below what my studies were powered to reliably detect, if similarly sized effects on enjoyment were present I may not have detected them.

A possible alternative explanation is that I manipulated the players' mental model of how to play the game without invoking role expectations of the experimental situation. This is akin to Zizzo's (2010) distinction between 'purely cognitive' demand effects, and 'social' demand effects in economics. My treatment of demand effects, following Orne (1962), as role expectations grounded in the sociology of framing considers all demand effects to involve a social element. In contrast, in Zizzo describes 'purely cognitive' demand effects as those that arise from cognitive aspects of *task construal*: sans social pressure, what decision do I think best optimises my utility from playing the game? Differences in a player's instruction in a game – i.e. what assistance is required to players – may have some effect on data quality and these must be considered when assessing the construct validity of game-elicited data. In other words, one might not just consider the *actual* virtual utility of actions within the Rational Game User Model, but their *expected* utility. In the study presented here, the instruction did not affect the strategic play of the game. However, within the Rational Game User Model we might expect that if an instruction (social norms) went against strategically optimal play (virtual utility), it might be ignored by players.

The exploratory correlations performed in each study found significant correlations between play framing and enjoyment. While this does not demonstrate causality (two equally plausible explanations are that play framing could lead to enjoyment, and that enjoyment could cause participants to perceive the activity as play), it does suggest that the operationalisation of play framing used was measuring something meaningful, and my failure to observe effects on play framing and between play framing and accuracy were not due to my choice of operationalisation. It is also consistent with Lieberoth's (2015) findings: his metacommunicative labelling of a non-game task as a game may have led participants to experience play framing, which led to enjoyment.

Exploratory correlations between play framing and accuracy in both experiments largely contradicted the intuition that differences in perceived play framing might have been associated with differences in accuracy: experiencing the game though a play frame was not associated with decreased accuracy. Indeed, in Study 3, though the results were only ap-

proaching significance, there was a *positive* correlation between play framing and accuracy. This is surprising as the game play frame is commonly associated with reduced concern for game-external consequences (Deterding, 2013). It may be that due to the instrumental nature of the input method, the particular order in which adjectives were supplied was not seen as consequential to participants in either frame. Thus studies whose data collection is a more obvious result of game*play* (rather than game *input*) may see effects where I did not.

6.4.1 Manufacturing Demand Effects

It may be easy to manufacture demand effects to increase data quality using instructions as the manipulation requires little effort. This could only be done if it would not constitute it's own threat to validity, i.e. the demand effects would need to be orthogonal or uncorrelated with the variable under test or true hypothesis (Zizzo, 2010). For instance, a survey could instruct participants to respond honestly – while carefully not implying that any particular answer would be considered by experimenters as the most honest (which in demand effects would be not be orthogonal to, and thus confound, the variable under test). However, Orne (1969, p. 154) cautions that "It can be extremely difficult to predict how, if at all, demand characteristics are altered by instructions". To make demand effects a reliable tool for use they would need to be demystified and preferably reduced to directly manipulable operationalisations.

There may be ways elicitation games could try to increase the effectiveness of demand effects. The manipulation being present and visible throughout the game may help, so that it is harder to overlook. Running in person (Rosnow & Rosenthal, 1997), making data provision (perceived to be) not anonymous (De Quidt et al., 2019), or providing payment (Mummolo & Peterson, 2019) may help, but not necessarily reliably, and such approaches may not often be practical. Naïve participants may exhibit greater demand effects (Chandler et al., 2015), and such participants are perhaps more likely to be recruited via gaming sites rather than participant portals, though this is far from certain.

I did not provide additional motivating context that might explain *why* particular input data was valuable beyond merely contribute to my research. A sense of meaningfulness may be easier for some games to instil than others. However, in the absence of such context, presenting the participants as 'doing a favour' for the researcher may help (De Quidt et al., 2018).

6.4.2 Limitations and Further Work

The sample size in both studies ($n = 107$, $n = 126$) was smaller than the sample sizes originally sought (168, which included 20 expected exclusions) due to unexpectedly high numbers of exclusions. This limits the interpretation of the negative results presented here. Though I did not observe an effect on enjoyment nor an effect of metacommunicative framing on accuracy, I cannot rule these out.

An alternative explanation for my results is that participants followed the instructions without any desire to satisfy a role expectation in the experimental context. In the context of a game, instructions are frequently used to onboard new players who lack experience with a particular game or a type of game. Just as during an experiment we expect to follow the instructions of an experimenter, when encountering a new game, we expect that necessary guidance may be provided in the form of instructions. We are unlikely to doubt its general applicability, even if we later learn more successful strategies through experience. In other words, games do not tend to sabotage or misdirect their players. However, given the rate at which participants didn't answer correctly, one would expect participants would soon identify the instruction was not necessary. Future work could test whether it is possible to manipulate the mental models of players in this way to affect data quality.

While metacommunicative framing before the experiment had little effect, considering the results of Lieberoth (Lieberoth, 2015) and of Study 4, I suggest that an important variable is where the manipulations occur in the game. Context from before or 'outside' of the game may be forgotten due to task demands of the game. Thus manipulations to improve data quality should be presented alongside the main task.

Demand effects depend on participant perceptions and beliefs about what the role expectations upon them are. As I did not observe the expected effect on perception of play framing, I may not have successfully manipulated this. It is also possible that participants saw through my manipulation. If participants are as adept at identifying demand characteristics as some authors propose, they may have recognised that the instructions given to them were themselves a manipulation for a wider hypothesis. This experiment to study demand effects was in some ways unlike a normal experiment and this may have an effect (Weber & Cook, 1972). The instructions may have appeared 'too simple', not sufficiently camouflaged to appear like the 'real' instructions for them to follow (Orne, 1962).

I was not able to detect an effect on enjoyment. It would be relevant to know if there is an effect of demand effects on enjoyment for two reasons. First for high-level

comparison between gamification and applied games methodologies. If gamification tends to evoke role expectations that lead to reduced enjoyment, this would hinder gamification as a motivational tool and suggest that applied games, typically not evoking such role expectations, have the potential to be more motivating. Second, it affects whether it is worthwhile to elicit demand effects as a conscious design choice to improve data quality in a game.

6.5 Conclusion

To understand the main threats to experimental validity from games, we need to isolate the effects of particular game features on data quality operationalised in a way that points to potential implications for bias and validity. In terms of the Rational Game User Model, we need to identify those features that must be included in our model of the player's rational decision making. I have taken the first step towards this goal in considering demand effects as arising from social norms. While not within the game, the role expectations players bring to elicitation games have been shown to affect the data quality they provide.

I found that demand effects may increase data quality from elicitation games. While my research used a 'whole' game, it suggests that demand effects may also play a role in data quality in gamified experiments and surveys. While I was unable to detect any effect of metacommunicative framing on accuracy, a maximal manipulation of demand effects through repeated explicit instructions found a positive effect ($d = .33$) on accuracy. Telling participants to respond in a certain way led them to do so. I detected no significant effects on enjoyment, but I did do not have the statistical power to confidently exclude comparably sized effects.

First, this result aids in interpreting the literature. In particular, while different studies report differing effects on data quality, it is important to consider how demand effects may have contributed to this. Second, it raises the possibility of consciously manufacturing demand effects to improve data quality. A manipulation as simple as giving participants the instruction to answer in a certain way may be enough to improve data quality. According to the Rational Game User Model we can understand this as creating an extrinsic utility for certain inputs via social norms. Further, this need not threaten the validity of the data collected so long as the Veracity principle of Intrinsic Elicitation is satisfied.

Finally, these results also contribute to interpreting the results of Studies 1 and 2. It is unlikely that metacommunicative framing was meaningfully responsible for the effects

on accuracy observed in Studies 1 and 2. The instructions given to participants in Study 1 differed between conditions, with the control task giving a direct instruction for grammatical (accurate) inputs. It is likely, given the result of Study 4, that this instruction was partially responsible for the decrease in accuracy.

Chapter 7

Discussion

Imagine games that motivate players to engage in science, not only as research assistants (as citizen science games have explored), but as experimental subjects. I gave a name to this ambition: *elicitation games*, games that through the mere act of playing generate valid human-subject data. In particular, I was interested in the kind of human-subject data that is not directly observable or verifiable such as preferences, beliefs, latent traits, abilities, judgements, and competences. The challenge such an ambition faced was immediately obvious: human-subject data has to be valid, and games introduce many (new) threats to validity. These cannot be resolved by existing applied games approaches, which are characterised by *validation*: game designs that check the data is valid post-hoc. While powerful, these approaches apply only to the solutions to modelable problems or where it is an intersubjective consensus, not a given individual's true opinion, that is desired. In contrast, when data cannot be validated – as with the kind of human-subject data discussed here – it must be valid in the manner of its collection.

When applied games have previously been used to collect human-subject data, the assumption has been that they will motivate players to provide valid data. The assumption of validity has perhaps been justified by the very limited range of games that have been created: only a few 'safe' options have been attempted. The assumption of motivation has been borne out in the success of some of these games, such as *Sea Hero Quest* (Spiers et al., 2021). However, while we have general theories of game motivation and generalised citizen science participation, we do not know what in particular motivates the moment-to-moment provision of data within a game, let alone how these in-game motivations may threaten the validity of data collected. Further, while we might expect that games would threaten validity, no doubt variously, possibly fatally, these threats have not been systematically

studied. This leaves the data elicited by such games shrouded in murky uncertainty. Lacking confidence and justification, it is no wonder so few elicitation games have thus far been designed.

7.1 Research Questions

Having identified the problem I set out to answer the question: "How do we design games that motivate their players to produce valid data?"

The answer to this question that was presented in this thesis is the Intrinsic Elicitation model. Having first grounded our discussion in validity and motivation in chapters 2 and 3, the Intrinsic Elicitation model was presented in chapter 4. Empirical results were obtained in chapters 5 and 6 that allow us to reflect on the utility of the model, and on the relevant threats to validity that must be considered. In the introduction, I introduced four research questions to guide the research of this thesis. For each of these questions I will briefly summarise what has been learned.

RQ1 What do we know about validity in games? From surveying the literature, it is clear that the use of games give rise to multivarious threats to validity for human-subjects research. At a high level, these can be classified into four characteristic properties of games: their systemic complexity, high levels of variance, the particular social framing of gameplay, and the use of players as a sample. First, games are complex systems, involving emergent interactions of multiple parts, making them difficult to control and manipulate. Second, games are highly variant as stimuli: not only do games and occasions of gameplay differ widely (due in part to their systemic complexity and the diversity of players), but games frequently differ significantly between moments of play. While many experiments can be viewed as a sequence of identical trials, this claim is less convincingly made of a game. Third, gameplay gives rise to normative behaviour and role expectations characteristic of games that differ markedly from those of experimental situations, most notably with regard to expectations of honest or diligent responding. Finally, particular threats to validity also arise from using game players as a sample distinct to the general population.

This review addresses a gap in the existing literature. Calls for systemic systematic research on validity issues have not yet been realised (Deterding, 2016b; Williams, 2010). Indeed, previous work on threats to validity has been scattered and piecemeal (e.g. Ferguson, 2015; Hoonhout, 2008; Latham et al., 2013). For example, much of this work has been

specific either to experimental paradigms (e.g. Hoonhout, 2008), or concentrated within particular research domains, such as entertainment games user research Louvel (2018) and educational games research S. P. Smith et al. (2015). A broad view incorporating all data collection with games has not previously been attempted.

The review I have presented makes a first step at systematising these scattered pieces. While my proposed systematisation remains open to refinement by future authors, I believe the general approach is the right one. In systematising threats to validity in games must deal with those characteristic properties of games *qua* games – independent of their domain of application – that give rise to threats to validity. This is in addition to general threats to validity in experimental research (Shadish et al., 2002), and issues such as publication bias (Ferguson, 2007) and reproducability (Munafò et al., 2017) which games studies must also consider.

More research is needed to enumerate and to empirically test the threats to validity characteristic of the use of games. Research is needed to undertake the research programme on the mapping of real and virtual worlds called for by Williams (2010), and as has been attempted in part in this thesis, one may add a second research programme, concerned not just with the meta-methodological preconditions of game-based research (when and where can games even function as research instruments?), but with its *design*. In the wider field of applied games, we now have substantive literatures on how to *integrate* persuasive, educational, or motivational purposes into the design of serious games and gamified systems (Bogost, 2007; Deterding, 2015; K. Squire, 2011). In research games, though this thesis has made a start, there is much still to do. By surveying the validity threats that arise in collecting data with games, I hope to have made a first step in closing this gap.

For the development of this thesis, the review suggested that, given the complexity and variance characteristic of games, we cannot give a satisfactory account of validity at the level of a whole game – at least for any game of reasonable complexity. To the extent that gameplay varies within a game, we are unable to consider the validity of the game and its data as a whole. As moments of gameplay diverge, the there is greater need to consider the validity threats of each part of the game. If we say this variance is characteristic of gameplay *per se*, then, in the extreme, validity or invalidity must be decided at the level of individual moments of gameplay. In particular, validity must be decided at those moments of gameplay where players choose to perform actions that provide data within a game. If so, this raises the question of what motivations influence the choice of data-providing

actions in a game.

RQ2 What motivates players to choose to perform actions that provide data within a game? While many models exist why players play games as a whole, few describe why players choose to perform specific actions at a given time, and none consider the motivation particularly for data-providing actions. Therefore, through a principled generation of theory from data I developed a model of why players speak to games, as a case study for why players provide data in games. This model identified two overarching motivations that leads players to speak to a game at a particular moment of play. First, as an actuation of a mechanic. Second, as a communication of information to another player or agent. Both of these motivation sub-models are further explicated in terms of interacting components that motivate or demotivate a particular utterance within a decision process. These two sub-models approach the question of speech motivation from complementary perspectives.

The Actuation sub-model suggests that game actions that provide data, such as speech input, are motivated (or demotivated) by the specific effect they are expected to have on the game or its wider context. This might be understood as the actuation of a mechanic triggering a formal change in game 'state'. From this perspective the data (e.g. speech) acts as a 'lever' to change the game world according to the rules of the game. Alternatively, the Communication model describes how speech may effect the information state of the game or its context, particularly affecting the information available to different players of the game in multiplayer games. In such games, this wider context of player knowledge may itself constitute part of the wider 'state' of the game that is affected by speech. Forms of data other than speech might also be seen as communicative and therefore be motivated in similar ways.

From this we see that it is not so much general properties or 'game characterstics' (c.f. Garris et al., 2002) such as 'Challenge' or 'Fantasy' that 'give rise to' or 'make likely' certain kinds of data (as it is common to identify such characteristics driving learning outcomes in educational games (Graesser, 2017; T. W. Malone, 1981)). Rather it is specific desired actuations of mechanics or specific contextual communicative needs at a moment of gameplay that *elicit* (with the implication of actively drawing out) data that satisfies that need or actuates that mechanic. Thus, when I speak of a game as a whole eliciting data, it is to be understood as an aggregation of these moments of data provision. This

mode of analysis was adopted by the Intrinsic Elicitation model.

This provides a model that can provide guidance for designing novel kinds of speech mechanics in games, going beyond existing speech design patterns (Allison et al., 2018) or the straightforward 'remapping' of non-speech inputs to the speech domain Harada et al., 2011; Mustaquim, 2013. While previous research has looked at speech communication motivations in *particular* types of game (J. H. Smith, 2006), such as roleplaying games (Drachen & Smith, 2008), this adds up to only a handful of games have been studied so far in total. My research provides both a broad and grounded view of the wide variety of existing games involving speech and thus is significantly more generalisable what exists in the literature.

Many similarities were identified between the model presented here and previous studies on speech interaction in games (e.g. Allison et al., 2019; Drachen and Smith, 2008) despite these not influencing my theory generation until it was substantively complete. These provide converging evidence on the nature of speech interaction in games. In particular it essentially supports Allison et al.'s (2019) frame analysis of voice in gameplay, which extends Conway and Trevillian (2015). In both the Actuation and Communication sub-models, an integration is performed of multiple factors operating *at different levels* of the gaming situation. There are both factors *strictly within the game*: Optimality and Endogenous Value; there are factors that operate at the *interface to the game*: Effort Minimisation and Procedural Value; and finally there are factors that belong to the *social context of gameplay*: Situational Norms and Social Value. These map to Strategic World, Functional World, and Social World in Conway and Trevillian and Allison et al.'s analyses.

The model's perspective also contrasts with the dominant situational or molecular approach to motivation in games research, which predominantly adopts Self-Determination Theory (Ryan & Deci, 2000; Tyack & Mekler, 2020) that does not speak to the motivations involved in individual moments or actions. By highlighting that speech interaction motivations are best understood by moment-to-moment factors, I hope to strengthen the calls by Kumari (2021) and Melhárt (2018) for more consideration of moment-to-moment motivation in games.

For use in designing elicitation games, the study is limited as it studied entertainment games. It is unknown if the presence of recording would change motivations for speech. Furthermore, it is directed specifically at speech data and does not straightforwardly generalise to other types of data. However, within this thesis, it provided the template for the

development of a subsequent more general theoretical model as follows. In both models we see similarities in what influences the elicitation of an utterance. First, usefulness of the utterance at that moment in the game, either via the state change it effects or the information it communicates, suggests that in-game, or virtual utility of data providing actions is a significant contributing factor. Second, the social context of gameplay is significant, both in constraining and shaping 'appropriate' actuations, and in providing value (importance) to a class of social information asymmetries. Third, a significant influence in the Actuation sub-model is the effort of a particular utterance. These three: virtual utility, social appropriateness (i.e. norms), and actuation effort, became significant in development of the Intrinsic Elicitation model.

RQ3 How does game design motivate players to provide valid data during gameplay? Games design practice commonly makes use of the assumption that players act rationally in pursuit of game goals, which facilitates the creation of enjoyable strategic choices. A rational model of the player explains that particular actions in a game are desirable when they maximise 'utility' in the game. Threats to validity can be understood in similar terms: as biases that illegitimately make one kind of response more 'rational' by introducing their own utilities. A rational choice model can thus integrate game design knowledge and validity knowledge. Moreover it can apply at the move-by-move level of individual actuations or communications in the game. Thus we can use it to understand the biases on, and thus the validity of, individual datums collected in a game.

In chapter 4, this is formulated as the Rational Game User Model, an extension of J. H. Smith's (2006) rational model of the player that integrates in-game and out-of-game utilities to explain what actuations a player would make at a particular moment of gameplay. Three sources of utility and disutility were suggested that have been of particular interest. First, virtual utility, an in-game utility, is the value of actuations towards the goals of the game. This might be represented, for example, in the number of points the action would provide. Second, effort, being the work involved in making an actuation, whether as little as pressing a button or more as in speaking an instruction, is an out-of-game source of disutility, which players (all else being equal) would try to minimise. Finally, social norms, another out-of-game source of utility and disutility, motivates player actuations that fit the social role expectations of the situation and demotivates those that violate them.

Based on this model, a design approach called Intrinsic Elicitation is then articulated which ensures the actuations chosen by the player express a certain class of data without biasing it. Three principles were introduced and theoretically justified: Necessity, Centrality, and Veracity. Necessity is the principle that actuating a mechanic in the game must unavoidably give data of a particular type, such that a player cannot choose *not* to provide data if they are to use it. For instance, if a mechanic involves entering three words, the order in which these are entered is necessarily also provided. The principle of Centrality asserts that a rational game user would choose to actuate this mechanic (and thus provide data), if it is strategically central to gameplay by achieving their goals or maximising their performance. Finally, Veracity states that the desired kind of data should have a higher overall utility than other kinds, and that equally valid datums should have equal utility. An example of this is a game where the order of words inputted is of interest, if one (valid) order is more effortful to input than another, we expect the data collected to be biased.

Game studies has adopted and been guided by various, often unstated, models of the player (Thorhauge, 2003). Four models of the player that are adopted in game studies have been identified by J. H. Smith (2006). These are the Susceptible Player (whose behaviour is influenced by the game), the Selective Player (who makes choices between types of media or games), the Active Player (who engages with the game in novel ways), and the Rational Player (who optimises their outcome in the game). It is this final model that the Rational Game User Model extends. The perspective these models of the player assume shape the kinds of research questions that games research considers. In particular they affects whether it is behaviour *in* the game (Active and Rational player models) or *around* the game (Susceptible and Selective player models) that is of interest. The Rational Game User Model contributes a new perspective on the player to the literature. This perspective frames games research with a complementary set of questions to those that came before, questions set at the interface between the virtual and the real world, such as: How is the players action choice in the game influenced by the context of play? and How does the player synthesise the dichotomy between virtual and real-world motivations? I have suggested that this perspective is particularly relevant to applied games research.

The Rational Game User Model is intentionally simple, which is good for generating clear and concrete design guidance. Moreover, the assumptions it makes underlie much existing game design practice, such as the essentially goal-directed character of gaming, as opposed to playing (Deterding, 2015; Juul, 2005), and that players can be thought

of as rational agents (J. H. Smith, 2006). However, these assumptions are known to be over-simplified. We know that players gain game enjoyment from more than just playing optimally, such as from curiosity, surprise, engrossment and relatedness (Boyle et al., 2012), and that rationality, according to behavioural economics, is bounded by biases and heuristics (Camerer & Loewenstein, 2004). In addition, systematic causal effects may exist between individual factors that have not been specified – e.g., self-determination theory would predict that adding tangible rewards may undermine intrinsic game enjoyment (Deci & Ryan, 2000). As such the model may need to be cautiously extended or adapted. However, theory is more pragmatically powerful when it needs fewer starting assumptions and data to make good-enough predictions (Healy, 2017). Its extension should be based foremost on what is required address real observed empirical limitations that make it less useful in practice.

This model explains validity in terms of the motivations that affect the player for the interaction that provides data. Validity of a game as a whole is understood as an aggregate of the validates of the individual instances of data provision. Because each data providing interaction is affected by various motivations, these must be understood in order to understand the validity of the game data. The Rational Game User model calls for the integration of game-internal and -external sources of utility into a single rational choice. However, it is underspecified with regards to which sources of validity must be considered. It suggests that virtual utility (score or chance of winning in the game) should be a major factor arising from within the game. It suggests that actuation effort (the effort to trigger a particular mechanic actuation) should be a major game-external factor. Other possibilities are raised such as social norms, which can be understood experimentally as demand effects. An empirical investigation of biasing factors was therefore warranted.

RQ4 What are practically relevant threats to validity in elicitation games? I ran a series of experimental studies to specify which motivations are the most significant in the Rational Game User Model, and thus most important in determining the validity of game-elicited data. The first two preregistered studies, reported in chapter 5 ($n = 96$, $n = 139$), one a conceptual replication of the first, compared a game for eliciting adjective order against a practice-as-usual control. The game was designed using Intrinsic Elicitation to control threats to the validity of its data. The studies found that the game was more enjoyable than a practice-as-usual control but also provided less accurate data.

These studies empirically tested the intuition that games, possessing characteristic threats to validity such as complexity, variance, and framing, are indeed likely to reduce the quality of data provided to a game relative to using an practice-as-usual experimental task. As the studies indeed found that there was a decrease in the accuracy of data in the game compared to the practice-as-usual condition, they support the claim that games are likely to affect validity, even when designed to avoid this, and thus systematic research to identify and empirically quantify validity issues in the use of games is needed to understand and control for these.

Studies 1 and 2 controlled virtual utility and actuation effort, two proposed validity threats considered in the Rational Game User Model. The data they collected had an accuracy significantly better than random: there was signal in the noise. However, the significant decrease in accuracy in the game suggests that there are other factors of primary importance when using the model. I noted that studies in the literature making use of *gamified* experiments generally find that data quality is maintained or improved (Friehs et al., 2020; Hawkins et al., 2013; Keusch & Zhang, 2017; Levy et al., 2016; Lumsden et al., 2016), in contrast to the findings reported here. I suggested it may be that demand effects of the experimental situation are responsible, as this also followed from the differences observed between Study 1 and Study 2.

I tested whether demand effects affect accuracy in game data in two ways. In Study 3, inspired by the large effect of 'framification' observed by Lieberoth (2015), I manipulated the metacommunicative framing to stimulate demand effects. There was no observed effect, suggesting that, in the already highly framed context of online experiment participation, metacommunicative framing is unlikely to have a significant impact on participants (and thus was not responsible for the effect seen in Studies 1 and 2). In Study 4 I gave some participants explicit instructions to provide data to the game in a particular form, a maximal positive control to identify how significant demand effects might actually be. This led to a significant increase in accuracy of $d = .33$; important, certainly, but not the whole story.

Demand effects are threats to validity that might impact games via instructions but not (in the online situation used) via labelling and metacommunicative framing. Gamification approaches apply game design elements but do little to seriously disguise the purpose of the game. In contrast, elicitation games, whole games that seek to generate data 'effortlessly' by mere play, seek primarily to engage players in the game for its own sake. For example,

while players of *Sea Hero Quest* (Spiers et al., 2021) might be aware of and motivated by the scientific data provision aspect, the mapping between in-game behaviours and outcomes is not obvious, and the gameplay is compelling enough to stand alone. Thus, as gamified experiments and surveys are likely to be more direct in their instructions to participants than elicitation games, gamification may be both a simpler and more effective method for eliciting valid data. Indeed, this might in part explain why gamification generally see maintained or improved data quality.

7.2 Evaluating Intrinsic Elicitation

By drawing together the insights from the preceding chapters, an answer – or a framework for an answer, with more research required – can be provided for the overall research question of this thesis: "How do we design games that motivate their players to produce valid data?". This answer takes the form of the Intrinsic Elicitation model, presented in chapter 4, further understood in light of the four subsequent empirical studies.

Intrinsic Elicitation gives three principles to be used in applied game design and evaluation intended to be sufficient to motivate players to produce valid data. These principles are:

1. **Necessity** Data provision is an unavoidable part of actuating one or more of the mechanics in the game
2. **Centrality** Such data providing mechanics are strategically central to the game
3. **Veracity** Different equally-valid data providing actuations are of equivalent (and maximal) overall utility

Evaluations of these principles rely on the Rational Game User Model, a rational-choice model of player actions where each possible mechanic actuation is assigned an overall utility from on both in- and out-of-game factors. While Necessity and Centrality can be considered primarily concerns of efficiency – the rate at which a game can elicit target data – violations of Veracity, whatever the factor responsible, can be interpreted as threats to validity. I have so far justified three factors to incorporate: virtual utility, actuation effort, and social norms.

Within this section, I will draw together evidence from the whole thesis and my experience of developing elicitation games to evaluate Intrinsic Elicitation and identify weaknesses

and places that need further development. I will structure this by discussing each of the principles in turn: Necessity, Centrality, and Veracity.

7.2.1 Necessity

Satisfying Necessity requires that it is not enough for a mechanic *to be able to* elicit data (perhaps evidenced by playtesting), but rather the actuation of the mechanic must *inherently require* providing the data. In this way, the use of a mechanic satisfying necessity should be sufficient for data of the desired type to be elicited. *Adjective Game* satisfies Necessity, as any actuation of the core block-clearing mechanic of the game requires an ordered sequence of words to be supplied, from which the relative order of two adjectives can always be determined. One could imagine variations on this game that would violate Necessity, such as if the player had the choice of how many words to enter.

Three limitations of Necessity arose in my experience of designing this game. First, it requires the designer make several assumptions and does not give guidance for these. Second, ideating mechanics satisfying necessity is challenging, and tools or approaches for ideation are motivated. Third, it excludes game designs which in practice still collect the desired data. I address these in turn.

Assumptions What do we need to assume about play for a given mechanic to satisfy Necessity, and what is reasonable to assume? *Adjective Game*, for example, could be played by a bot rather than a human player. In this case the data we collect would only have the surface appearance of encoding adjective orders; it would not exhibit the human behaviour which I claimed was an unavoidable consequence of play. Yet this seems a reasonable assumption to make – it is unlikely that the experiments reported here were affected by bots (yet were the experiments run on MTurk, for instance, we might be more concerned (Chmielewski & Kucker, 2020)). On the other hand the assumption that players will always do exactly as you want is clearly not reasonable. What defines whether an assumption is reasonable?

Any claim to Necessity is dependant on assumptions. First, we assume something about what it means to play the game (e.g. do we consider the possibility of players adapting the game rules to make 'house rules', potentially eliminating our data-providing mechanic?). Second, we assume something about the expected context of such play (e.g. what skills and competencies are we assuming players posses for their actions to encode meaningful

information?). Such assumptions are baked into our claims to Necessity, which ultimately manifest in our claims to validity of our results and the generalisability of the game as a research tool.

As a **design principle**, use of appropriate assumptions is important for the success of the game design at collecting the desired data. While an assumption that players will not cheat may serve in a multiplayer board game, it may be inadvisable in a single-player digital game. Designers need to work with appropriate assumptions to minimise the likelihood of failure. Knowledge of what assumptions are reasonable might be approached systematically. Much could be likely incorporated from existing literature on experimental design and game studies. The synthesis of such literature aimed at the designer of elicitation games would be helpful to guide design.

As an **analytic principle**, the use of appropriate assumptions moderates what claims to generalisability a game can make. The assumptions made in asserting Necessity limit the contexts where one might expect that game to be effective. While an applied game need not generalise outside of its intended context of application, if that context is excessively narrow (e.g. to use only within a controlled experiment) it would threaten the ecological validity claim the game makes to being a whole game (if such a claim is necessary). Benefits of developing a standalone game, such as scalability, necessitate a resilience to common contexts in which such games are played. One again, a systematisation of such knowledge from the game studies literature would be helpful to those analysing and interpreting the results of elicitation games.

Playtesting with the intended audience of the game may be one way to surface what assumptions don't hold in real life. Though, to keep with the cautious spirit of the principle, this should be approached with the spirit of falsification: attempt to falsify claims to Necessity (i.e. disprove an assumption), and gain confidence in confirmed assumptions only so far as they have endured such testing. In some cases it might be efficient to direct testers to attempt to violate the assumptions, such as if our assumption was, for example "there is no way to trigger this speech recognition system without providing such-and-such an input."

Mechanic Evaluation and Ideation Developing an enjoyable game mechanic that satisfies Necessity is the first challenge for the design of an Experiment Game to overcome. From my own design experience, this was the point at which the greatest proportion of

seemingly promising ideas failed. As such I needed to ideate a lot of game mechanics likely to satisfy Necessity. At present, the Intrinsic Elicitation model does not suggest any means of doing so. Without this, it is hard to tell how widely applicable the model is; are there domains where no such mechanics exist? Developing tools to guide mechanic ideation would be helpful for the design of future elicitation games. I suggest a number of starting points for this based on my experience of designing elicitation games.

The principle of Necessity was most useful for *counterfactual thinking*, the process of evaluating hypothetical designs based on a mental model (Oulasvirta & Hornbæk, 2021). Its simplicity allowed it to act as a fast filter on ideas: likely failing ideas could be discarded quickly without consuming development time. However, this required knowing what data was Necessary in a given interaction, for which I employed domain-specific knowledge. For my linguistic case study, this meant considering the purpose the language structure serves in communication more than replicating experimental paradigms. For example I did not approach *Adjective Game* by gamifing a picture description task, rather I was inspired by the formal semantics of adjectives and their role in defining sets. I would imagine that similar domain-specific knowledge would be required to design e.g. what mechanics in a social game might encode personality traits. As such, domain-general tools for ideating elicitation game mechanics may not be very useful in the absence of domain-specific knowledge.

Performance-conditional rewards are in-game rewards dependant on measureable player performance. Where possible, these might be employed in the identification of elicitation mechanics as some variant of 'perform the task as well as you can'. For example, a game to identify colourblindness could involve catching only coloured balls of a particular colour. A game to measure working memory could involve memorising an increasing number of shapes. Similarly, because *Sea Hero Quest* (Spiers et al., 2021) can measure player performance at navigating efficiently, the data elicitation mechanic can be 'navigate efficiently to the target'.

It would be helpful to delimit in what games performance conditional rewards are likely to be possible. While we can think of this case-by-case in terms of likely threats to validity claims, that is relatively laborious and requires domain-specific expertise. Hypothetical examples where performance-conditional rewards are plausible seem to involve data that is essentially a distance from an ideal value (e.g. closeness of a reaction time to 0, accuracy of input to expected answer, closeness of jump distance to infinity). In many cases these

seem to arise from normative value judgements. For example, it is seen by educators as more desirable to answer a maths question accurately than to express a creative answer[1]. The difficulty I found in identifying mechanics satisfying Necessity may have been because contemporary linguistics – my case study – prises descriptivism, meaning it does not place value judgments on performance. On the other hand, economics, which largely assumes the acquisition of wealth is desirable, is likely to be generally more amenable to performance-conditional rewards.

Unwilling to use performance-conditional rewards, the game mechanic that I settled upon was one where the feedback dimension was orthogonal to the data being collected. This was achieved in *Adjective Game*, where the *set* of words selected drives the feedback system, but the *order* of categories is the data desired. It may be that this 'principle of orthogonality' can be generalised to other kinds of data. For instance, imagine a game to elicit ice cream flavour preference where you run an ice cream parlour. In this game, each day players to select 3 ice creams to sell from different categories: e.g. single scoop, double scoop, sorbet, gluten free, etc. Each category of ice cream contains a range of possible flavours. Players try and make the selection that will make the most money on a given day. In this game we could collect data about *flavour* of ice cream selected (e.g. chocolate vs. vanilla) while driving game feedback systems based on the *category* of ice cream selected.

Exclusion of valid designs One game I designed but did not report on in this thesis showed me that sometimes games reliably provide relevant data without this being strictly necessary. This was a case where players do not *need* to provide data to actuate a mechanic but, in testing, did so anyway. The game was a multiplayer card game called *Pastry Chefs*. In this game a player might instruct another player (e.g. John) to give a card that they hold (e.g. the two of hearts) to another player (e.g. Mary). To do so they might say "John, give the two of hearts to Mary" (prepositional construction), or "John, give Mary the two of hearts" (double object construction). As such they have produced a one of the two structures that an English dative expression can take (Oehrle, 1976). Yet, while both of the above are natural ways to actuate this mechanic, it is not *necessary* to use one of them. Saying only "two of hearts" accompanied by some pointing and grunting might

[1]Some mathematicians might assert that this is a failing of contemporary educational practice (Lockheart, 2008)

be sufficient to convey the same instruction. Yet although this mechanic does not satisfy Necessity, playtesting suggested it is a promising candidate to develop into an elicitation game. The data I collected with it indeed showed a large number of dative expressions, of both types.

Such over-exclusion represents an inefficiency in the articulation of Necessity as a design principle. However, how can we tell the difference between a Necessity-violating mechanic with potential, and one that is a waste of time? Simple playtesting is one option. When doing so, it pays to regularly test with players who are unaware of the desired data. Game analytics (data gathered from real play) would be reliable but is not possible until a significant amount of development work has already been expended – and potentially wasted if players don't behave as desired. Neither serve as an analytical principle, however, and cannot be used to think counterfactually about possible game designs.

7.2.2 Centrality

Centrality is the principle that the data-eliciting mechanic must be strategically central to the game to be frequently actuated (or actuated at all). Like Necessity it is straightforward to understand, which is a strength in using it for design. In *Adjective Game* centrality is satisfied by default as the data providing mechanic is the only mechanic available in the game. Such a solution avoids much complexity otherwise inevitable from balancing the relative utilities of multiple mechanics.

For a mechanic to be strategically central, there must first be strategy – desirable states and undesirable states. In the Rational Game User Model, player actions are primarily positively motivated by their virtual utility. Actuations can only have virtual utility – and thus motivate their actuation – if some ways of acting (or abstaining from action) are better than others. For example, different inputs in *Adjective Game* lead to differing success in clearing the level and scoring points, yet were there no difference between acting and watching the game play itself, the mechanic would unlikely be used (and the game swiftly abandoned). As such, Centrality remains a concern for design, even in a game with only a single mechanic.

The creation of strategy summarises a significant amount of game design. In *Adjective Game* this was relatively straightforward: no other mechanics were required so it was merely necessary to design a compelling puzzle dynamics. Little of this process was specific to elicitation games: the entertainment game *Two Dots* (Playdots Inc., 2014) served as

primary inspiration for *Adjective Game*. It may sometimes be necessary to introduce secondary mechanics which are balanced in virtual utility to create compelling choices, as was the case with the game *Pastry Chefs*, mentioned above. In that game, cards were only valuable to pass between players if they could also be played as an action to score points. Such secondary mechanics need not compromise the principle of Centrality so long as they are less frequently the strategic choice.

Centrality can be evaluated through playtesting. Though this proved unnecessary in *Adjective Game* due to its single mechanic, this was effectively employed in the development of *Pastry Chefs*. To the extent that playtesters correctly identify the optimal strateg(ies) of a game, their evidence is a reliable predictor of other players' behaviour within the (idealised) player assumptions of the Rational Game User Model. In some circumstances evidence from automated game testing (and in more limited circumstances – e.g. some puzzle games – a formal proof) might demonstrate that a particular mechanic is essential for the completion of the game.

7.2.3 Veracity

Veracity is the principle that different equally-valid data providing actuations should be of equivalent overall utility, and must be of greater overall utility that invalid data providing actuations. For example, in *Adjective Game* there is no difference in virtual utility or actuation effort between different orders. Randomisation is used to avoid any minor differences in effort that may arise from positions of words on the screen threatening the validity of the data collected, though at a cost of increased variance.

Veracity depends on intended use, the specifics of presentation, and potentially endless theoretically possible biases. As such there is a need to identify the empirically most significant biases to focus on in design and analysis. I have suggested that effort and virtual utility would be two key potential drivers of invalidity. However, I have also demonstrated that these alone are insufficient. In comparing *Adjective Game* to a practice-as-usual experiment implemented in the same interface, the game collected data that was less valid (operationalised as lower accuracy). This doesn't mean that the game lacked Veracity *per se*: there still seemed to be a significant bias towards accurate responding, but this was not as strong as in an equivalent experimental task.

One further threat to validity considered in this thesis was demand effects, which are incorporated into the Rational Game User model as social norms, an out-of-game

extrinsic source of utility/disutility. Two approaches were taken to evaluating the influence of demand effects. First, manipulating metacommunicative framing did not have an effect on accuracy, suggesting the observed decrease in accuracy in previous experiments was not due to this. Second, direct instructions did have a significant effect on accuracy. This effect was smaller than the decrease originally observed, so it is but one additional factor.

Social norms have thus been shown to demand consideration when evaluating Veracity. However, it does not seem necessary to consider the surrounding metacommunicative frame – at least if the game is delivered in the context of an online experiment – as part of these social norms as the effect of these on individual data-providing actuations is likely very small. The effects of stronger metacommunicative frames, such as might be present in an in-person experiment, must still be empirically tested.

In contrast, it does seem necessary to consider the socially normative effect of direct instructions when evaluating Veracity. The presentation of a clear experimental instruction to answer in a particular way was shown to have a significant effect on individual data-providing actuations. Elicitation game designers can then consider how instructions can be used to bias players towards desired types of data. As this was not observed to have an effect on enjoyment, such benefits in the validity of data may come at little cost. Instructions that conflict with strategic or low-effort play, however, might be expected to be less effective at improving validity (due to being overpowered within the Rational Game User Model) and I suspect also more likely to harm enjoyment, though this hypothesis has not yet been empirically tested. Until then, it would be prudent not to expect too much of instructions.

In designing with Veracity we make use of the Rational Game User Model as a theoretical tool: a mental model within which we can evaluate hypothesised designs through counterfactual thinking (Oulasvirta & Hornbæk, 2021). As such we will not (often) have quantitative data describing the *relative* influence of the biasing factors in the model, particularly not within our specific game. Two approaches to ensuring Veracity can then be undertaken. First, we could undertake the empirical work to evaluate the relative strength of the biases within our game or a closely related game – maximal positive controls might for instance show that in our game even strong instructions have no effect on validity. Alternatively, we could justify a purely theoretical claim by arguing that Veracity is achieved for every biasing factor in isolation. For instance, both virtual utility and effort might be argued to each be equivalent between different data-providing actuations, while social

norms might be argued to be biased towards our desired category of data.

In addition to the factors discussed above, other potential biases might be incorporated into the Rational Game User Model, and thus our determinations of Veracity. These might include opportunity cost, meaningfulness, financial rewards, novelty, and many more. At present the elicitation game designer lacks such guidance as to which of these they should include in their evaluation of Veracity. To address this, further controlled experiments are needed to evaluate the impact of the above biases in different game contexts, from which at a later stage empirically justified design guidance might be developed.

7.2.4 Summary

Intrinsic Elicitation is a design approach for elicitation games. It contributes a framework for collecting a type of data – human-subject data – that is not compatible with existing models and templates of data collection with applied games. While the collection of such data with a game can be understood in general experimental terms, Intrinsic Elicitation extends the meta-methodological literature with a model and design approach tailored to the issues present when using games. It integrates game design and experimental concerns into a single model. I have demonstrated it as an analytical tool and as a framework for interpreting my experimental studies. My hope that it will prove useful in the design and analysis of elicitation games.

While I have not performed any design evaluation of the Intrinsic Elicitation design approach, from my own design work, I found the first two of its three design principles (Necessity and Centrality) straightforward to understand and implement. The Veracity principle proved more difficult to realise as the principle itself provided very little practical guidance as to what factors to include and it presently offers no pragmatic stopping point for when different actuation options for the data-providing mechanic can be considered reasonably balanced. How utility-balanced is *enough* for the mechanic actuation to be sensitive to the latent property we wish to elicit?

I think these issues stem from the fact that the approach is really an articulation of sensible principles or design goals, but lacks underlying methods that would walk a practitioner towards accomplishing these goals. As enjoyment and validity are integrated (and the responsibility for each cannot thus be separated between game designer and researcher), these tools must be accessible to game designers. Thus, future work might expand on the Intrinsic Elicitation approach with such methods and processes.

7.3 Generalising the Intrinsic Elicitation Model

In the introduction to this thesis I situated my research at the intersection of Applied Games and Games for Human-Subject Research. Within this intersection I introduced elicitation games. Intrinsic Elicitation was developed as a model for the design and analysis of such elicitation games. The characteristic constraints imposed by that intersection guided the resulting model, leading to a rational-choice-based moment-by-moment account of motivation and behaviour. Now this model has been created I will reflect on its potential utility in other domains.

7.3.1 Applied Games for Data Collection

Existing methodologies for collecting data from applied game, such as scientific discovery games and human computation games collect data of different a different *type* than elicitation games. However, the novel perspective of Intrinsic Elicitation might prove instructive for their design.

Scientific Discovery Games such as *FoldIt* make use of a computational model to reward players relative to the value of the data they provide (Cooper, Khatib, et al., 2010). *FoldIt* can be well understood at a macro level: the data is the net effect of many mechanic actuations. However, following Intrinsic Elicitation, we could switch focus to those individual mechanic actuations. Why are players motivated to use a particular tool (mechanic) at a particular time? We can understand this a combination between virtual utility (neatly encoded in the score of the computational model), and external utilities such as actuation effort. This suggests that were mechanics that were highly effortful to use (or violated social norms, etc.), they may be dispreferred despite leading to higher quality solutions. Designers of such games might therefore consider not only virtual utility (computational scoring model) but also external utility factors such as effort, social norms, etc.. Not only for the games a whole (e.g. the benefit of meaningfulness for contributing to science), but for individual mechanics in the game, for the sake of efficiency.

Human Computation Games such as *The ESP game* (von Ahn & Dabbish, 2004) often rely on intersubjective consensus to validate the data they collect (Quinn & Bederson, 2011). Such agreement mechanisms have been understood as achieving *validation* at the level of the whole game. However, applying Intrinsic Elicitation, we can also the look at

agreement mechanisms in an elicitation game for achieving per-user *validity*. This would be appropriate if the data we sought from a player was their (individual) expectation of the consensus belief, rather than their unbiased personal belief. For instance, in an economics or a market research study, we might wish to know not what (monetary) value an individual places on a product, but what they believe other people would pay for it.

To collect this data, an agreement mechanic must elicit data from each player and the use of this mechanic must be central to the game, corresponding to Necessity and Centrality. Second, the desired data (e.g. honest prediction attempts of other peoples' monetary value judgements) must be actuations of this agreement mechanic with of higher overall utility that undesired data. To achieve this, data matching the intersubjective consensus can be rewarded to ensure it is in the rational interest of the player to provide data that they believe will agree. So long as the other utilities (e.g. social norms, effort) contributing to a player's actuation satisfy Veracity, an agreement mechanic would be expected to collect valid human-subject data for this purpose.

Games for Educational Assessment Educational games often incorporate assessments, allowing e.g. teachers to observe the performance of their students. As such they elicit human-subject data, generally for which performance-conditional feedback can be provided to motivate players to engage. Under Veracity, this feedback is liable to bias player responses towards the 'correct' answer. However, in some domains this bias may not be not desirable, for example in the assessment of a creative art form.

An assessment in an educational game for a creative subject like music composition likely cannot provide performance-conditional feedback, yet it might be that a game could motivate students to e.g. compose music if they otherwise were unmotivated to do so. If used as an assessment, educators need to be able to fairly assess student ability, without this being obscured by the design of the game. We might therefore consider how such a game might systematically bias players inputs. For example, if the game rewards players for composing in C major, this might disadvantage a student who would otherwise have preferred G minor[2]. As a framework to analyse such biases, we could adopt the principles of Intrinsic Elicitation.

More broadly this issue relates to 'gaming the system', which has been defined in education as "attempting to succeed in the learning environment by exploiting properties of

[2]Let alone atonal music in free time.

the system rather than by learning the material and trying to use that knowledge to answer correctly" (Baker, 2005, p. 6). Within the Rational Game User model, we can model players as being biased towards a correct answer as the most utility-optimising input, irrespective of what learning they are supposed to be demonstrating. Such a situation can be understood as a violation of the Intrinsic Elicitation principle of Necessity: use of the data elicitation mechanic (the test) in this case does not necessarily encode information about learning. Intrinsic Elicitation would suggest – though I am conscious of oversimplifying an entire field of research – that novel mechanics could be identified that seek to satisfy Necessity under the assumptions of classroom learning. Such mechanics need not appear to be assessments; indeed in principle, any mechanic that encodes the desired data could be used. The challenge would be in convincing educators of the construct validity of the elicited data, for which Intrinsic Elicitation may provide a framework.

7.3.2 Human-Subject Research

Orne (1962, p.777) remarked at the great power describing an activity as an experiment had. The response he received to asking someone do do 5 pushups was normally "Why?", but if framed as an experiment, the answer was "Where?". Describing an activity as a game has a similar power. A quality assurance factory worker may expect to be paid to put confectionery products into matching sets but be happy to pay to play *Candy Crush Saga* (King, 2012). Yet while experiments are expected to be purposeful, games are expected to be fun.

We can understand lab or online experiments as (mostly) un-fun games. As games, players are situationally motivated to take one action or another, provide one response or another, based on the contextual utilities of each at that moment. Perhaps they start off motivated by meaningfulness, but over time their motivation drifts, spending the time to provide accurate data becomes an opportunity cost, perhaps the experimenter steps out of the room and the beneficial influence of social norms on data quality goes with them.

Typically experiment designs are analysed and reported as a whole rather than as a succession of moments of participant experience. The change in perspective is perhaps interesting, particularly if there are phenomena easy to miss at the whole-experiment level that occur at the moment-by-moment level. Consider motivation. While motivation is important for data quality, it is usually only considered for recruiting and retaining participants. Yet a well-motivated participant might still spend some period of the experiment

experiencing boredom and, for that time, provide poorer quality data. Summative self-report measurements, such as of motivation or enjoyment, if delivered only at the end of the experiment may obscure the variation in this variable over time. Such variation could in cases prove a potent threat to validity, particularly if participants are their own controls.

Game design could be used as a lens when designing an experiment. Whether or not the experiment is intended to be enjoyable, this lens could be used to cut through the assumption that players will even be attempting to provide the data we want. They might on the contrary be rationally motivated to give poor-quality data. A game designer would not see that as a problem with the participant ("Unmotivated participants give poor quality data"), but with the experiment ("Unmotivating experiments collect poor quality data").

7.4 Summary

Intrinsic Elicitation can be applied both as an analytical tool for post-hoc theoretical evaluation or justification of an elicitation game, and as a design tool for novel elicitation games. My aim was to generate simple, useful, and unnuanced theory that can be most effectively used to guide design. While this model has primarily been justified for use to design and evaluate games for eliciting human-subject data – elicitation games –, I have summarised above some implications of adopting the same perspective in adjacent fields. It is easy to make suggestions, of course. The hard work of developing and extending the Intrinsic Elicitation model – as well as applying it in practice – remains to be done.

Chapter 8

Conclusion

When a scientific discovery game like *FoldIt* (Cooper, Khatib, et al., 2010) rallies tens of thousands of citizen scientists to solve scientific puzzles, it should not matter if some players produce bad solutions. When a human computation game like *The ESP Game* solicits 1.2 million image labels from 13,000+ players in four months (von Ahn & Dabbish, 2004), we want to be able to trust the consensus is not a fix. When an elicitation game like *Sea Hero Quest* (Spiers et al., 2021) charts the course of its four-millionth sailor, we do not want to worry about whether this one is playing at home or on the bus. For each type of game – for each type of data – we want to be able to justify the inferences we draw from it and avoid threats to validity. We can do so only if certain criteria for ensuring validity of the data are met.

What are these criteria? The above cited games above illuminate three *types* of data, each requiring different justification to ensure their trustworthiness. If we can express what it means for our data to be valid in a formal, way, then we simply *validate* the data we collect. If, on the other hand, we want to gather an intersubjective consensus then we may use agreement-based game mechanics (or similar) to validate it as we collect it[1]. Finally, when our data is about non-observable properties of individuals, such as latent individual dispositions – human-subject data – we must craft the rational motivations within the game such that the major sources of positive and negative utility, within and outside the game align with the data we want. In short, we want to align quality data provision with the fun of gameplay. This final strategy was formalised within this thesis as the Rational

[1]We must justify our assumptions in each case. In the former case, we must justify why we think the computational model is accurate. In the latter, why it is not possible for players to circumvent our agreement mechanics.

Game User Model and a design approach called Intrinsic Elicitation was synthesised that interprets this as three simple design principles: Necessity, Centrality, and Veracity.

When collecting human-subject data, we naturally want high *quality* data – data that is fit for use (Tayi & Ballou, 1998) which often requires it have a minimum of unwanted statistical variance. On the other hand, a certain *quantity* of data is required – sample size, in experimental parlance – which comes at a cost, most often financial (Faber & Fonseca, 2014). These are related: data with higher statistical variance requires a larger sample size to detect a statistically significant effect on our measure of interest. More concerning is the potential for such variance – if it is not randomly distributed between conditions – to threaten internal validity. Yet to an extent this is inevitable: any approach to increase sample size that reduces experimental control is liable to increase unwanted variance.

Applied games for the data collection has been a solution of quantity over quality: motivate participation at the cost of controlling the precise manner of collection, or, in other words, trade the potential for increased sample size against increased unwanted statistical variance. Indeed, characteristic properties of games *qua* games give rise to potential systemic threats to data quality, be it from the systemic complexity and variance of games themselves, or from the diversity of players and playing situations. This was discussed theoretically when surveying known and potential threats to validity in data collection games in chapter 2. It was empirically supported in two experiments in chapter 5, which found that, while I had followed a theoretically justified approach to minimise threats to validity (chapter 4), the design of an experimental task into a game led to decreased accuracy of data provided. This suggests a trade-off in the use of games for data collection between enjoyment (and the actual and potential benefits thereof) and data quality.

Scientific discovery games, such as *FoldIt* (Cooper, Khatib, et al., 2010) and human computation games such as *The ESP Game* (von Ahn & Dabbish, 2004) have addressed this trade-off using a motivate-then-validate approach that I characterised as gamification+validation. In this way the potential threat to validity that this unwanted variance represents is avoided by post-hoc validation of the data, while any number of motivating game features can be incorporated to maximise enjoyment. The variance itself, so long as it does not significantly harm the efficiency of data collection, can be left unchecked, an ancillary concern to the main thrust of research.

In contrast, to collect human-subject data in an elicitation game we require a quantity-

of-quality approach. This characterises the Intrinsic Elicitation design approach presented in chapter 4. Intrinsic Elicitation gives us a framework to understand the integrated concerns of motivation and validity in an applied game to collect human-subject data. Of course, merely adopting a framework does not make the trade-off disappear; proof if this if needed was shown in the experiments of chapter 5. However, it gives a framework for research into what specific (and ideally controllable) factors are responsible for this trade-off, and a design approach for avoiding them once identified.

My first two experiments (chapter 5) showed that there was a decrease in data quality in an elicitation game, and suggested that this may be to do with the different social norms and roles associated with gameplay in comparison to experimental participation. This was tested in two experiments in chapter 6 though the results were equivocal. While manufacturing demand characteristics through metacommunicative framing was unsuccessful at achieving an effect (suggesting significant limits to metacommunicatively framing a game in an already strongly framed online experimental platform), merely giving direct instructions was enough to increase data quality in the game. Effect sizes indicate that this is far from the full story, however.

This thesis argued that elicitation games *qua* games – and perhaps unlike the gamification of experimental tasks – threaten data quality, though do not necessarily threaten the validity, of human-subject data.

In this final chapter, I next summarise its main contributions in section 8.1. I then discuss general limitations in section 8.2. I conclude with some directions for future research in section 8.3.

8.1 Contribution

People who make experiments look at them as experiments – even if they are adopting surface features of game framing to motivate participants to perform arbitrary behaviours within a closed system. People who make applied games look at them as games. Yet what is an experiment but a poorly designed game? What is a game but a chance to observe the latent competences and preferences of the player? This thesis set out to bridge the gap between these two communities of practice with an integrated approach to understanding games for data collection as *both* games *and* experiments simultaneously.

Elicitation Games In chapter 1, I identified elicitation games as a class of data collection game with unique challenges. Elicitation games are games designed to collect human-subject data: non-observable or latent properties about individuals. Such games are particularly vulnerable to threats to validity as the data they collect cannot be validated post-hoc. There are no previous design templates and taxonomies available for these kinds of games, as these are based around the idea of validation.

Threats to Validity with Games In chapter 2, I provide the first review of threats to validity characteristic of using games. Games are complex and involve emergent behaviour. Games very significantly, both between and within games. Gameplay adopts a particular social framing distinct to experimental participation which sees behaviour as less consequential. Finally, players of games are diverse and distinct from the general population. These threats to validity characteristic of games extends to all quantitative uses of games in data collection, including elicitation games.

Speech Motivation In chapter 3, I developed a grounded theory of speech motivation. Speech in or to games is motivated in two ways. First, as an actuation of a mechanic to efficiently achieve a desired effect, while acting in a way that is socially appropriate. Second, as the communication of information to resolve a perceived, valuable information asymmetry within the constraints and entitlements of the game rules. While this model contributed to the development of this thesis, it is also useful for those who want to design games using voice as an input medium, or collect spoken linguistic data with a game. It also provides converging evidence with studies on voice and player communication in games about on factors affecting in-game speech.

Intrinsic Elicitation In chapter 4, I developed a theoretical model that integrates applied games knowledge and social science methodological knowledge. The Rational Game User Model is a rational-choice model of the player, built on and extending the work of J. H. Smith (2006). Within this model, the player is a rational agent who selects mechanic actuations to maximise combined game-internal and -external utility. In-game – virtual – utility is assumed to contribute positively to game enjoyment. Intrinsic Elicitation draws out three design principles from this model for ensuring data collected by an elicitation game is valid: Necessity (a mechanics actuations must encode data), Centrality (that mechanic must be strategically central to the game), and Veracity (the most

utility-maximising actuation within the Rational Game User Model must be the one desired). This provides a framework for designing and analysing elicitation games and for theoretically justifying their validity.

Accuracy and Enjoyment Trade-off Two experiments in chapter 5 empirically supported the intuition that elicitation games would lead to a trade-off with data quality. In comparison with an experimental control task, an elicitation game designed to elicit adjective orders was more enjoyable but collected data with lower accuracy. Yet the data it collected was significantly more accurate than random. This informs the ongoing debate on whether game features affect data quality. In particular, it suggests that the use of whole games might reduce data quality, which may stand in contrast to the use of gamification.

Demand Effects and Accuracy One apparent reason for this trade-off was the difference in social norms between (whole-game) gameplay and experimental participation giving rise to demand effects in the experimental control task. The experiments in chapter 6 tested this by adding experimental metacommunicative framing and direct instructions to a game. The only significant effect was that direct instructions improved accuracy. While games may harm data quality if they do not give the same kinds of instruction that experiments do, much of the trade-off between accuracy and enjoyment evidenced in chapter 5 remains to be explained by other factors.

Overall, this thesis has suggested that, under typical assumptions of gameplay, games generate their own potential systematic biases to data quality. In other words, games generate incentive to misalign behaviour with desired data. Intrinsic Elicitation provides a framework for research and design to address these issues. As biasing factors are identified and incorporated into the model of the player, the success of games that satisfy Intrinsic Elicitation at collecting valid data should improve.

8.2 Limitations

The approach of this thesis has been to identify the hardest possible case study. Among human-subject datums, intuitively the hardest to elicit validly are latent properties, preferences or beliefs. Adjective order is one of these. This makes the Intrinsic Elicitation model all the more interesting as it has been demonstrated with a type of data that is rarely

collected with a game, opening the possibility of other games to collect similar datums that have so far been completely overlooked by applied games research. Whether similar results would be achieved with datums participants are relatively more conscious of – and may consciously falsify – such as political beliefs, is not yet known. My results apply only to the 'worst case' – they may not extend to competences, such as typing speed, as here it would be possible to give performance-conditional rewards. Valuable future work would be to extend the approach taken here to relevant but unexplored data types and context, including knowledge (e.g. geography trivia), competences (e.g. reaction times), properties vulnerable to dissimulation (e.g. political voting intention) and non-conscious properties (e.g. neuroticism).

8.2.1 Limitations of the Case Study

Throughout, this thesis has used linguistic data as its primary example of human-subject data. The experiments have used a single, novel game *Adjective Game* for eliciting this data. This poses a threat to the generalisability of the results presented here to other games and other datums. Is the trade-off between enjoyment and accuracy common to all elicitation games, or just those that elicit linguistic data (or adjective order in particular), or indeed limited to just *Adjective Game* itself?

Linguistic data may be particularly sensitive to instrumental behaviour of players, and thus more liable to observe a trade-off, because of the resilience of natural language semantics to violations of its structural norms. While a native speaker can immediately identify an ungrammatical string of adjectives as ungrammatical, they can still readily interpret its intended meaning. Such meaning-preserving violations of the syntactic norms of language might thus be seen as an acceptable substitute in a game where players are primarily concerned with efficiently communicating the semantic meaning, not the syntactic form, of their actuations. We might expect this to contrast to kinds of data where efficient or instrumental responding is not merely less appropriate but is rather incorrect. Were players averse to giving false answers in the game, we might expect accuracy to be preserved, despite the game context. Alternatively, the trade-off might work in the opposite direction: accuracy may be preserved only at cost to enjoyment. Future research should attempt to replicate this trade-off with a diversity of types of game collecting a diversity of types of data.

Finally, I used only a single game for my experiments. The demonstrated generalis-

ability of the Intrinsic Elicitation design approach and of the experimental results is thus limited. More convincing evidence of this would require applying the approach in designing a diversity of games for collecting different kinds of data. However, the use of a single case study allowed consistency of comparison between my experiments. In particular, I was able to compare effect sizes for the decrease in data quality in different game configurations. At for the game itself, I like to believe it captures the essence of both a valid experimental task and a genuinely interesting puzzle game.

8.2.2 Limitations of the Experiments

I articulated in chapter 2 that players and gameplay situations are diverse, yet I make sweeping assumptions about them in the Rational Game User Model. Do players really enjoy maximising performance at the game? Are their actions really rational? This rational perspective affects my treatment of games, gameplay, and players. In my qualitative work, I am perhaps not going to identify potential games as games. I may have constructed distinctions among games because of the priority of this, which may not be justified. Empirically, I performed my experiments on a sample of participants from Prolific, yet online play in a paid experimental context does not necessarily generalise to all play.

8.3 Future Work

It has been the attempt of this thesis to establish a starting point for designing games for human-subject data collection. Much work remains to be done, as I summarise briefly below. I begin with directions to further develop Intrinsic Elicitation and the Rational Game User model. I then discuss new directions in considering motivation in games. Finally I point towards future work that is required in understanding validity in games.

8.3.1 Intrinsic Elicitation

The first necessary next step for the Rational Game User Model is to further test its predictions. Here, I consider the suggestions of Siu et al. (2017) instructive: designing a series of studies and replications where we define and hold player experience and task completion metrics constant while varying individual factors of the model in controlled A/B design experiments, with data collection goals where the ground truth can be cross-validated against preexisting data – for instance prior responses of participants to standard

personality instruments or replicating well-established spontaneous inclinations.

A second necessary step is testing the usefulness and ease of use of the design approach of Intrinsic Elicitation with data collection designers. The studies reported here suggested that one area where the model is limited at present is framing. This is something for which the Rational Game User model is at present too abstract to give clear guidance. While in principle we can explain framing effects as arising from social norms as an external source of utility, this cannot easily be translated into specific instructions or requirements for design.

The Rational Game User model is open for extension with additional variables such as curiosity and social norms. It may be that incorporating these elements to the model's calculation of utility can aid the development of applied games for data collection for which pure 'game-based' rewards such as points are unsuitable. It may be that players will provide better data to a game if they risk social censure, for example.

Finally, if the model and approach prove reliable and useful for human-subject data collection, a third step would be to test its generalisability for data collection games like human computation games as well.

8.3.2 Motivation in Games and Experiments

I have demonstrated the usefulness of adopting a rational model of motivation. Such rational models require an understanding of motivation in games at a fine grained, moment-to-moment level, and for motivation for particular actions within games. Studies of motivation in both games and experiments have looked primarily at generalised motivation – e.g. motivation to participate in citizen science, or overall enjoyment of an activity – and not the motivation either at specific moments of play, or motivation to perform specific behaviours at moments in time. Yet such moment-to-moment motivations has been argued here to affect the quantity and validity of data elicited. While it is clear that in games we cannot take continued motivation to act in the way we want on trust, this generalises naturally to non-game experiments. After all, experiments are just games that are not (usually) very fun.

Further, given the fuzzy boundary between games and experiments it makes more sense than ever to pursue a research program on the mapping between motivation for behaviours in games and in real life. When does this mapping (not) occur? Such knowledge would be necessary for elicitation games to identify which game mechanics encode what data about

the player. What are the relevant differences between a game and experimental context for this? I have suggested that the differing social norms of game and experiment are one such difference.

As a game provides an incentive structure – which we can capture with idealised abstractions such as the Rational Game User Model – it may be particularly informative to pursue a research program on when and how players violate rational behaviour in a game, as it will be these occasions when some other bias must be affecting their responses. For example, with *Adjective Game*, I observed that, despite no incentive to do so, players were more likely to provide one adjective order over any other, this corresponding to their intuitions of adjective order.

8.3.3 Validity in Games

Finally, I have explored in this thesis how the use of games for data collection gives rise to threats to validity. Such threats might understandably deter researchers tempted to use games who are concerned about justifying their methodological decisions, especially to audiences unfamiliar with game-based methodologies. Whether for direct data collection, as with elicitation games, or as a more enjoyable way of presenting an experimental stimulus, games have the potential to broaden participation in research and make it more enjoyable. As such it would be a shame if anxieties about validity stopped us from exploring the potential of games in a diversity of research fields.

To achieve this, we need more work on how games can be used in research when validity is a concern. While this thesis has taken some steps on this road by surveying threats to validity, and by constructing a theoretical model, I have, like most authors on the topic, for the most part speculated on *possible* threats to validity that seem *plausible*, without first empirically quantifying their real impact. Such speculations should not be used to imprison ourselves and others in familiar, nonthreatening paradigms. Indeed, if we would be able to run more studies with more participants by doing so, would it not be rational to give up a little essential validity to purchase a little more temporary enjoyment?

Better a diamond with a flaw than a pebble without.

Confucius (attrib.)

Appendix A

Experimental Materials

A.1 Adjective Game

Figures A.1 & A.2 Game interfaces from the first and subsequent studies during play, shown upon completing an input. Players assemble a three-word string by tapping/clicking word bubbles at the bottom of the screen. Shapes identified by this string are then cleared from the board above, with bonuses for contiguous groups of shapes. This triggers multiple forms of feedback: a bar filling and turning green when completing a word string; matching shapes are highlighted; the score at the top is increasing; the number of moves left is increasing (A.1) and decreasing (A.2) depending on the size of groups cleared; a bar around the edge of the screen fills corresponding to progression through the level; and stars have burst out of the cleared group. Highlighted shapes will disappear and the remaining shapes will drop down into empty spaces, with new shapes falling from above.

Figures A.3 – A.5 Screenshots from levels 1, 2, and 3 of the *Adjective Game* tutorial. The tutorial consists of three levels. Each level presents multiple pieces of text in response to player actions. The third and final level of the tutorial adds additional input buttons as the player progresses through solving the puzzle.

Figure A.6 Screenshot from the on-screen help. This modal dialog appears upon clicking the juicy question mark in the top right corner of the screen. This was included in Studies 2, 3 and 4.

A.2 Studies 1 & 2

Figures A.7 & A.8 Experimental control interfaces from the first (A.7) and second (A.8) study. A target shape, always the top left, is indicated by a double red box in a 2x2 grid of shapes. Each other shape differs from the first in one dimension: shape, size, colour, or filling. Participants again tap/click on word bubbles in the bottom screen to assemble a word string identifying the target shape. In Study 1 (A.7), word bubbles are arranged in a grid, in Study 2 (A.8), word bubbles are arranged in columns by type.

Figures A.9 & A.10 The first and second page of the integrated participant instructions in Studies 1 and 2. This is the first thing participants see once they have agreed to begin the study on Prolific.

Figure A.11 Demographics questionnaire in Study 1 and 2. This follows after the instructions in Figure A.10.

A.3 Study 3

Figures A.12 & A.13 Participant information from the game-framed and experiment-framed conditions. This was included as a requirement for ethical approval, and also captures informed consent. The text in both is identical to that presented on Prolific to recruit participants. The only scope available to adapt this screen was its visual style.

Figure A.14 Initial instruction screen for game-framed condition only. This followed the participant information screen and preceded the 'loading' screen.

Figures A.15 & A.16 Adapted loading screens from the game and experiment-framed conditions. Each enforced a wait of 10 seconds. The game-framed loading screen shows an animated *Adjective Game* board being played automatically. The experiment-framed loading screen shows a number counting down from 10 in the bottom right hand corner. This is replicated by the "Next" button on completion. The game-framed condition presents a button labelled "start" in the bottom centre of the screen, which matches the visual style of the game inputs.

A.4 Study 4

Figures A.17 & A.18 Screenshots from the with- and without-instruction conditions in Study 4. This modal dialogue box appears immediately before beginning the game. Players must click the checkbox for the button labelled "Next" to appear, allowing them to continue.

Figure A.19 Screenshot from the with-instruction condition of *Adjective Game* in Study 4 showing the in-game instruction presented to participants. The same instruction is present on every level. This instruction fades in at the beginning of each level to draw attention to it in keeping with the visual style of the game.

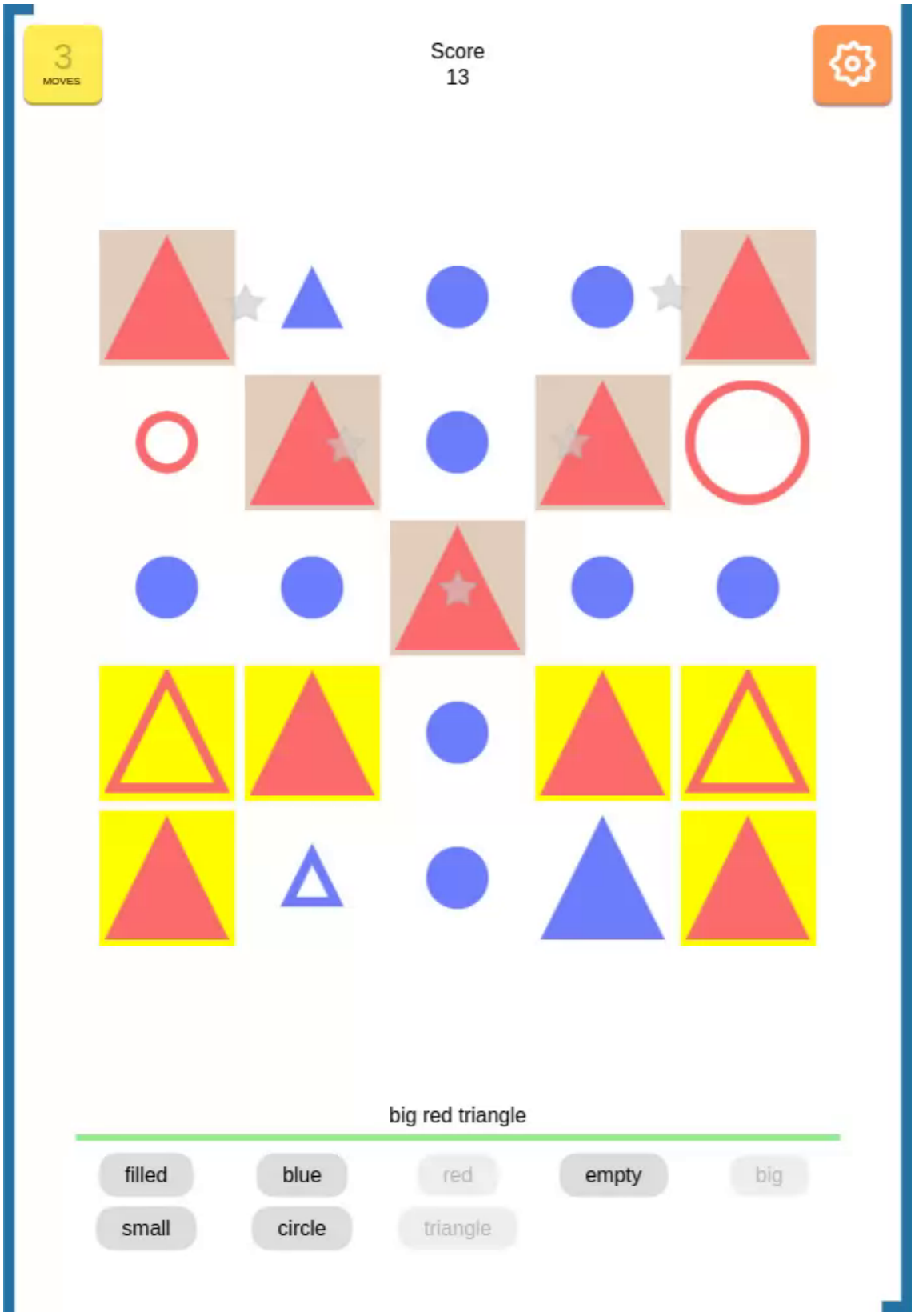


Figure A.1: Study 1: Game condition

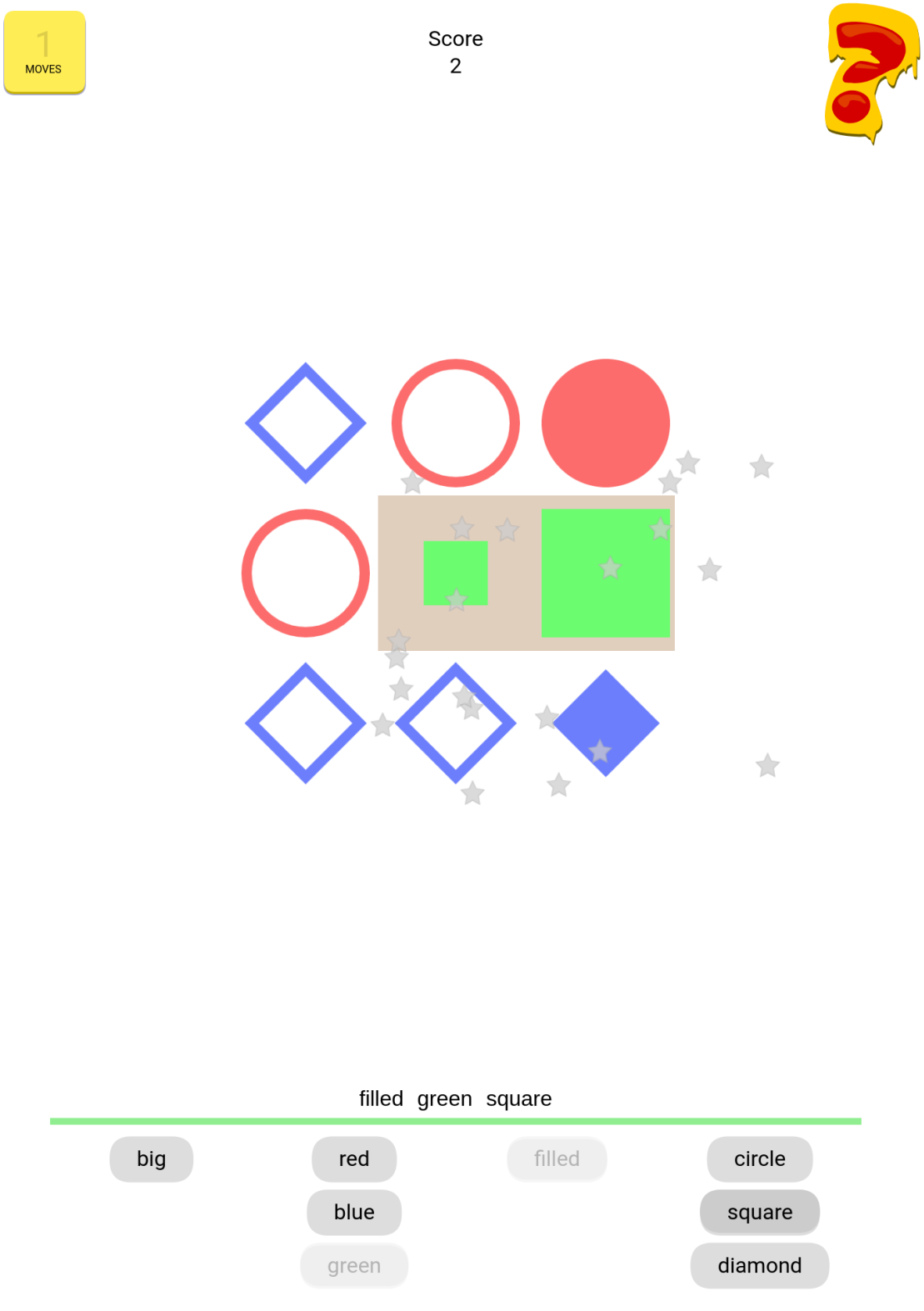


Figure A.2: Study 2-4: Game condition

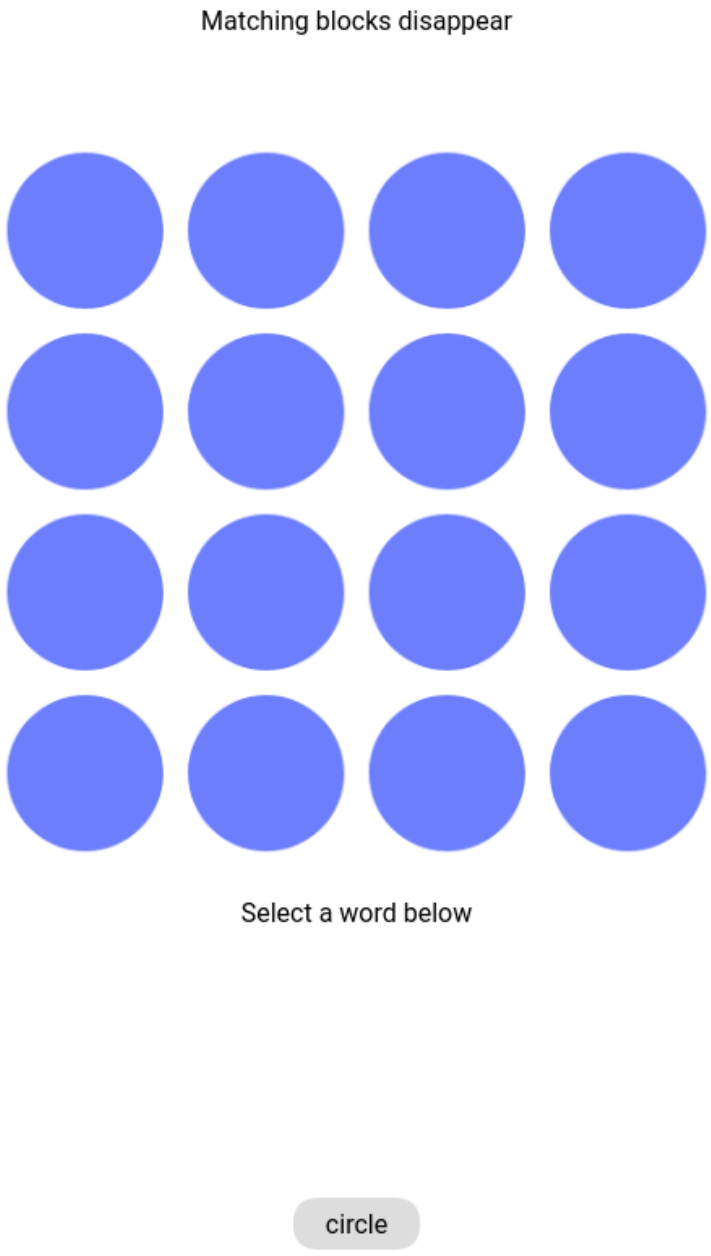


Figure A.3: Tutorial level 1

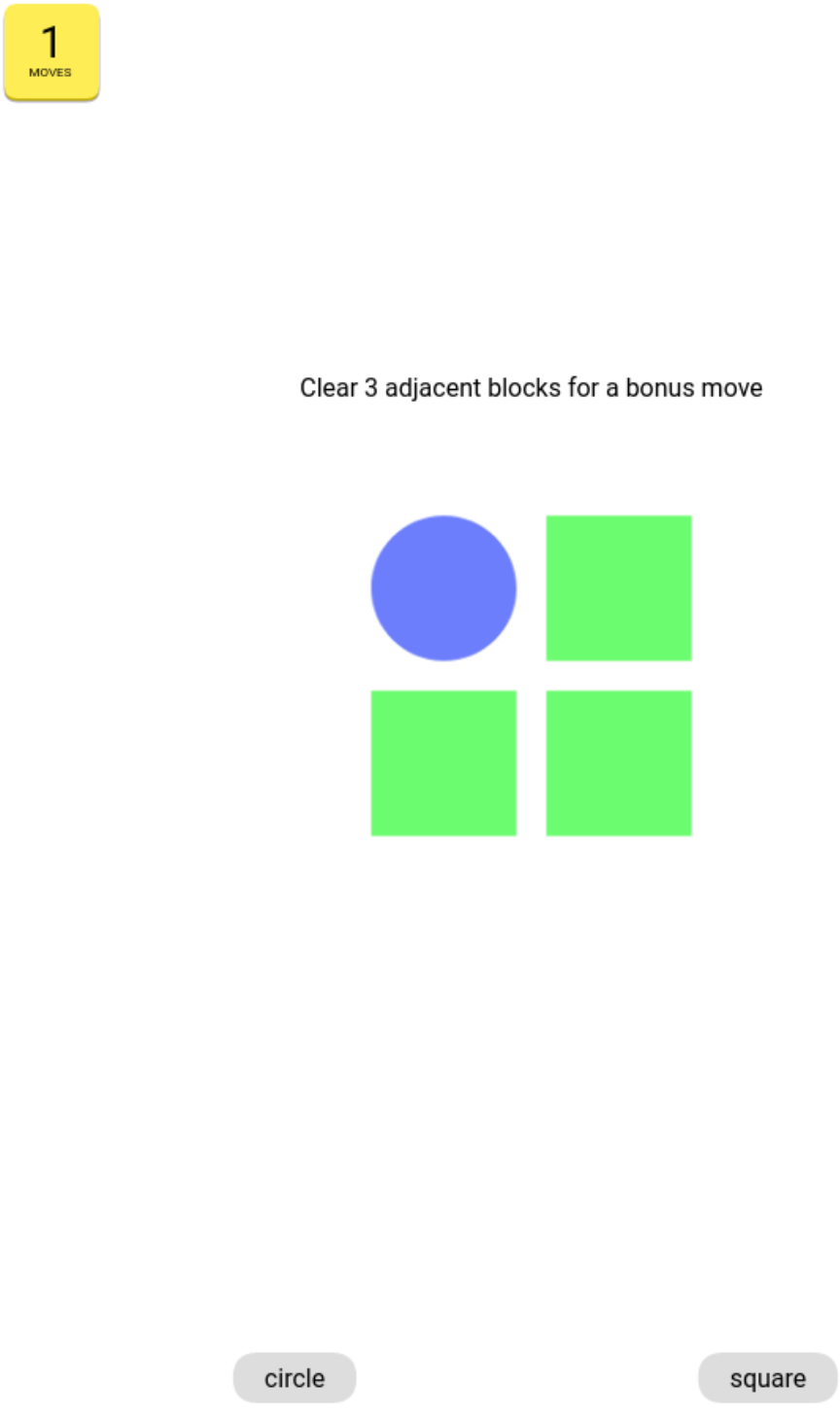


Figure A.4: Tutorial level 2

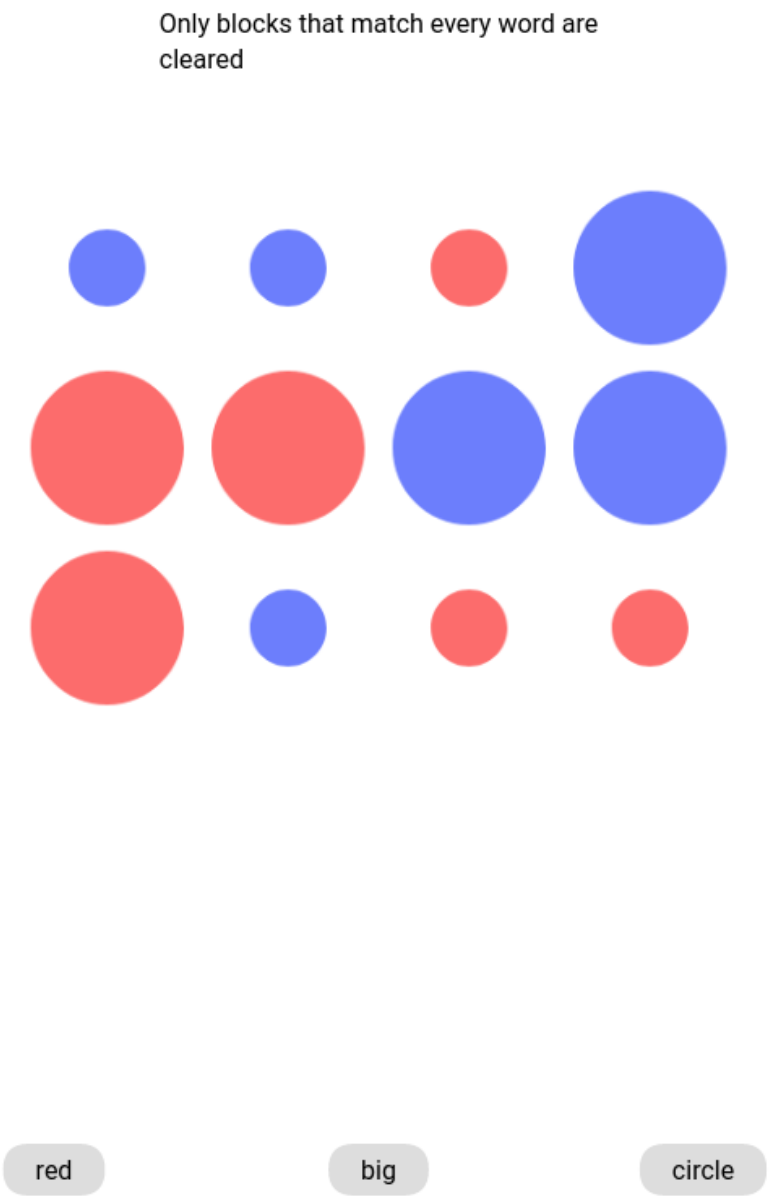


Figure A.5: Tutorial level 3

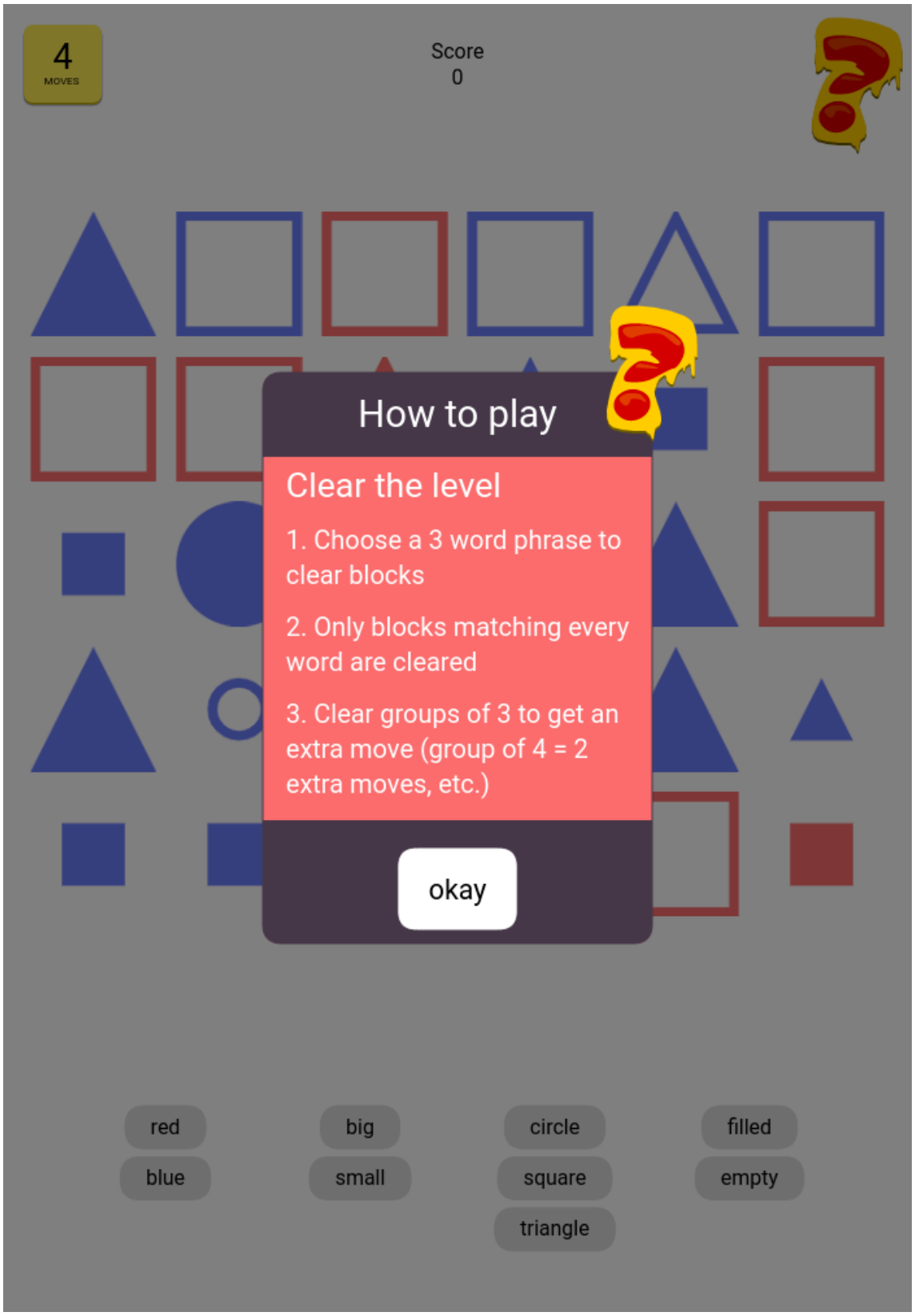


Figure A.6: On-screen help dialog in Studies 2, 3 and 4

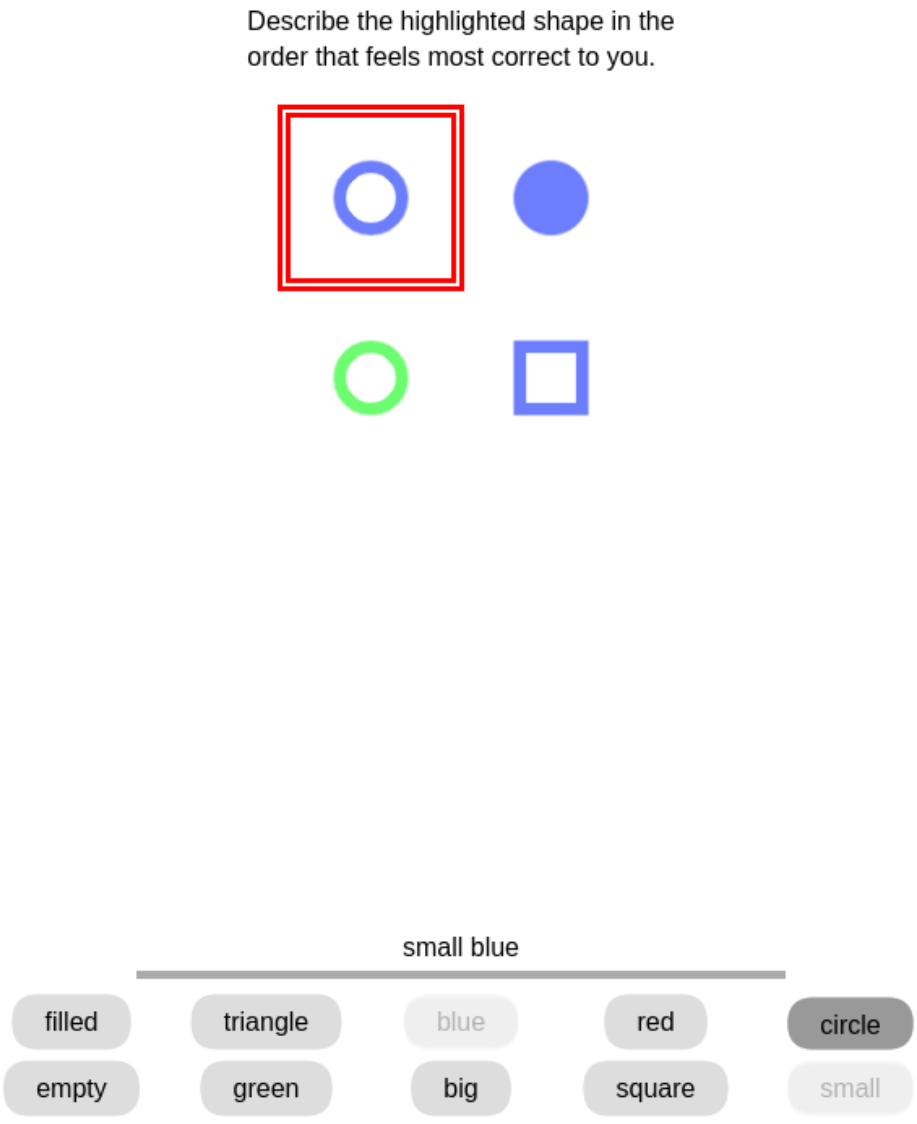


Figure A.7: Study 1: Control condition

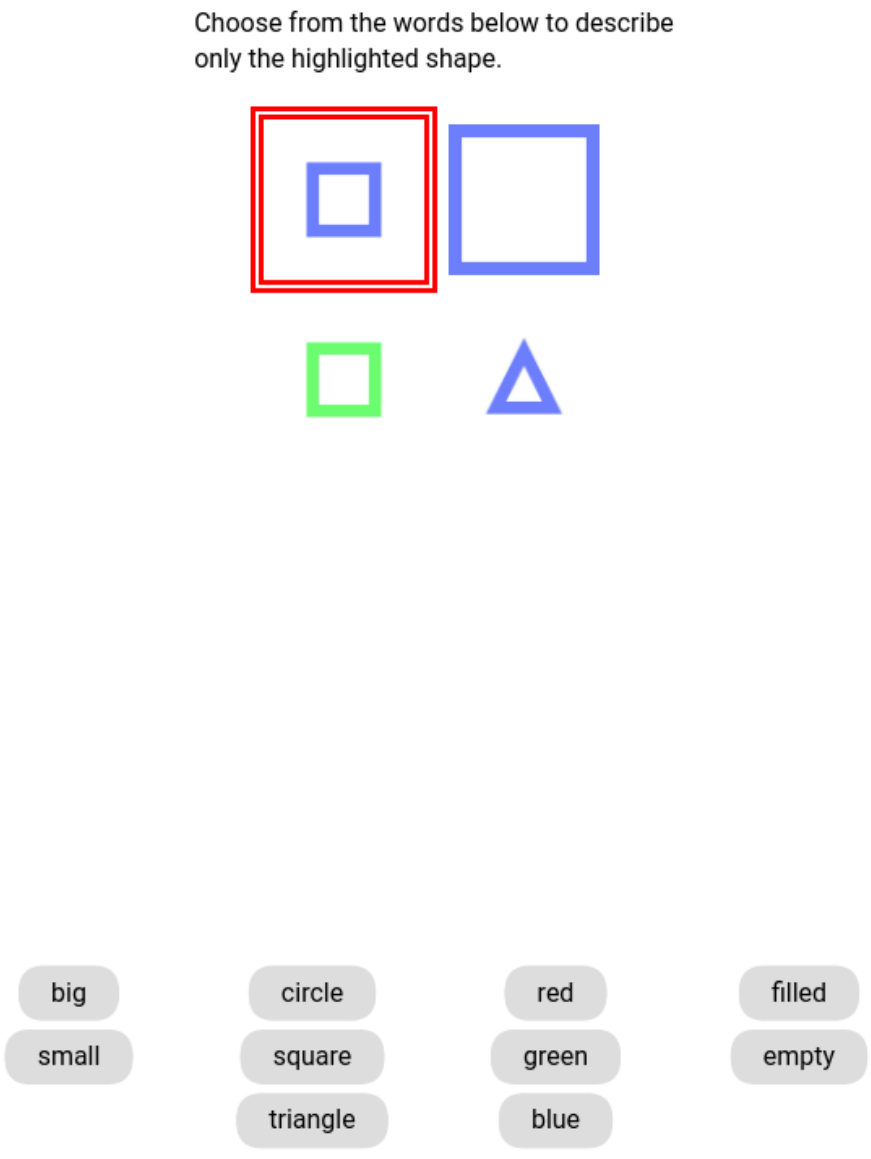


Figure A.8: Study 2: Control condition

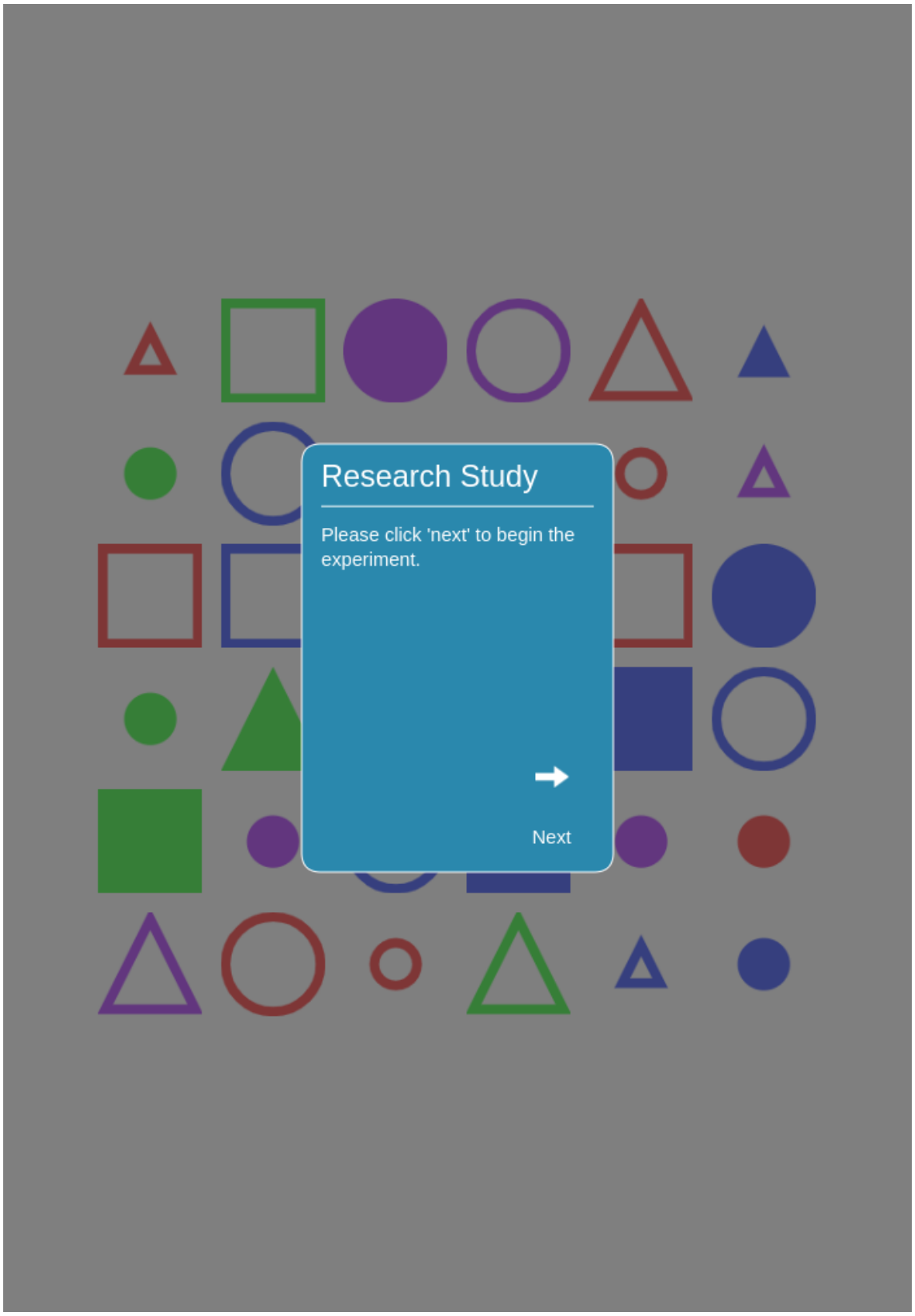


Figure A.9: Study 1 intro

Figure A.10: Study 2 intro

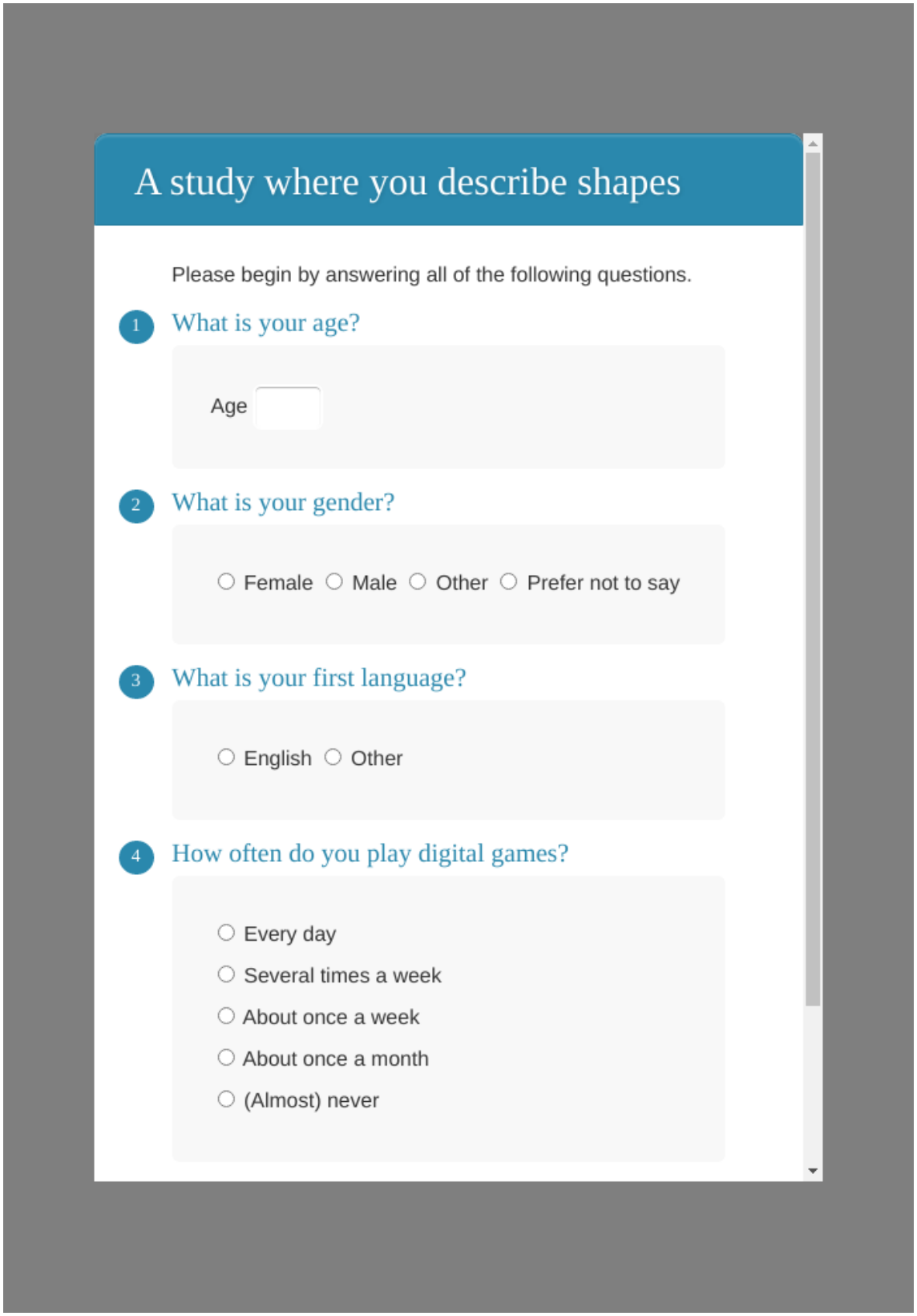


Figure A.11: Questionnaire in studies 1 and 2

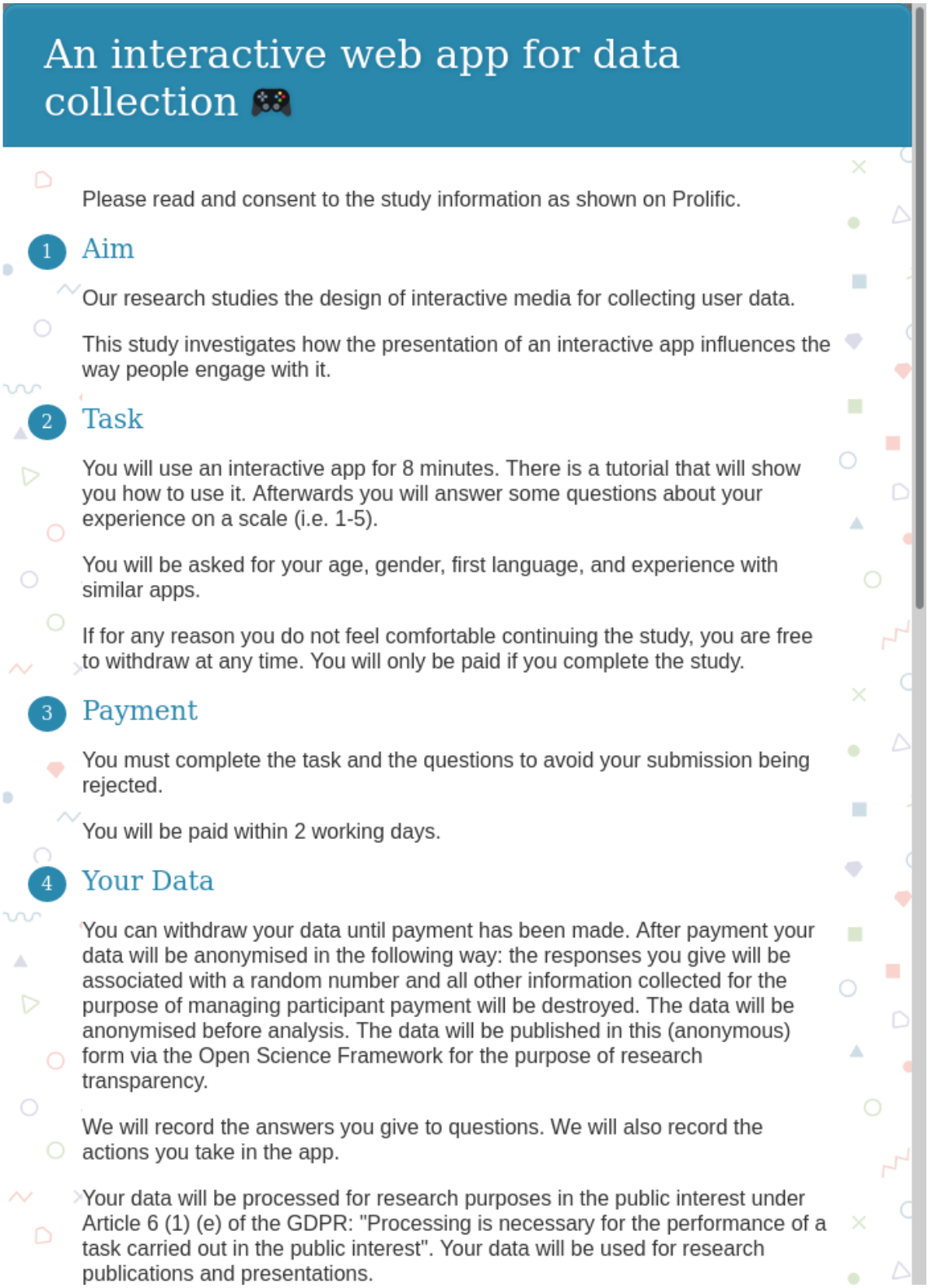


Figure A.12: Participant information from the game-framed condition

An interactive web app for data collection

Please read and consent to the study information as shown on Prolific.

1 Aim

Our research studies the design of interactive media for collecting user data.

This study investigates how the presentation of an interactive app influences the way people engage with it.

2 Task

You will use an interactive app for 8 minutes. There is a tutorial that will show you how to use it. Afterwards you will answer some questions about your experience on a scale (i.e. 1-5).

You will be asked for your age, gender, first language, and experience with similar apps.

If for any reason you do not feel comfortable continuing the study, you are free to withdraw at any time. You will only be paid if you complete the study.

3 Payment

You must complete the task and the questions to avoid your submission being rejected.

You will be paid within 2 working days.

4 Your Data

You can withdraw your data until payment has been made. After payment your data will be anonymised in the following way: the responses you give will be associated with a random number and all other information collected for the purpose of managing participant payment will be destroyed. The data will be anonymised before analysis. The data will be published in this (anonymous) form via the Open Science Framework for the purpose of research transparency.

We will record the answers you give to questions. We will also record the actions you take in the app.

Your data will be processed for research purposes in the public interest under Article 6 (1) (e) of the GDPR: "Processing is necessary for the performance of a task carried out in the public interest". Your data will be used for research publications and presentations.

All data you provide will be stored and processed securely in the UK and the European Economic Area in full compliance with data protection legislation.

5 Contact Details

Figure A.13: Participant information from the experiment-framed condition

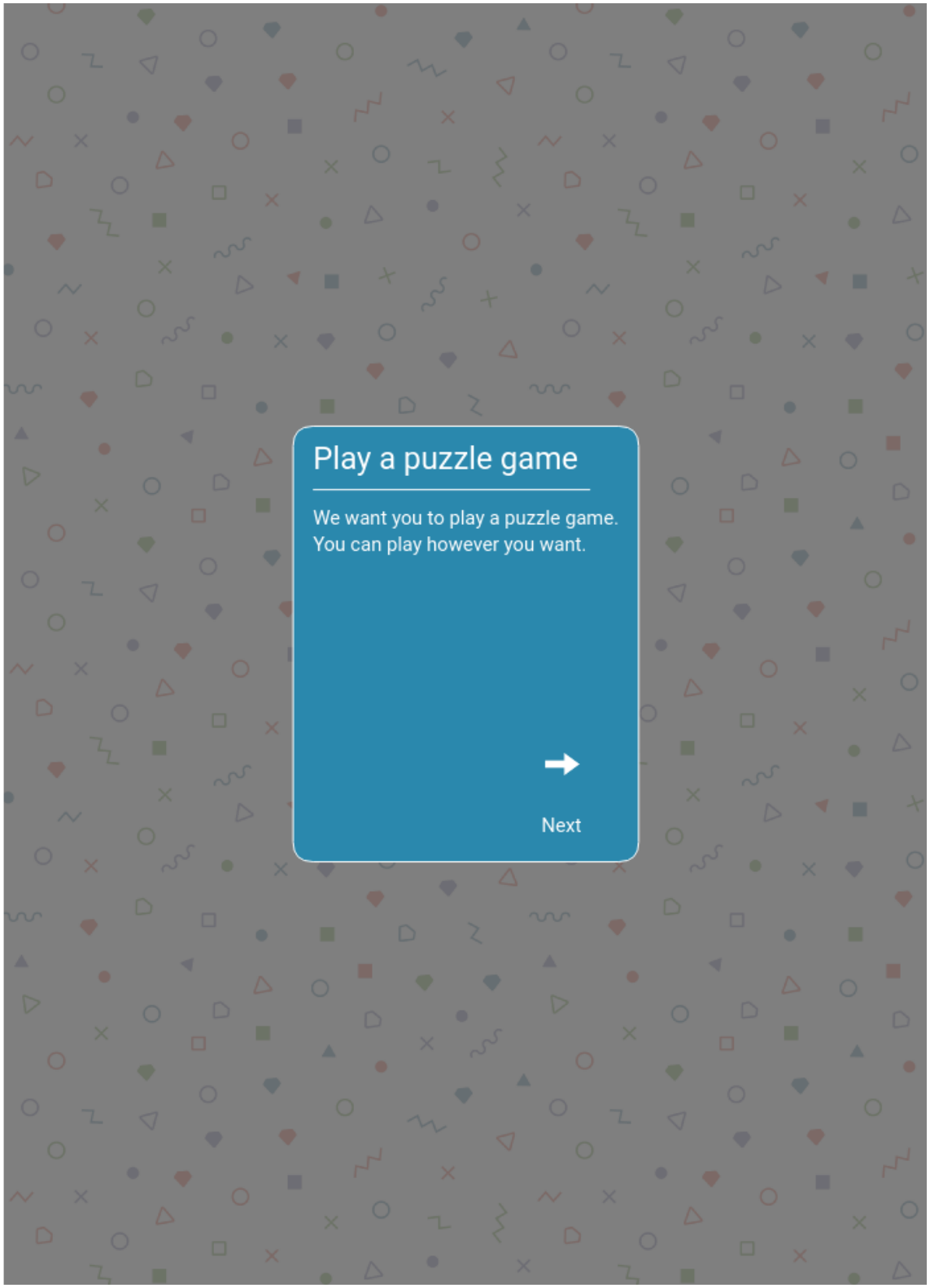


Figure A.14: Participant information from the game-framed condition

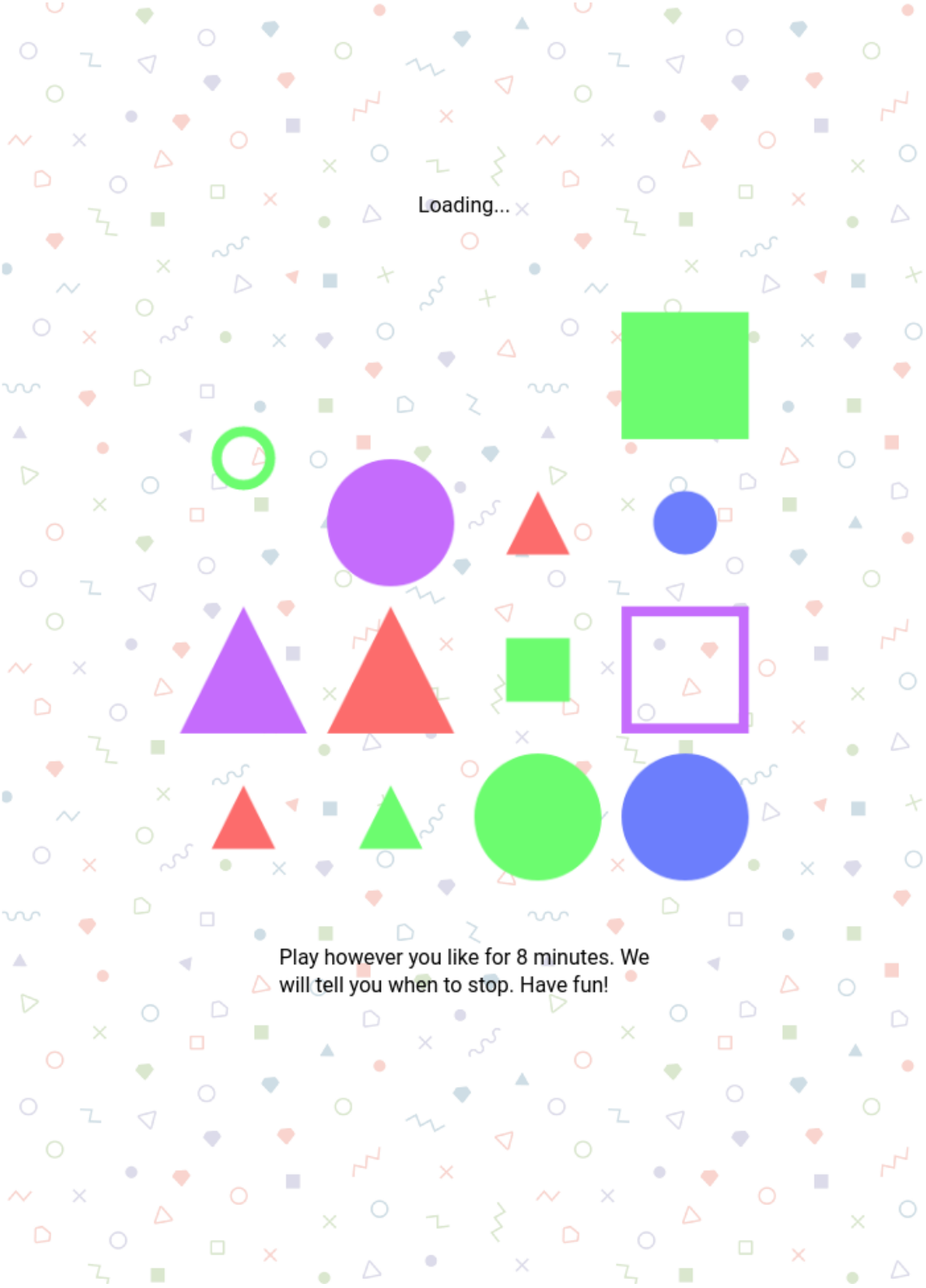


Figure A.15: Loading screen from the game-framed condition

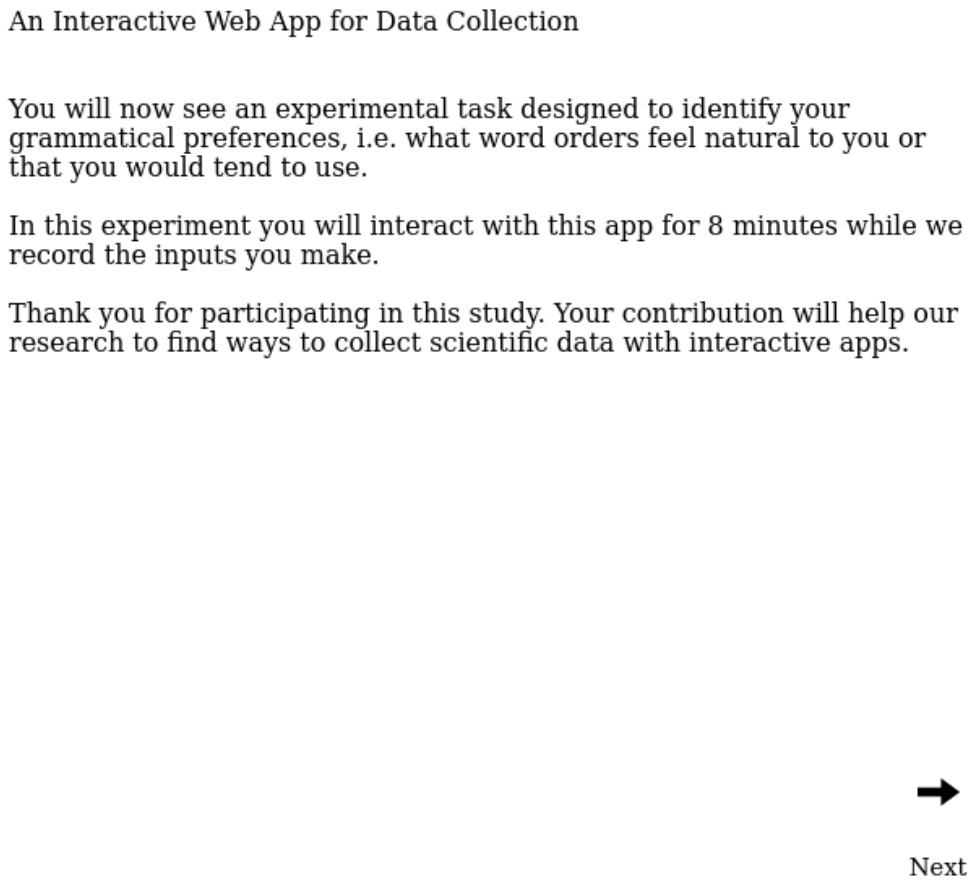


Figure A.16: Loading screen from the experiment-framed condition

Figure A.17: Starting instruction in the with-instruction condition

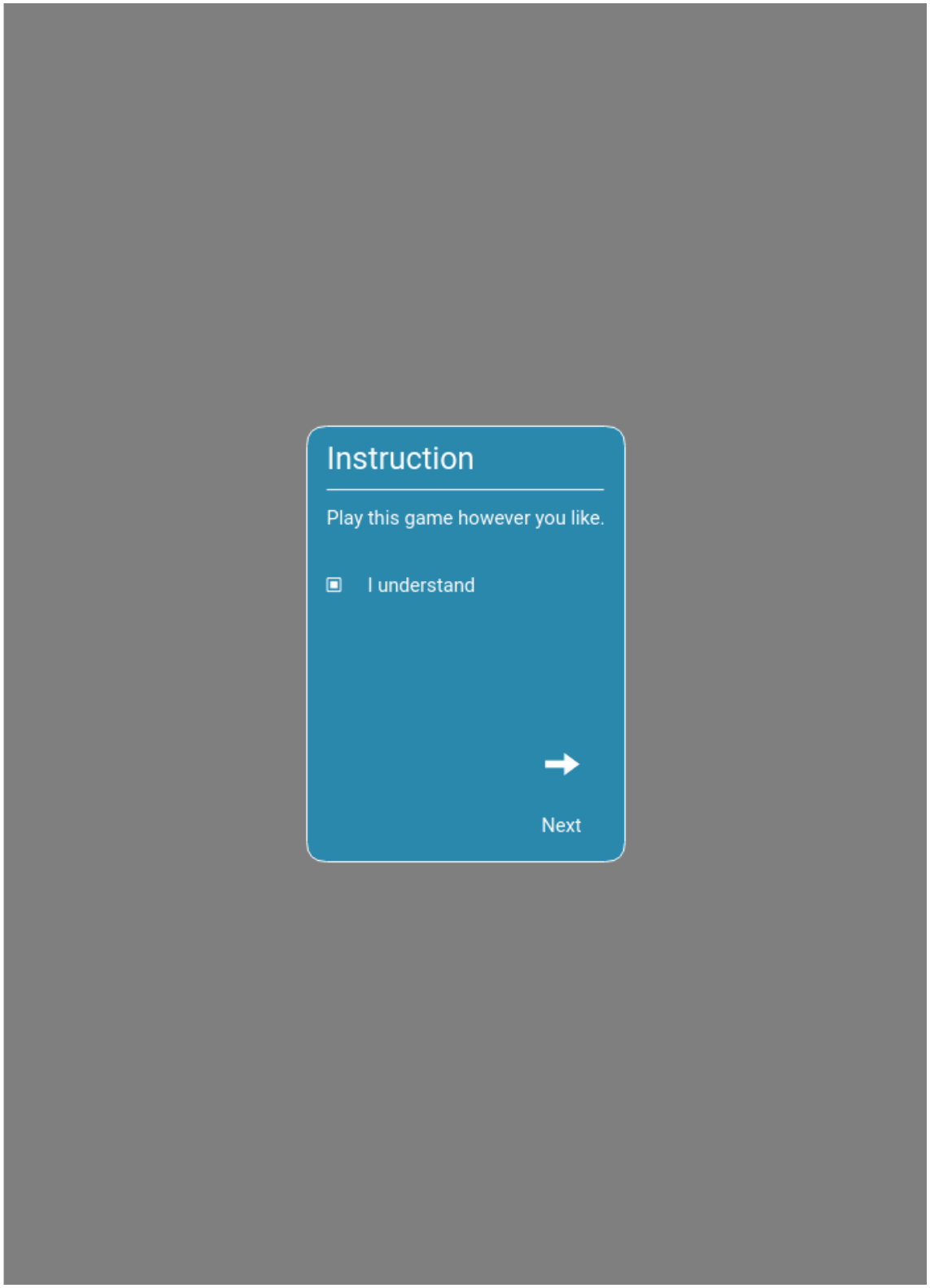


Figure A.18: Starting instruction in the without-instruction condition

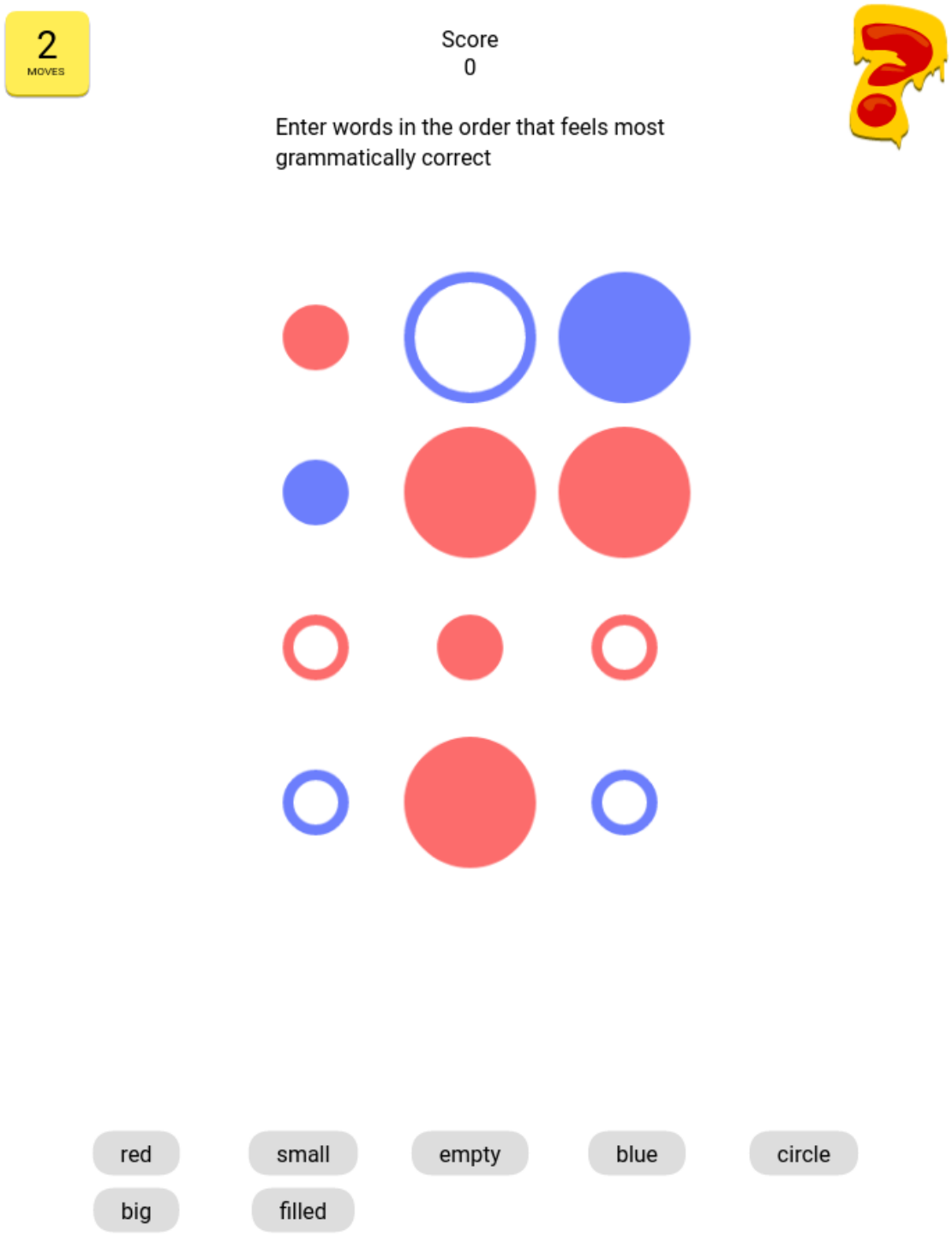


Figure A.19: On-screen instruction in the with-instruction condition

References

Adachi, P. J., & Willoughby, T. (2011a). The effect of video game competition and violence on aggressive behavior: Which characteristic has the greatest influence? *Psychology of Violence*, *1*(4), 259–274. https://doi.org/10.1037/a0024908

Adachi, P. J., & Willoughby, T. (2011b). The effect of violent video games on aggression: Is it more than just the violence? *Aggression and Violent Behavior*, *16*(1), 55–62. https://doi.org/10.1016/j.avb.2010.12.002

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, *47*(4), 823–837.

Al Hashimi, S. (2007). Preferences and patterns of paralinguistic voice input to interactive media. *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments: 12th International Conference, HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings, Part III 12*, 3–12.

Alexander, W., & Tait, R. (2002). *Cranium hoopla* [Board game]. Cranium Inc.

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J., & Kievit, R. A. (2021). Raincloud plots: A multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. *Wellcome Open Res*, (63). https://doi.org/10.12688/wellcomeopenres.15191.2

Allison, F. (2020). *Voice interaction game design and gameplay* (Doctoral dissertation). The University of Melbourne.

Allison, F., Carter, M., & Gibbs, M. (2020). Word play: A history of voice interaction in digital games. *Games and Culture*, *15*(2), 91–113.

Allison, F., Carter, M., Gibbs, M., & Smith, W. (2018). Design patterns for voice interaction in games. *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 5–17.

Allison, F., Newn, J., Smith, W., Carter, M., & Gibbs, M. (2019). Frame analysis of voice interaction gameplay. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Almaniac [Board game]. (1990). ALMANiAC, Inc.

Alzheimer's Research UK. (n.d.). *Sea hero quest* [Accessed: 2022-02-11]. https://www.alzheimersresearchuk.org/research/for-researchers/resources-and-information/sea-hero-quest/

Ambridge, B., & Rowland, C. F. (2013). Experimental methods in studying child language acquisition. *WIREs Cognitive Science*, *4*(2), 149–168. https://doi.org/https://doi.org/10.1002/wcs.1215

Anand, P., Chung, S., & Wagers, M. (2011). Widening the net: Challenges for gathering linguistic data in the digital age [Response to NSF SBE 2020: Future Research in the Social, Behavioral & Economic Sciences]. https://people.ucsc.edu/~schung/anandchungwagers.pdf

Anderson, C. A., & Bushman, B. J. (2001). Effects of Violent Video Games on Aggressive Behavior, Aggressive Cognition, Aggressive Affect, Physiological Arousal, and Prosocial Behavior: A Meta-Analytic Review of the Scientific Literature. *Psychological Science*, *12*(5), 353–359. https://doi.org/10.1111/1467-9280.00366

Anderson, C. A., & Dill, K. E. (2000). Video Games and Aggressive Thoughts, Feelings, and Behavior in the Laboratory and in Life. *Journal of Personality and Social Psychology*, *78*(4), 772–790.

Anderson, M. (2014). *Black rabbit dice* [Board game].

Arechar, A. A., & Rand, D. G. (2021). Turking in the time of covid. *Behavior Research Methods*, *53*, 2591–2595.

Arnold, H. J. (1976). Effects of performance feedback and extrinsic reward upon high intrinsic motivation. *Organizational Behavior and Human Performance*, *17*(2), 275–288.

Aronson, E. (1985). Cognitive evaluation theory: Perceived causality and perceived competence. *Intrinsic motivation and self-determination in human behavior*, 43–51.

Auyang, S. Y. (1999). *Foundations of Complex-System Theories in Economics, Evolutionary Biology, and Statistical Physics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511626135

Bache, C., & Davidsen-Nielsen, N. (2010). *Mastering english: An advanced grammar for non-native and native speakers* (Vol. 22). Walter de Gruyter.

Baker, R. S. (2005). *Designing intelligent tutors that adapt to when students game the system* (Doctoral dissertation). Carnegie Mellon University Pittsburgh.

Baker, R. S. (2011). Gaming the System: A Retrospective Look. *Philippine Computing Journal*, *6*(2), 9–13.

Barthel, K. (2013). Perceived color difference-ein spielerisches experiment zur erfassung empfundener farbunterschiede. *19. Workshop Farbbildverarbeitung, Berlin*.

Bartle, R. (1996). Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD research*, *1*(1), 19.

Bartle, R. (2004). *Designing Virtual Worlds*. New Riders.

Bartle, R. (2010). A "digital culture, play and identity: A world of warcraft reader" reader. *Game Studies*, *10*(1), 17.

Baruch, A., May, A., & Yu, D. (2016). The motivations, enablers and barriers for voluntary participation in an online crowdsourcing platform. *Computers in Human Behavior*, *64*, 923–931. https://doi.org/10.1016/j.chb.2016.07.039

Baum, S. R. (2002). Sensitivity to sub-syllabic constituents in brain-damaged patients: Evidence from word games. *Brain and language*, *83*(2), 237–248.

Baum, W. M. (2002). From molecular to molar: a paradigm shift in behavior analysis. *Journal of the experimental analysis of behavior*, *78*(1), 95–116. https://doi.org/10.1901/jeab.2002.78-95

Bauza, A. (2010). *Hanabi* [Card game]. R&R Games.

Bell, M. L., Kenward, M. G., Fairclough, D. L., & Horton, N. J. (2013). Differential dropout and bias in randomised controlled trials: When it matters and when it may not. *Bmj*, *346*.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of Serious Games: An Overview. *Advances in Human-Computer Interaction*, *2013*, 1–11.

Berger, A., Jones, L., Rothbart, M. K., & Posner, M. I. (2000). Computerized games to study the development of attention in childhood. *Behavior Research Methods, Instruments, & Computers*, *32*(2), 297–303.

Bergström, K. (2010). The implicit rules of board games: On the particulars of the lusory agreement. *Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments*, 86–93.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political analysis*, *20*(3), 351–368.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American psychologist*, *37*(3), 245.

Berkowitz, L., & Troccoli, B. T. (1986). An examination of the assumptions in the demand characteristics thesis: With special reference to the velten mood induction procedure. *Motivation and Emotion*, *10*(4), 337–349.

Bevan, G., & Hood, C. (2006). What's measured is what matters: Targets and gaming the system in the english public health care system. *Public Administration*, *84*(3), 517–538. https://doi.org/10.1111/j.1467-9299.2006.00600.x

Bidwill, M. (2015). There came an echo – don't yell. [Accessed 2022-02-11]. *TechRaptor*. https://techraptor.net/gaming/review/there-came-echo-review-dont-yell

Bizzocchi, J., & Tanenbaum, J. (2011). Well read: Applying close reading techniques to gameplay experiences. *Well played 3.0: Video games, value and meaning*, *3*.

Björk, S., & Holopainen, J. (2006). Games and design patterns. *The game design reader*, 410–437.

Blizzard Entertainment. (2016). *Overwatch* [Video game].

Bogost, I. (2007). *Persuasive Games: The Expressive Power of Videogames*. MIT Press.

Bostan, B. (2009). Player motivations: A psychological perspective. *Computers in Entertainment (CIE)*, *7*(2), 1–26.

Boyle, E. A., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, *28*(3), 771–780. https://doi.org/10.1016/j.chb.2011.11.020

Broome, J. (1991). "Utility". *Economics and Philosophy*, *7*, 1–12.

Brown, E., & Cairns, P. (2004). A Grounded Investigation of Game Immersion. *CHI '04 Extended Abstracts on Human Factors in Computing*, 1297–1300. https://doi.org/10.1145/985921.986048

Brown, H. R., Zeidman, P., Smittenaar, P., Adams, R. A., McNab, F., Rutledge, R. B., & Dolan, R. J. (2014). Crowdsourcing for cognitive science – The utility of smartphones. *PloS one*, *9*(7), e100662.

Browning, R. C., Baker, E. A., Herron, J. A., & Kram, R. (2006). Effects of obesity and sex on the energetic cost and preferred speed of walking. *Journal of Applied Physiology*, *100*(2), 390–398. https://doi.org/10.1152/japplphysiol.00767.2005

Buikstra, B., Kortmann, R., Klaassen, E., Rook, L., & Verbraeck, A. (2015). In Pursuit of Evidence: a Design and Empirical Study of a Gamified Online Marketplace. *AcademicMindTrek"15*, 73–80.

Cairns, P., Cox, A., Berthouze, N., Dhoparee, S., & Jennett, C. (2006). Quantifying the experience of immersion in games. *cognitive science of games and gameplay workshop at cognitive science*, 26–29.

Camerer, C. F., Hogarth, R. M., Budescu, D. V., & Eckel, C. (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty*, *19*(1-3), 7–42. https://doi.org/10.1023/A:1007850605129

Camerer, C. F., & Loewenstein, G. (2004). Behavioral Economics: Past, Present, Future. In *Advances in behavioral economics* (pp. 2–51). Princeton University Press.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

Capcom. (2008). *Apolo justice: Ace attorney* (Version Nintendo DS) [Video game].

Capcom Vancouver. (2013). *Dead rising 3* (Version XBox One) [Video game]. Microsoft Studios.

Carnagey, N. L., & Anderson, C. A. (2005). The effects of reward and punishment in violent video games on aggressive affect, cognition, and behavior. *Psychological Science*, *16*(11), 882–9. https://doi.org/10.1111/j.1467-9280.2005.01632.x

Carter, M., Allison, F., Downs, J., & Gibbs, M. (2015). Player Identity Dissonance and Voice Interaction in Games. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 265–269. http://dl.acm.org/citation.cfm?id=2793144

Case, D. A., Ploog, B. O., & Fantino, E. (1990). Observing behavior in a computer game. *Journal of the Experimental Analysis of Behavior*, *54*(3), 185–199.

Celino, I., & Cerizza, D. (2012). Urbanopoly: collection and quality assesment of geo-spatial linked data via a human computation game. *Proceedings of the 10th Semantic Web Challenge*, 148–163.

Celino, I., Cerizza, D., Contessa, S., Corubolo, M., Dell'Aglio, D., Valle, E. D., & Fumeo, S. (2012). Urbanopoly–A Social and Location-Based Game with a Purpose to Crowdsource Your Urban Data. *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 910–913.

Champely, S. (2020). Pwr: Basic functions for power analysis [Version 1.3-0]. https://CRAN.R-project.org/package=pwr

Chandler, J., Mueller, P., & Paolacci, G. (2012). Non-naivety among experimental participants on amazon mechanical turk. *ACR North American Advances*.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological science*, *26*(7), 1131–1139.

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM*, *50*(4), 31–34.

Chmielewski, M., & Kucker, S. C. (2020). An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473.

Chomsky, N., Gallego, Á. J., & Ott, D. (2019). Generative grammar and the faculty of language: Insights, questions, and challenges. *Catalan Journal of Linguistics*, *2019*, 229–261.

Cinque, G. (2002). *Functional structure in DP and IP* (Vol. 1). Oxford University Press.

Cinque, G. (2010). *The syntax of adjectives: A comparative study*. The MIT Press.

Cler, G. J., Mittelman, T., Braden, M. N., Woodnorth, G. H., & Stepp, C. E. (2017). Video game rehabilitation of velopharyngeal dysfunction: A case series. *Journal of Speech, Language, and Hearing Research*, *60*(6S), 1800–1809.

Cole, T., & Gillies, M. (2022). More than a bit of coding:(un-) grounded (non-) theory in HCI. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–11.

Consalvo, M. (2007). *Cheating: Gaining Advantage in Videogames*. MIT Press.

Conway, S., & Trevillian, A. (2015). "blackout!" unpacking the black box of the game event. *Transactions of the Digital Games Research Association*, *2*(1).

Cooper, S. (2014). *A framework for scientific discovery through video games*. Morgan & Claypool.

Cooper, S. (2015). Massively Multiplayer Research: Gamification and (Citizen) Science. In S. P. Walz & S. Deterding (Eds.), *The gameful world: Approaches, issues, applications* (pp. 487–500). MIT Press.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., & Players, F. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*(7307), 756–60. https://doi.org/10.1038/nature09304

Cooper, S., Treuille, A., Barbero, J., Leaver-Fay, A., Tuite, K., Khatib, F., Snyder, A. C., Beenen, M., Salesin, D., Baker, D., et al. (2010). The challenge of designing scientific discovery games. *Proceedings of the Fifth international Conference on the Foundations of Digital Games*, 40–47.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research* [3rd Edition]. Sage publications.

Costikyan, G. (2013). *Uncertainty in games*. MIT Press.

Coutrot, A., Schmidt, S., Coutrot, L., Pittman, J., Hong, L., Wiener, J. M., Hölscher, C., Dalton, R. C., Hornberger, M., & Spiers, H. J. (2019). Virtual navigation tested on a mobile app is predictive of real-world wayfinding navigation performance. *PloS one*, *14*(3), e0213272.

Crawford, V. P. (2002). Introduction to experimental game theory. *Journal of Economic Theory*, *104*(1), 1–15.

Creative Assembly. (2014). *Alien isolation* [Video game]. Sega.

Crowston, K., & Prestopnik, N. R. (2013). Motivation and data quality in a citizen science game: A design science evaluation. *Proceedings of the 2013 46th Hawaii international conference on system sciences*, 450–459.

Crystal Dynamics. (2014). *Tomb radier: Definitive edition* (Version XBox One) [Video game]. Square-Enix.

Crytek. (2013). *Ryse: Son of Rome* (Version XBox One) [Video game]. Microsoft Studios.

Cunningham, S. J., & Jones, M. (2005). Autoethnography: A tool for practice and education. *Proceedings of the 6th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: making CHI natural*, 1–8.

Curtis, V. (2015). Motivation to participate in an online citizen science game: A study of Foldit. *Science Communication*, *37*(6), 723–746.

Cutting, J., & Cairns, P. (2020). Investigating game attention using the distraction recognition paradigm. *Behaviour & Information Technology*, 1–21.

Cyan. (1993). Myst.

Daft, R. L., Lengel, R. H., & Trevino, L. K. (1987). Message equivocality, media selection, and manager performance: Implications for information systems. *MIS quarterly*, 355–366.

D'Angiulli, A., & LeBeau, L. S. (2002). On boredom and experimentation in humans [PMID: 12956142]. *Ethics & Behavior*, *12*(2), 167–176. https://doi.org/10.1207/S15327019EB1202_4

Darley, J. M., & Latané, B. (1968). Bystander Intervention in Emergencies: Diffusion of Responsibility. *Journal of Personality and Social Psychology*, *8*(4), 377–383. https://doi.org/10.1037/h0025589

Das, R., Keep, B., Washington, P., & Riedel-Kruse, I. H. (2019). Scientific discovery games for biomedical research. *Annual review of biomedical data science*, *2*, 253–279.

De Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, *108*(11), 3266–3302.

De Quidt, J., Vesterlund, L., & Wilson, A. J. (2019). Experimenter demand effects. In *Handbook of research methods and applications in experimental economics*. Edward Elgar Publishing.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, *125*(6), 627.

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of research in personality*, *19*(2), 109–134.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, *11*(4), 227–268.

Denisova, A., & Cairns, P. (2015). The placebo effect in digital games: Phantom perception of adaptive artificial intelligence. *Proceedings of the 2015 annual symposium on computer-human interaction in play*, 23–33.

Derboven, J., Huyghe, J., & De Grooff, D. (2014). Designing voice interaction for people with physical and speech impairments. *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 217–226.

DeRight, J., & Jorgensen, R. S. (2015). I just want my research credit: Frequency of suboptimal effort in a non-clinical healthy undergraduate sample [PMID: 25494327]. *The Clinical Neuropsychologist*, *29*(1), 101–117. https://doi.org/10.1080/13854046.2014.989267

Deterding, S., Canossa, A., Harteveld, C., Cooper, S., Nacke, L., & Whitson, J. (2015). Gamifying research: Strategies, opportunities, challenges, ethics. *Conference on Human Factors in Computing Systems - Proceedings*, *18*.

Deterding, S. (2013). *Modes of play: A frame analytic account of video game play* (Doctoral dissertation). Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky.

Deterding, S. (2014). *Modes of play : A frame analytic account of video game play* (Doctoral dissertation). University of Hamburg.

Deterding, S. (2015). The Lens of Intrinsic Skill Atoms: A Method for Gameful Design. *Human-Computer Interaction*, *30*(3-4), 294–335. https://doi.org/10.1080/07370024.2014.993471

Deterding, S. (2016a). Contextual autonomy support in video game play: A grounded theory. *Proceedings of the 2016 CHI conference on human factors in computing systems*, 3931–3943.

Deterding, S. (2016b). Gameplay: Map or Frame? *CHI 2016 Workshop Games as an HCI Method*.

Deterding, S. (2017). Alibis for adult play: A goffmanian account of escaping embarrassment in adult play. *Games and Culture*. https://doi.org/10.1177/1555412017721086

Deterding, S. (2019). Interaction tension: A sociological model of attention and emotion demands in video gaming. *Media and Communication*, 226–236.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, 9–15.

Devlin, S., Cowling, P. I., Kudenko, D., Goumagias, N., Nucciareli, A., Cabras, I., Fernandes, K. J., & Li, F. (2014). Game Intelligence. *Computational Intelligence and Games*, 1–8. https://doi.org/10.1109/CIG.2014.6932917

Dillon, J., Dranove, D., Halpern, E., Hantoot, B., Munk, D., Pinsof, D., Temkin, M., & Weinstein, E. (2011). *Cards against humanity* [Card game]. Cards Against Humanity LLC.

Dix, A. (2010). Human–computer interaction: A stable discipline, a nascent science, and the growth of the long tail. *Interacting with computers*, *22*(1), 13–27.

Drachen, A., & Smith, J. H. (2008). Player talk—the functions of communication in multplayer role-playing games. *Computers in Entertainment*, *6*(4), 1. https://doi.org/10.1145/1461999.1462008

Ducheneaut, N., & Moore, R. J. (2004). The social side of gaming: A study of interaction patterns in a massively multiplayer online game. *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, 360–369.

Dunbar, G., Hill, R., & Lewis, V. (2001). Children's attentional skills and road behavior. *Journal of experimental psychology: Applied*, *7*(3), 227.

Duval, J., Rubin, Z., Segura, E. M., Friedman, N., Zlatanov, M., Yang, L., & Kurniawan, S. (2018). Spokeit: Building a mobile speech therapy experience. *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–12.

EA Canada. (2013). *Fifa 14* (Version Xbox One) [Video game]. Electronic Arts.

Echeverría, A., Barrios, E., Nussbaum, M., Améstica, M., & Leclerc, S. (2012). The Atomic Intrinsic Integration Approach: A structured methodology for the design of games for the conceptual understanding of physics. *Computers & Education*, *59*(2), 806–816.

Eckel, C. (2014). Economic games for social scientists. In *Laboratory experiments in the social sciences* (pp. 335–355). Elsevier.

Edvin & cuber3. (2005). Blow boat [Video game]. https://www.newgrounds.com/portal/view/251798

Eidos Montréal. (2016). *Deux ex: Mankind divided* (Version PlayStation 4) [Video game]. Square Enix.

Eisenbeiss, S. (2009). Contrast is the name of the game: Contrast-based semi-structured elicitation techniques for studies on children's language acquisition.

Eisenbeiss, S. (2010). Production methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 11–34). John Benjamins.

Elias, G. S., Garfield, R., & Gutschera, K. R. (2012). *Characteristics of Games*. MIT Press.

Ellis, C., Adams, T. E., & Bochner, A. P. (2011). Autoethnography: An overview. *Historical social research/Historische sozialforschung*, 273–290.

Elson, M., Breuer, J., Looy, J. V., Kneer, J., Quandt, T., Van Looy, J., Kneer, J., & Quandt, T. (2013). Comparing Apples and Oranges? Evidence for Pace of Action as a Confound in Research on Digital Games and Aggression. *Psychology of Popular Media Culture*, *4*(2). https://doi.org/10.1037/ppm0000010

Elson, M., & Quandt, T. (2016). Digital games in laboratory experiments: Controlling a complex stimulus through modding. *Psychology of Popular Media Culture*, *5*(1), 52–65. https://doi.org/10.1037/ppm0000033

Empire Wiki. (n.d.). *Calls* [Accessed 2022-02-11]. https://www.profounddecisions.co.uk/empire-wiki/Calls

Engelhardt, C. R., Hilgard, J., & Bartholow, B. D. (2015). Acute exposure to difficult (but not violent) video games dysregulates cognitive control. *Computers in Human Behavior*, *45*, 85–92. https://doi.org/10.1016/j.chb.2014.11.089

Ernest, J. (2011). *The big idea* [Board game]. Funforge.

Eveleigh, A., Jennett, C., Blandford, A., Brohan, P., & Cox, A. L. (2014). Designing for dabblers and deterring drop-outs in citizen science. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2985–2994.

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental press journal of orthodontics*, *19*(4), 27–29.

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, *26*(2), 132–139.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Ferguson, C. J. (2015). Do Angry Birds Make for Angry Children? A Meta-Analysis of Video Game Influences on Children's and Adolescents' Aggression, Mental Health, Prosocial Behavior, and Academic Performance. *Perspectives on Psychological Science*, *10*(5), 646–666. https://doi.org/10.1177/1745691615592234

Ferguson, C. J., Colon-Motas, K., Esser, C., Lanie, C., Purvis, S., & Williams, M. (2017). The (Not So) Evil Within? Agency in Video Game Choice and the Impact of Violent Content. *Simulation and Gaming*, *48*(3), 329–337. https://doi.org/10.1177/1046878116683521

Ferguson, C. J. (2007). The good, the bad and the ugly: A meta-analytic review of positive and negative effects of violent video games. *Psychiatric Quarterly*, *78*(4), 309–316. https://doi.org/10.1007/s11126-007-9056-9

Fernández-Aranda, F., Jiménez-Murcia, S., Santamaría, J. J., Gunnard, K., Soto, A., Kalapanidas, E., Bults, R. G., Davarakis, C., Ganchev, T., Granero, R., et al. (2012). Video games as a complementary therapy tool in mental disorders: Playmancer, a european multicentre study. *Journal of Mental Health*, *21*(4), 364–374.

Foroughi, C. K., Serraino, C., Parasuraman, R., & Boehm-Davis, D. A. (2016). Can we create a measure of fluid intelligence using Puzzle Creator within Portal 2? *Intelligence*, *56*, 58–64. https://doi.org/10.1016/j.intell.2016.02.011

Frey, A., Hartig, J., Ketzel, A., Zinkernagel, A., & Moosbrugger, H. (2007). The use of virtual environments based on a modification of the computer game quake iii arena® in psychological experimenting. *Computers in Human Behavior*, *23*(4), 2026–2039.

Friehs, M. A., Dechant, M., Vedress, S., Frings, C., & Mandryk, R. L. (2020). Effective gamification of the stop-signal task: Two controlled laboratory experiments. *JMIR Serious Games*, *8*(3), e17810. https://doi.org/10.2196/17810

Früh, W., & Schönbach, K. (2005). Der dynamisch-transaktionale Ansatz III: Eine Zwischenbilanz. *Publizistik*, *50*(1), 4–20. https://doi.org/10.1007/s11616-005-0115-7

Galli, L. (2014). Matching Game Mechanics and Human Computation Tasks in Games with a Purpose. *Proceedings of the 2014 ACM International Workshop on Serious Games.*, 9–14. https://doi.org/10.1145/2656719.2656727

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & gaming*, *33*(4), 441–467.

Gary, K., Stoll, R., Rallabhandi, P., Patwardhan, M., Hamel, D., Amresh, A., Pina, A., Cleary, K., & Quezado, Z. (2017). MHealth games as rewards: Incentive or distraction? *ACM International Conference Proceeding Series*, *Part F1286*, 7–10.

Gaston, J., & Cooper, S. (2017). To Three or not to Three: Improving Human Computation Game Onboarding with a Three-Star System. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 5034–5039.

Geoghegan, H., Dyke, A., Pateman, R., West, S., & Everett, G. (2016). *Understanding Motivations for Citizen Science. Final Report on behalf of the UK Environmental Observation Framework (UKEOF)* (tech. rep.). http://www.ukeof.org.uk/resources/citizen-science-resources/MotivationsforCSREPORTFINALMay2016.pdf

Glaser, B. G., & Strauss, A. L. (2010). *The discovery of grounded theory: Strategies for qualitative research*. Transaction Publishers.

Goffman, E. (1972). *Encounters: Two Studies in the Sociology of Interaction*. Penguin.

Goffman, E. (1983). The interaction order: American sociological association, 1982 presidential address. *American sociological review*, *48*(1), 1–17.

Goffman, E. (1986). *Frame Analysis: An Essay on the Organization of Experience*. Northeastern University Press.

Goh, D. H.-L., & Lee, C. S. (2011a). Perceptions, quality and motivational needs in image tagging human computation games. *Journal of Information Science*, *37*(5), 515–531.

Goh, D. H.-L., & Lee, C. S. (2011b). Understanding playability and motivational needs in human computation games. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7008 LNCS*, 108–117.

Goh, D. H.-L., Pe-Than, E. P. P., & Lee, C. S. (2017). Perceptions of virtual reward systems in crowdsourcing games. *Computers in Human Behavior*, *70*, 365–374.

Gollwitzer, P. M., & Oettingen, G. (2012). Goal Pursuit. In R. M. Ryan (Ed.), *The oxford handbook of human motivation* (pp. 208–231). Oxford University Press.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224.

Graesser, A. C. (2017). Reflections on serious games. In *Instructional techniques to facilitate learning and motivation of serious games* (pp. 199–212). Springer.

Grammenos, D., Savidis, A., & Stephanidis, C. (2005). Ua-chess: A universally accessible board game. *Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction, Las Vegas, Nevada (July 2005)*.

Grammenos, D., Savidis, A., & Stephanidis, C. (2007). Unified design of universally accessible games. *International Conference on Universal Access in Human-Computer Interaction*, 607–616.

Grammenos, D., Savidis, A., & Stephanidis, C. (2009). Designing universally accessible games. *Computers in Entertainment (CIE)*, *7*(1), 1–29.

Granic, I., Lobel, A., & Engels, R. C. M. E. (2014). The benefits of playing video games. *The American psychologist*, *69*(1), 66–78.

Gruenstein, A., McGraw, I., & Sutherland, A. (2009). A self-transcribing speech corpus: Collecting continuous speech with an online educational game. *International Workshop on Speech and Language Technology in Education*.

Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating Children to Learn Effectively : Exploring the Value of Intrinsic Integration in Educational Games. *The Journal of the Learning Sciences*, *20*(August), 169–206.

Habgood, M. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *The Journal of the Learning Sciences*, *20*(2), 169–206.

Halloran, J., Fitzpatrick, G., Rogers, Y., & Marshall, P. (2004). Does it matter if you don't know who's talking? multiplayer gaming with voiceover ip. *CHI'04 extended abstracts on Human factors in computing systems*, 1215–1218.

Hämäläinen, P., Mäki-Patola, T., Pulkki, V., & Airas, M. (2004). Musical computer games played by singing. *Proc. 7th Int. Conf. on Digital Audio Effects (DAFx'04), Naples*.

Hamari, J., & Tuunanen, J. (2014). Player Types: A Meta-synthesis. *ToDIGRA*, *1*(2), 29–53. https://doi.org/10.26503/todigra.v1i2.13

Harada, S., Wobbrock, J. O., & Landay, J. A. (2011). Voice games: Investigation into the use of non-speech voice input for making computer games more accessible. *IFIP Conference on Human-Computer Interaction*, 11–29.

Harmonix. (2015). *Rock band 4* [Video game].

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hawkins, G. E., Rae, B., Nesbitt, K. V., & Brown, S. D. (2013). Gamelike features might not improve data. *Behavior Research Methods*, *45*(2), 301–318.

Healy, K. (2017). Fuck Nuance. *Sociological Theory*, *35*(2), 118–127.

Heeks, R. (2010). Understanding "Gold Farming" and Real-Money Trading as the Intersection of Real and Virtual Economies. *Journal of Virtual Worlds Research*, *2*(4), 1–27.

Hersch, B. (1989). *Taboo* [Board game]. Hasbro.

Hilbig, B. E. (2016). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. https://doi.org/10.3758/s13428-015-0678-9

Hilgard, J. (2021). Maximal positive controls: A method for estimating the largest plausible effect size. *Journal of Experimental Social Psychology*, *93*, 104082.

Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of Anderson et al. (2010). *Psychological Bulletin*, *143*(7), 757–774. https://doi.org/10.1037/bul0000074

Hoonhout, J. (2008). Let the Game Tester Do the Talking: Think Aloud and Interviewing to Learn About the Game Experience. In K. Isbister & N. Schaffer (Eds.), *Game usability: Advice from the experts for advancing the player experience* (pp. 65–78). Morgan Kaufmann.

Hudson Soft. (2005). *Mario party 6* [Video game]. Nintendo.

Hughes, L. (2005). Beyond the rules of the game: Why are rooie rules nice?

Huizinga, J., & Hull, R. F. C. (1949). *Homo Ludens. A Study of the Play-element in Culture.[Translated by RFC Hull.]* Routledge & Kegan Paul.

Hunicke, R. (2005). The case for dynamic difficulty adjustment in games. *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, 429–433. https://doi.org/10.1145/1178477.1178573

Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. *Proceedings of the AAAI Workshop on Challenges in Game AI*, *4*(1), 1722.

Iacovides, I., Jennett, C., Cornish-Trestrail, C., & Cox, A. L. (2013). Do games attract or sustain engagement in citizen science? a study of volunteer motivations. In *Chi'13 extended abstracts on human factors in computing systems* (pp. 1101–1106). ACM.

id Software. (1992). Wolfenstein3D.

id Software. (1996). Quake.

Igarashi, T., & Hughes, J. F. (2001). Voice as sound: Using non-verbal voice input for interactive control. *Proceedings of the 14th annual ACM symposium on User interface software and technology*, 155–156.

Imangi Studios. (2011). *Temple run* (Version iOS).

Iridium Studios. (2015). *There came an echo* [Video game].

Isen, A. M., Shalker, T. E., Clark, M., & Karp, L. (1978). Affect, accessibility of material in memory, and behavior: A cognitive loop? *Journal of personality and social psychology*, *36*(1), 1.

Jamieson, P., Hall, J., & Grace, L. (2012). Research Directions for Pushing Harnessing Human Computation to Mainstream Video Games. *Meaningful Play 2012*.

Järvelä, S., Ekman, I., Kivikangas, J. M., & Ravaja, N. (2012). Digital games as experiment stimulus. *Proceedings of DiGRA Nordic*, 6–8.

Järvelä, S., Ekman, I., Kivikangas, J. M., & Ravaja, N. (2014). A practical guide to using digital games as an experiment stimulus. *Transactions of the Digital Games Research Association*, *1*(2).

Jenkins, J. G. (1946). Validity for What? *Journal of Consulting Psychology*, *10*, 93–98. https://doi.org/10.1037/h0059212

Johnson, D., Deterding, S., Kuhn, K.-A., Staneva, A., Stoyanov, S., & Hides, L. (2016). Gamification for health and wellbeing: A systematic review of the literature. *Internet Interventions*, *6*, 89–106. https://doi.org/10.1016/j.invent.2016.10.002

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*(1 SPEC. ISS.), 103–129.

Johnson, W. L. (2010). Serious use of a serious game for language learning. *International Journal of Artificial Intelligence in Education*, *20*(2), 175–195.

Jones, M. B., Kennedy, R. S., & Bittner Jr, A. C. (1981). A video game for performance testing. *The American Journal of Psychology*, 143–152.

Juul, J. (2005). *Half-Real: Video Games between Real Rules and Fictional Worlds*. MIT Press.

Juul, J. (2008). The magic circle and the puzzle piece. *Conference Proceedings of the Philosophy of Computer Games*, 55–67.

karaokeparty. (2010). Micropong [Video game]. https://www.newgrounds.com/portal/view/535064

Kawrykow, A., Roumanis, G., Kam, A., Kwak, D., Leung, C., Wu, C., Zarour, E., players, P., Sarmenta, L., Blanchette, M., et al. (2012). Phylo: A citizen science approach for improving multiple sequence alignment. *PloS one*, *7*(3), e31362.

Kelders, S. M., Kok, R. N., Ossebaard, H. C., & Van Gemert-Pijnen, J. E. (2012). Persuasive system design does matter: A systematic review of adherence to web-based interventions. *J Med Internet Res*, *14*(6), e152. https://doi.org/10.2196/jmir.2104

Kennedy, R., Bittner, A. C., & Jones, M. B. (1981). Video-game and conventional tracking. *Perceptual and motor skills*.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the mturk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629.

Keusch, F., & Zhang, C. (2017). A review of issues in gamified surveys. *Social Science Computer Review*, *35*(2), 147–166.

Khazaal, Y., van Singer, M., Chatton, A., Achab, S., Zullino, D., Rothen, S., Khan, R., Billieux, J., & Thorens, G. (2014). Does self-selection affect samples' representativeness in online surveys? An investigation in online video game research. *Journal of Medical Internet Research*, *16*(7). https://doi.org/10.2196/jmir.2759

Kiesling, M. (2017). *Azul* [Board game]. Plan B Games.

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *Internet and Higher Education*, *8*(1), 13–24. https://doi.org/10.1016/j.iheduc.2004.12.001

killthemouse. (2005). Deathmetal sim!!! [Video game]. https://www.newgrounds.com/portal/view/231172

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers and Education*, *87*, 340–356. https://doi.org/10.1016/j.compedu.2015.07.009

King. (2012). *Candy crush saga*.

Kirkwood, M. W., Kirk, J. W., Blaha, R. Z., & Wilson, P. (2010). Noncredible effort during pediatric neuropsychological exam: A case series and literature review [PMID: 20628928]. *Child Neuropsychology*, *16*(6), 604–618. https://doi.org/10.1080/09297049.2010.495059

Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., & Ravaja, N. (2011). A review of the use of psychophysiological methods in game research. *journal of gaming & virtual worlds*, *3*(3), 181–199.

Klimmt, C. (2006). *Computerspielen als Handeln: Dimensionen und Determinanten des Erlebens interaktiver Unterhaltungsangebote*. Herbert von Halem.

Klimmt, C., Vorderer, P., & Ritterfeld, U. (2007). Interactivity and Generalizability: New Media, New Challenges. *Communication Methods and Measures*, *1*(3), 169–179. https://doi.org/10.1080/19312450701434961

Kokkinakis, A. V., Cowling, P. I., Drachen, A., & Wade, A. R. (2017). Exploring the relationship between video game expertise and fluid intelligence. *Plos One*, *12*(11). https://doi.org/10.1371/journal.pone.0186621

Kordyaka, B., & Kruse, B. (2021). Curing toxicity–developing design principles to buffer toxic behaviour in massive multiplayer online games. *Safer Communities*.

Koster, R. (2005). *Theory of Fun for Game Design*. Paraglyph Press.

Kultima, A., Niemelä, J., Paavilainen, J., & Saarenpää, H. (2008). Designing game idea generation games. *Proceedings of the 2008 conference on future play: Research, play, share*, 137–144.

Kumari, S. (2021). *Design inspiration for motivating uncertainty in games using stage magic principles* (Doctoral dissertation). University of York.

Kuznekoff, J. H., & Rose, L. M. (2013). Communication in multiplayer gaming: Examining player responses to gender cues. *New Media & Society*, *15*(4), 541–556.

Lafrenière, M.-A. K., Verner-Filion, J., & Vallerand, R. J. (2012). Development and validation of the gaming motivation scale (gams). *Personality and individual differences*, *53*(7), 827–831.

Landers, R. N., Auer, E. M., Collmus, A. B., & Armstrong, M. B. (2018). Gamification Science, Its History and Future: Definitions and a Research Agenda. *Simulation and Gaming*. https://doi.org/10.1177/1046878118774385

Lapp, K. (2017). *Magic maze* [Board game]. Sit Down!

Latham, A. J., Patston, L. L. M., & Tippett, L. J. (2013). Just how expert are "expert" video-game players? Assessing the experience and expertise of video-game players across "action" video-game genres. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00941

Lazzaro, N. (2004). Why we play games: Four keys to more emotion in player experiences.

Leacock, M. (2010). *Forbidden island* [Board game]. Gamewright Games.

Lee, J., Kladwang, W., Lee, M., Cantu, D., Azizyan, M., Kim, H., Limpaecher, A., Gaikwad, S., Yoon, S., Treuille, A., Das, R., & null null. (2014). Rna design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences*, *111*(6), 2122–2127. https://doi.org/10.1073/pnas.1313039111

Leiner, D. J. (2013). Too Fast, Too Straight, Too Weird: Post Hoc Identification of Meaningless Data in Internet Surveys. *SSRN Electronic Journal*, (November 2013). https://doi.org/10.2139/ssrn.2361661

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*(3), 187–193.

Levine, R. V., & Norenzayan, A. (1999). The pace of fife in 31 countries. *Journal of Cross-Cultural Psychology*, *30*(2), 178–205. https://doi.org/10.1177/0022022199030002003

Levy, L., Solomon, R., Johnson, J., Wilson, J., Lambeth, A. J., Gandy, M., Moore, J., Way, J., & Liu, R. (2016). Grouches, extraverts, and jellyfish: Assessment validity and game mechanics in a gamified assessment. *DiGRA/FDG: Proceedings of the first international joint conference of DiGRA and FDG*, 1–16.

Lieberoth, A. (2015). Shallow gamification: Testing psychological effects of framing an activity as a game. *Games and Culture*, *10*(3), 229–248. https://doi.org/10.1177/1555412014559978

Light, B., et al. (2011). Interpreting digital gaming practices: Singstar as a technology of work.

Linehan, C., Kirman, B., & Roche, B. (2015). Gamification as behavioral psychology. *The gameful world: Approaches, issues, applications*, 81–105.

Linehan, C., Roche, B., Lawson, S., Doughty, M., Kirman, B., et al. (2009). A behavioural framework for designing educational computer games.

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among us-and india-based workers on mechanical turk. *Behavior research methods*, *47*(2), 519–528.

Little, B. (2000). Seaman us release walkthrough [Accessed 2022-02-10]. https://gamefaqs.gamespot.com/dreamcast/198567-seaman/faqs/8562

Littman, R. A., & Rosen, E. (1950a). Molar and molecular. *Psychological Review*, *57*(1), 58–65. https://doi.org/10.1037/h0056560

Littman, R. A., & Rosen, E. (1950b). Molar and molecular. *Psychological Review*, *57*(1), 58.

Lockheart, P. (2008). A mathematician's lament. https://www.maa.org/sites/default/files/pdf/devlin/LockhartsLament.pdf

Loughrey, K., & Broin, D. O. (2018). Are we having fun yet? Misapplying motivation to gamification. *2018 IEEE Games, Entertainment, Media Conference (GEM)*, 1–9.

Louvel, G. (2018). 'Play as if you were at home': dealing with biases and test validity. In A. Drachen, P. Mirza-Babaei, & L. E. Nacke (Eds.), *Games user research* (pp. 393–402). Oxford University Press.

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Serious Games*, *4*(2), e11. https://doi.org/10.2196/games.5888

Lyytinen, H., Ronimus, M., Alanko, A., Poikkeus, A.-M., & Taanila, M. (2007). Early identification of dyslexia and the use of computer game-based practice to support reading acquisition. *Nordic Psychology*, *59*(2), 109–126.

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive science*, *5*(4), 333–369.

Malone, T. (1980). What makes things fun to learn? Heuristics for designing instructional computer games. *Proceedings of the 3rd ACM SIGSMALL symposium*, 162–169.

Manninen, T. (2004). *Rich interaction model for game and virtual environment design*. Oulun yliopisto.

Mccabe, D. L., Treviño, L. K., & Butterfield, K. D. (2001). Cheating in Academic Institutions: A Decade of Research. *Ethics & Behavior*, *11*(3), 219–232.

McGee, K., Merritt, T., & Ong, C. (2011). What we have here is a failure of companionship. *Proceedings of the 23rd Australian Computer-Human Interaction Conference on - OzCHI '11*, 198–201. https://doi.org/10.1145/2071536.2071568

McGraw, I., Gruenstein, A., & Sutherland, A. (2009). A self-labeling speech corpus: Collecting spoken words with an online educational game. *Tenth Annual Conference of the International Speech Communication Association*.

McGraw, I., & Seneff, S. (2008). Speech-enabled card games for language learners. *AAAI*, 778–783.

McLaughlin, A., Gandy, M., Allaire, J., & Whitlock, L. (2012). Putting fun into video games for older adults. *Ergonomics in Design*, *20*(2), 13–22. https://doi.org/10.1177/1064804611435654

McMahan, R. P., Ragan, E. D., Leal, A., Beaton, R. J., & Bowman, D. A. (2011). Considerations for the use of commercial video games in controlled experiments. *Entertainment Computing*, *2*(1), 3–9.

Mekler, E. D., Bopp, J. A., Tuch, A. N., & Opwis, K. (2014). A systematic review of quantitative studies on the enjoyment of digital entertainment games. *Proceedings of the SIGCHI conference on human factors in computing systems*, 927–936.

Melhárt, D. (2018). Towards a comprehensive model of mediating frustration in videogames. *Game Studies*, *18*(1).

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & cognition*, *37*(8), 1077–1087.

Michie, S., & Johnston, M. (2013). Behavior change techniques. In *Encyclopedia of behavioral medicine* (pp. 182–187). Springer. https://doi.org/10.1007/978-1-4419-1005-9_1661

Mohseni, M. R., Liebold, B., & Pietschmann, D. (2015). Extensive modding for experimental game research. In *Game research methods* (pp. 323–340). ETC Press.

Molina, R., Unsworth, K., Hodkiewicz, M., & Adriasola, E. (2013). Are managerial pressure, technological control and intrinsic motivation effective in improving data quality? *Reliability Engineering & System Safety*, *119*, 26–34.

Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2008). Verification Strategies for Establishing Reliability and Validity in Qualitative Research. *International Journal of Qualitative Methods*, *1*(2), 13–22. https://doi.org/10.1177/160940690200100202

Muller, M. J., & Druin, A. (2012). Participatory design: The third space in human–computer interaction. In *The human–computer interaction handbook* (pp. 1125–1153). CRC Press.

Mummolo, J., & Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, *113*(2), 517–529.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie Du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Muramatsu, J., & Ackerman, M. S. (1998). Computing, social activity, and entertainment: A field study of a game mud. *Computer Supported Cooperative Work (CSCW)*, *7*(1), 87–122.

Murray, D. M. (1998). *Design and Analysis of Group-Randomized Trials*. Oxford University Press.

Mustaquim, M. M. (2013). Automatic speech recognition-an approach for designing inclusive games. *Multimedia tools and applications*, *66*(1), 131–146.

Nahid Golafshani. (2003). Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report*, *8*(4), 597–607.

Nakamura, J., & Csikszentmihalyi, M. (2014). The concept of flow. In *Flow and the foundations of positive psychology* (pp. 239–263). Springer.

Nintendo. (1987). *Super Mario bros* [Video game].

Nintendo. (2005a). Big brain academy [Video game].

Nintendo. (2005b). Brain age: Train your brain in minutes a day! [Video game].

Nintendo. (2009). *The legend of Zelda: Spirit tracks* (Version Nintendo DS) [Video game].

Norte, S., & Lobo, F. G. (2008). Sudoku access: A sudoku game for people with motor disabilities. *Proceedings of the 10th international ACM SIGACCESS Conference on Computers and Accessibility*, 161–168.

Ochs, J. (2010). *Snake oil* [Board game]. Out of the Box Publishing.

Oehrle, R. T. (1976). *The grammatical status of the english dative alternation* (Doctoral dissertation). Massahusetts Institute of Technology.

Office Create. (2006). *Cooking mama* (Version Nintendo DS) [Video game]. Majesco Entertainment.

O'Kane, A. A., Rogers, Y., & Blandford, A. E. (2014). Gaining empathy for non-routine mobile device use through autoethnography. *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, 987–990.

Oladimeji, P., Thimbleby, H., Curzon, P., Iacovides, I., & Cox, A. (2012). Exploring unlikely errors using video games: An example in number entry research.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, *45*(4), 867–872.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist*, *17*(11), 776.

Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls. Accademic Press, New York, NY.

Orne, M. T. (1981). The significance of unwitting cues for experimental outcomes: Toward a pragmatic approach. *Annals of the New York Academy of Sciences*, *364*(1), 152–159.

Orne, M. T., & Whitehouse, W. G. (2000). Demand characteristics. In A. Kazdin (Ed.), *Encyclopedia of psychology* (pp. 469–470). American Psychological Association.

Osborne, M. J., & Rubinstein, A. (1994). *A Course in Game Theory*. MIT Press.

Oulasvirta, A., & Hornbæk, K. (2021). Counterfactual thinking: What theories do in design. *International Journal of Human–Computer Interaction*, 1–15.

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

pandas development team, T. (2020). Pandas-dev/pandas: Pandas [Version 1.4.1]. https://doi.org/10.5281/zenodo.3509134

Pavlas, D., Jentsch, F., Salas, E., Fiore, S. M., & Sims, V. (2012). The play experience scale: Development and validation of a measure of play [PMID: 22624288]. *Human Factors*, *54*(2), 214–225. https://doi.org/10.1177/0018720811434513

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/https://doi.org/10.1016/j.jesp.2017.01.006

Pe-Than, E. P. P., Goh, D. H.-L., & Lee, C. S. (2012). A survey and typology of human computation games. *Proceedings of the 9th International Conference on Information Technology, ITNG 2012*, 720–725. https://doi.org/10.1109/ITNG.2012.124

Pham, M. T. (1996). Cue representation and selection effects of arousal on persuasion. *Journal of Consumer Research*, *22*(4), 373–387. https://doi.org/10.1086/209456

Playdots Inc. (2014). *Two dots*.

Ploog, B. O., Banerjee, S., & Brooks, P. J. (2009). Attention to prosody (intonation) and content in children with autism and in typical children using spoken sentences in a computer game. *Research in Autism Spectrum Disorders*, *3*(3), 743–758.

PopCap Games. (2001). Bejeweled [Video game].

Prestopnik, N., Crowston, K., & Wang, J. (2017). Gamers, citizen scientists, and data: Exploring participant contributions in two games with a purpose. *Computers in Human Behavior*, *68*, 254–268.

Prestopnik, N. R., Crowston, K., & Wang, J. (2014). Exploring Data Quality in Games With a Purpose. *iConference 2014 Proceedings*, (2008), 1–15. https://doi.org/10.9776/14066

Prestopnik, N. R., & Tang, J. (2015). Points, stories, worlds, and diegesis: Comparing player experiences in two citizen science games. *Computers in Human Behavior*, *52*, 492–506. https://doi.org/10.1016/j.chb.2015.05.051

Provo, F. (2000). Seaman review [Accessed 2022-02-10]. *Gamespot*. https://www.gamespot.com/reviews/seaman-review/1900-2613244/

Przybylski, A. K., Rigby, C. S., Deci, E. L., & Ryan, R. M. (2014). Competence-impeding electronic games and players' aggressive feelings, thoughts, and behaviors. *Journal of Personality and Social Psychology*, *106*(3), 441–457. https://doi.org/10.1037/a0034820

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A Motivational Model of Video Game Engagement. *Review of General Psychology*, *14*(2), 154–166.

Quinn, A., & Bederson, B. (2011). Human computation: a survey and taxonomy of a growing field. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1403–1412. https://doi.org/10.1145/1978942.1979148

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Cardamone, C., Murray, P., Schawinski, K., Szalay, A. S., & Vandenberg, J. (2013). Galaxy zoo: Motivations of citizen scientists. *arXiv preprint arXiv:1303.6886*.

Rapp, A. (2018). Autoethnography in human-computer interaction: Theory and practice. In *New directions in third wave human-computer interaction: Volume 2-methodologies* (pp. 25–42). Springer.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological bulletin*, *114*(3), 510.

Reeve, J. (2014). *Understanding motivation and emotion*. John Wiley & Sons.

Reeves, S., & Sherwood, S. (2010). Five design challenges for human computation. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries - NordiCHI '10*.

Reimers, S., & Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods*, *39*(3), 365–370. https://doi.org/10.3758/BF03193004

Rettie, R. (2004). Using Goffman's frameworks to explain presence and reality. *Presence 2004: Seventh Annual International Workshop*, 117–124.

Rico, J., & Brewster, S. (2010). Gesture and voice prototyping for early evaluations of social acceptability in multimodal interfaces. *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 1–9.

Riffenburgh, R. H. (2011). *Statistics in medicine*. Elsevier.

Rigby, S., & Ryan, R. M. (2011). *Glued to games: How video games draw us in and hold us spellbound: How video games draw us in and hold us spellbound*. AbC-CLIo.

rigel2010. (2009). Jet pass [Video game]. https://www.newgrounds.com/portal/view/513886

Riot Games. (2009). League of Legends.

Roepke, A. M., Jaffee, S. R., Riffle, O. M., McGonigal, J., Broome, R., & Maxwell, B. (2015). Randomized controlled trial of superbetter, a smartphone-based/internet-based self-help tool to reduce depressive symptoms. *Games for health journal*, *4*(3), 235–246.

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., & Vukovic, M. (2011). An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media An*, 321–328.

Rosnow, R., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. WH Freeman.

Ross, J., & Tomlinson, B. (2010). How games can redirect humanity's cognitive surplus for social good. *Computers in Entertainment*, *8*(4), Art. 25. https://doi.org/10.1145/1921141.1921145

Roth, C., Vermeulen, I., Vorderer, P., & Klimmt, C. (2012). Exploring Replay Value: Shifts and Continuities in User Experiences Between First and Second Exposure to an

Interactive Story. *Cyberpsychology, Behavior, and Social Networking*, *15*(7), 378–381. https://doi.org/10.1089/cyber.2011.0437

Roubira, J.-L. (2008). *Dixit* [Card game]. Libellud. https://www.libellud.com/wp-content/uploads/2019/06/DIXIT_RULES_EN.pdf

Rudchenko, D., Paek, T., & Badger, E. (2011). Text text revolution: A game that improves text entry on mobile touchscreen keyboards. *International Conference on Pervasive Computing*, 206–213.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, *55*(1), 68.

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*(4), 344–360. https://doi.org/10.1007/s11031-006-9051-8

Salen, K., & Zimmerman, E. (2004). *Rules of Play: Game Design Fundamentals*. MIT Press.

Salinäs, E.-L. (2002). Collaboration in multi-modal virtual worlds: Comparing touch, text, voice and video. In *The social life of avatars* (pp. 172–187). Springer.

Sarkar, A., Williams, M., Deterding, S., & Cooper, S. (2017). Engagement Effects of Player Rating System-Based Matchmaking for Level Ordering in Human Computation Games. *FDG'17*. https://doi.org/10.1145/3102071.3102093

Sarkar, S. (2015). There came an echo's voice controls may be the best way to play [Accessed 2022-02-11]. *Polygon*. https://www.polygon.com/2014/4/15/5606644/there-came-an-echo-preview-voice-controls-pax-east-2014

Sawin, D. A., & Scerbo, M. W. (1995). Effects of instruction type and boredom proneness in vigilance: Implications for boredom and workload. *Human factors*, *37*(4), 752–765.

Schell, J. (2009). *The Art of Game Design: A Book of Lenses*. Morgan Kaufman.

Schlenker, B. R., & Bonoma, T. V. (1978). Fun and Games : The Validity of Games for the Study of Conflict. *The Journal of Conflict Resolution*, *22*(1), 7–38. https://doi.org/10.1177/002200277802200102

Schmidt, R., Emmerich, K., & Schmidt, B. (2015). Applied games–in search of a new definition. *International Conference on Entertainment Computing*, 100–111.

Schrader, C., & Bastiaens, T. J. (2012). The influence of virtual presence: Effects on experienced cognitive load and learning outcomes in educational computer games.

Computers in Human Behavior, *28*(2), 648–658. https://doi.org/10.1016/j.chb.2011.11.011

Schrier, K. (2016). *Knowledge games: How playing games can solve problems, create insight, and make change*. JHU Press.

Schultheiss, O. C. (2008). Implicit motives. In *Handbook of personality: Theory and research* (pp. 602–633). Guilford.

Scott, J. (2000). Rational Choice Theory. In G. Browning, A. Halcli, & F. Webster (Eds.), *Understanding contemporary society: Theories of the present* (pp. 671–85). Sage.

Searle, J. R. (1965). What is a speech act. *Perspectives in the philosophy of language: a concise anthology*, *2000*, 253–268.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shaker, N., Togelius, J., & Nelson, M. J. (2016). *Procedural Content Generation in Games*. Springer. https://doi.org/10.1007/978-3-319-42716-4

Sharek, D., & Wiebe, E. (2011). Using flow theory to design video games as experimental stimuli. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 1520–1524.

Shaw, A. (2012). Do you identify as a gamer? Gender, race, sexuality, and gamer identity. *New Media and Society*, *14*(1), 28–44. https://doi.org/10.1177/1461444811410394

Sherry, J. L., Lucas, K., Greenberg, B. S., & Lachlan, K. (2006). Video game uses and gratifications as predictors of use and game preference. *Playing video games: Motives, responses, and consequences*, *24*(1), 213–224.

Sicart, M. (2008). Defining game mechanics. *Game Studies*, *8*(2), 1–14.

Sicart, M. (2014). *Play Matters*. MIT Press.

Sicart, M. (2015). Loops and Metagames: Understanding Game Design Structures. *Foundations of Digital Games*.

Silva, A., Mamede, N., Ferreira, A., Baptista, J., & Fernandes, J. (2011). Towards a serious game for portuguese learning. *International Conference on Serious Games Development and Applications*, 83–94.

Silvia, P. J. (2006). *Exploring the Psychology of Interest*. Oxford University Press.

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do "brain-training" programs work? *Psychological Science in the Public Interest*, *17*(3), 103–186.

Siu, K., & Riedl, M. O. (2016). Reward Systems in Human Computation Games. *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, 266–275.

Siu, K., Zook, A., & Riedl, M. O. (2017). A Framework for Exploring and Evaluating Mechanics in Human Computation Games. *FDG'17*.

Six to Start. (2012). *Zombies, run!* [Video game].

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., Pistrang, N., & Sanchez-Vives, M. V. (2006). A virtual reprise of the stanley milgram obedience experiments. *PloS one*, *1*(1), e39.

Slegers, K., Maurer, B., Bleumers, L., Krischkowsky, A., Duysburgh, P., & Blythe, M. (2016). Game-based HCI Methods: Workshop on Playfully Engaging Users in Design. *CHI Extended Abstracts on Human Factors in Computing Systems*, 3484–3491. https://doi.org/10.1145/2851581.2856476

Smith, J. L. (2004). Understanding the process of stereotype threat: A review of mediational variables and new performance goal directions. *Educational Psychology Review*, *16*(3), 177–206. https://doi.org/10.1023/B:EDPR.0000034020.20317.89

Smith, J. H. (2006). *Plans and Purposes: How Videogame Goals Shape Player Behaviour* (Doctoral dissertation). IT University of Copenhagen.

Smith, S. P., Blackmore, K., & Nesbitt, K. (2015). A Meta-Analysis of Data Collection in Serious Games Research. In C. S. Loh (Ed.), *Serious games analytics* (pp. 31–55). Springer. https://doi.org/10.1007/978-3-319-05834-4_2

Sniderman, S. (1999). Unwritten rules. *The Life of Games*, *1*(1), 2–7.

Soboczenski, F., Hudson, M., & Cairns, P. (2016). The effects of perceptual interference on number-entry errors. *Interacting with Computers*, *28*(2), 208–218.

Spencer, D. A. (2003). Love's labor's lost? The disutility of work and work avoidance in the economic analysis of labor supply. *Review of Social Economy*, *61*(2), 235–250+273.

Spencer, J. P., & Hund, A. M. (2002). Prototypes and particulars: Geometric and experience-dependent spatial categories. *Journal of Experimental Psychology: General*, *131*(1), 16.

Spencer, J. P., & Hund, A. M. (2003). Developmental continuity in the processes that underlie spatial recall. *Cognitive Psychology*, *47*(4), 432–480.

Spiers, H. J., Coutrot, A., & Hornberger, M. (2021). Explaining world-wide variation in navigation ability from millions of people: Citizen science project sea hero quest. *Topics in Cognitive Science*.

Sporka, A. J., Kurniawan, S. H., Mahmud, M., & Slavík, P. (2006). Non-speech input and speech recognition for real-time control of computer games. *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, 213–220.

Sproat, R., & Shih, C. (1991). The cross-linguistic distribution of adjective ordering restrictions. In *Interdisciplinary approaches to language* (pp. 565–593). Springer.

Squire, K. (2006). From content to context: Videogames as designed experience. *Educational researcher*, *35*(8), 19–29. http://edr.sagepub.com/content/35/8/19.short

Squire, K. (2011). *Video Games and Learning: Teaching and Participatory Culture in the Digital Age*. Teachers College Press.

Squire, K. D. (2008). Video Games and Education: Designing learning systems for an interactive age. *Educational Technology Magazine: The Magazine for Managers of Change in Education, 48*(2), 17–26.

Sra, M., Xu, X., & Maes, P. (2018). Breathvr: Leveraging breathing as a directly controlled interface for virtual reality games. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Stenros, J., Paavilainen, J., & Mäyrä, F. (2009). The many faces of sociability and social play in games. *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era on - MindTrek '09*, 82. https://doi.org/10.1145/1621841.1621857

Subrahmanyam, K., & Greenfield, P. M. (1994). Effect of video game practice on spatial skills in girls and boys. *Journal of Applied Developmental Psychology*, *15*(1), 13–32. https://doi.org/10.1016/0193-3973(94)90004-3

Tan, C. T., Rosser, D., & Harrold, N. (2013). Crowdsourcing facial expressions using popular gameplay. *SIGGRAPH Asia 2013 Technical Briefs*. https://doi.org/10.1145/2542355.2542388

Tan, C. T., Sapkota, H., & Rosser, D. (2014). Befaced: A casual game to crowdsource facial expressions in the wild. *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, 491–494. https://doi.org/10.1145/2559206.2574773

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*, *41*(2), 54–57.

Tennent, P., Rowland, D., Marshall, J., Egglestone, S. R., Harrison, A., Jaime, Z., Walker, B., & Benford, S. (2011). Breathalising games: Understanding the potential of breath control in game interfaces. *Proceedings of the 8th international conference on advances in computer entertainment technology*, 1–8.

Teuber, K. (2020). *Catan* [Board game]. Catan Studio. https://www.catan.com/sites/prod/files/2021-06/catan_base_rules_2020_200707.pdf

Thackray, R. I. (1981). The stress of boredom and monotony: A consideration of the evidence. *Psychosomatic medicine*.

Thorhauge, A. M. (2003). Player, reader and social actor. *Proceedings of the 5th International Digital Arts and Culture Conference*.

Tinati, R., Luczak-Roesch, M., Simperl, E., & Hall, W. (2017). An investigation of player motivations in eyewire, a gamified citizen science project. *Computers in Human Behavior*, *73*, 527–540.

Tower, W. (2007). The days and knights of Tom Murphy. *Washington Post*. https://www.washingtonpost.com/wp-dyn/content/article/2007/09/25/AR2007092501981_pf.html

Treiman, R. (1983). The structure of spoken syllables: Evidence from novel word games. *Cognition*, *15*(1-3), 49–74.

Tuite, K. (2014). Gwaps: Games with a problem. In M. Mateas, T. Barnes, & I. Bogost (Eds.), *Proceedings of the 9th international conference on the foundations of digital games, fdg 2014, liberty of the seas, caribbean, april 3-7, 2014*. Society for the Advancement of the Science of Digital Games.

Turkay, S., & Adinolf, S. (2015). The effects of customization on motivation in an extended study with a massively multiplayer online roleplaying game. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *9*(3). https://doi.org/10.5817/CP2015-3-2

Tyack, A., & Mekler, E. D. (2020). Self-determination theory in hci games research: Current uses and open questions. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–22.

Ubisoft Toronto. (2013). *Tom Clancy's splinter cell: Blacklist* (Version XBox One) [Video game]. Ubisoft.

Vaccarino, D. X. (2008). *Taboo* [Board game]. Rio Grande Games.

Vahlo, J., Kaakinen, J. K., Holm, S. K., & Koponen, A. (2017). Digital Game Dynamics Preferences and Player Types. *Journal of Computer-Mediated Communication*, *22*(2), 88–103. https://doi.org/10.1111/jcc4.12181

Vallerand, R. J., & Ratelle, C. F. (2002). Intrinsic and extrinsic motivation: A hierarchical model.

Valve. (2000). *Counter-strike* [Video game].

Van Berkel, N., Goncalves, J., Hosio, S., & Kostakos, V. (2017). Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–21.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Van Vught, J., & Glas, R. (2018). Considering play: From method to analysis. *Transactions of the Digital Games Research Association*, *4*(2).

Van Vught, J. F. (2016). *Neoformalist game analysis a methodological exploration of single-player game violence* (Doctoral dissertation). University of Waikato.

Vandercruysse, S., & Elen, J. (2017). Towards a game-based learning instructional design model focusing on integration. In *Instructional techniques to facilitate learning and motivation of serious games* (pp. 17–35). Springer.

van Grinsven, V. T. (2015). *Motivation in Business Survey Response Behaviour. Influencing motivation to improve survey outcome.*

Vermeir, J. F., White, M. J., Johnson, D., Crombez, G., & Van Ryckeghem, D. M. (2020). The effects of gamification on computerized cognitive training: Systematic review and meta-analysis. *JMIR serious games*, *8*(3), e18644.

Vésteinsdóttir, V., Joinson, A., Reips, U.-D., Danielsdottir, H. B., Thorarinsdottir, E. A., & Thorsdottir, F. (2019). Questions on honest responding. *Behavior Research Methods*, *51*(2), 811–825.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Visual Computing, HTW Berlin. (n.d.). *Apetopia* [Web game]. http://apetopia.visual-computing.com/

Vivarium. (1999). *Seaman* (Version Sega Dreamcast) [Video game]. Sega.

von Ahn, L. (2005). *Human computation* (Doctoral dissertation). Carnegie Mellon University.

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326.

von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, *51*(8), 58–67. https://doi.org/10.1145/1378704.1378719

von Ahn, L., Kedia, M., & Blum, M. (2006). Verbosity: A game for collecting common-sense facts. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 75–78. https://doi.org/10.1145/1124772.1124784

von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom: A game for locating objects in images. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 55–64. https://doi.org/10.1145/1124772.1124782

Wadley, G., Carter, M., & Gibbs, M. (2015). Voice in virtual worlds: The design, use, and influence of voice chat in online play. *Human–Computer Interaction*, *30*(3-4), 336–365. https://doi.org/10.1080/07370024.2014.987346

Wang, X., Goh, D. H.-L., Lim, E.-P., & Vu, E.-P. (2014). Player acceptance of human computation games: An aesthetic perspective. *16th International Conference on Asia-Pacific Digital Libraries, ICADL 2014*, *8839*, 233–242.

Wardaszko, M., Ćwil, M., Chojecki, P., & Dąbrowski, K. (2019). Analysis of matchmaking optimization systems potential in mobile esports. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2468–2475.

Washburn, D. A. (2003). The games psychologists play (and the data they provide). *Behavior Research Methods, Instruments, & Computers*, *35*(2), 185–193.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, *77*(4), 273.

Wenemark, M., Persson, A., Noorlind Brage, H., Svensson, T., & Kristenson, M. (2011). Applying Motivation Theory to Achieve Increased Respondent Satisfaction, Response Rate and Data Quality in a Self-administered Survey. *Journal of Official Statistics*, *27*(2), 393–414.

Werbach, K., & Hunter, D. (2012). *For the Win: How Game Thinking Can Revolutionize Your Business*. Wharton Digital Press.

Williams, D. (2010). The mapping principle, and a research framework for virtual worlds. *Communication Theory*, *20*(4), 451–470. https://doi.org/10.1111/j.1468-2885.2010.01371.x

Williams, D., Caplan, S., & Xiong, L. (2007). Can you hear me now? the impact of voice in an online gaming community. *Human communication research*, *33*(4), 427–449.

Williams, D., Yee, N., & Caplan, S. E. (2008). Who plays, how much, and why? Debunking the stereotypical gamer profile. *Journal of Computer-Mediated Communication*, *13*(4), 993–1018. https://doi.org/10.1111/j.1083-6101.2008.00428.x

Wizards of the Coast. (2014). *Dungeons & dragons: Player's handbook (5th edition)*.

Wouters, P., & Van Oostendorp, H. (2017). Overview of instructional techniques to facilitate learning and motivation of serious games. In *Instructional techniques to facilitate learning and motivation of serious games* (pp. 1–16). Springer.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van Der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, *105*(2), 249–265. https://doi.org/10.1037/a0031311

Wright, T., Boria, E., & Breidenbach, P. (2002). Creative player actions in fps online video games: Playing counter-strike. *Game studies*, *2*(2), 103–123.

Wu, J. (2016). *The copycat* [Board game].

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in cognitive sciences*, *8*(10), 451–456.

Yee, N. (2001). The virtual skinner box. *The Norrathian Scrolls: A Study of EverQuest*.

Yee, N. (2006). Motivations for play in online games. *CyberPsychology & behavior*, *9*(6), 772–775.

Zendle, D., Cairns, P., & Kudenko, D. (2015). Higher Graphical Fidelity Decreases Players' Access to Aggressive Concepts in Violent Video Games. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 241–251. https://doi.org/10.1145/2793107.2793113

Zendle, D., Cairns, P., & Kudenko, D. (2018). No priming in video games. *Computers in Human Behavior*, *78*, 113–125.

Zendle, D., Kudenko, D., & Cairns, P. (2018). Behavioural realism and the activation of aggressive concepts in violent video games. *Entertainment Computing*, *24*, 21–29.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98.