The University Of Sheffield.

Clinical Judgement: an investigation of clinical decision-making

By:

Benjamin Michael

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Clinical Psychology

The University of Sheffield
Faculty of Science
Clinical Psychology Unit, Department of Psychology

Submission Date: June 2019

I

This page is left intentionally blank.

Declaration

I hereby declare that this thesis has been submitted for the award of Doctorate in Clinical Psychology at the University of Sheffield. It has not been, and will not be, submitted for any other qualification or to any other academic institution. I confirm this work is my own and all other sources have been referenced appropriately.

This page is left intentionally blank.

Word Count

Literature Review	**8,272**
Including references and tables	14,690
Empirical Study	**8,310**
Including references and tables	11,619
Total	**16,582**
Including references and tables	26,309

This page is left intentionally blank.

Lay Summary

Mental Health Professionals (MHPs) make many important decisions in their daily practice which direct the care pathway and type of treatment delivered. Despite access to patient outcomes, the evidence base, clinical guidelines and diagnostic criteria research suggests many decisions are likely to be influenced by a number of dynamic and context variables. Consequently, clinical judgements and decisions are often prone to inaccuracy. The manner in which heuristics and biases influence the decisions of MHPs has not been fully investigated and the methods previously used have been rudimentary. By conducting a systematic review (including meta-analysis where possible) and two empirical studies this thesis aimed to contribute to the evidence-base regarding clinical judgement and decision-making.

The first part of this thesis describes the results of a systematic review and meta-analysis yielding 24 papers investigating the factors that have been shown to adversely influence accuracy of clinical judgement and decision-making. Results showed that several variables were implicated in this process. These include causal assumptions/causal theories, representativeness, contextual information, and race/culture. The decisions MHPs make are often inaccurate and evidence suggests that heuristics and biases are a probable cause for this.

The second part of this thesis reports two studies. Their aim, to design and test a trial-based methodology to assess the influence of bias on decisions regarding treatment allocation and progression. The first study developed an innovative 'real time' scenario-based approach (referred to as a 'dynamic measure') to assess clinical judgement and reasoning traits of Psychological Wellbeing Practitioners (PWPs) working as part of the Improved Access to Psychological Therapy (IAPT) programme. A non-systematic review of the cognitive biases and heuristics literature was conducted to develop a preliminary draft of the dynamic measure. This included a case vignette of a fictional male client referred to step 2 of the IAPT programme. Ethnographic decision-tree modelling was employed in the final stage of development. This incorporated qualitative thematic analysis of a focus group and a pilot study to develop two final versions of the dynamic measure. In

the second study 133 PWPs took part in an online survey completing two decision-making tasks (experimental and control). This was so that decisions when encountering a particularly challenging scenario during low intensity treatment could be compared with when treatment was relatively straightforward. Tasks included typical decisions a PWP is required to make during on-going clinical practice (e.g. assessing patient suitability for treatment, degree of alignment to treatment protocol and decisions when a client is not showing reliable improvement by session 4). The convergent validity of the dynamic measure as a test of heuristics and biases was not established but divergent validity was. Results suggest that the degree of treatment fidelity demonstrated by therapists and reasons they might sometimes prolong or conclude treatment may be due to an interaction between the therapist and the context. Given that the present study was explorative and convergent validity was not achieved further research is required.

This page is left intentionally blank.

Acknowledgements

I would like to thank my supervisors, Dr Stephen Kellett and Dr Jaime Delgadillo, for their honesty, enthusiasm and clear thinking throughout this project.

I am very grateful to all the teaching staff from the Sheffield University IAPT Programmes who helped me during the develop phase of my research.

I am thankful to each and every Psychological Wellbeing Practitioner who agreed to take part.

I really appreciate all the help from PWP Course Directors and staff during the recruitment phase. Further thanks go to staff members within the Psychological Professions Network, BPS PWP Training Committee and Health Education England who assisted me with this process.

To my loving wife, Hannah and two beautiful children, Grace and Jacob, whose births bookended my clinical training: I could not have done this without you. Your love, patience, honesty and solidarity mean everything.

Thank you and love always to my ever supportive Mum and Dad. I couldn't have got here without you.

Finally thank you to all the supervisors, teachers and mentors who have helped me along the way. I am so grateful for the trust and support you have shown towards me and believing that I am good enough to become a Clinical Psychologist.

This page is left intentionally blank.

Contents

Part One: Literature Review

How accurate are the decisions that mental health professionals' make?

A systematic review and meta-analysis

Abstract

Objective: Many patient and clinician factors are suggested as adversely influencing accuracy of clinical judgement and decision-making. This review aimed to identify studies reporting the cognitive processes of clinicians, validity of their judgments and utility of their decisions related to patient care.

Method: Three databases (Medline, PsycInfo, Scopus) were searched systematically. Studies eligible for inclusion explored clinical decisions mental health professionals (MHPs) make related to patient care and where a measure of decision/judgment accuracy was included. Studies were excluded where inclusion criteria were not met. An adapted critical appraisal tool was employed to assess methodological quality and was shown to demonstrate good inter-rater reliability in this review. Findings were synthesized in a narrative summary and where possible, meta-analysis.

Results: A total of 24 papers met eligibility criteria, a small set were eligible for meta-analysis (k = 4, N = 1,956). Variables highlighted as influencing clinical decision-making were grouped into twelve categories. Meta-analyses indicated the relationship between contextual information and diagnosis was significant (r = 0.41 (95% CI 0.37, 0.45), p = < 0.0001. Meta-analysis also showed the relationship between client race and diagnosis was significant (r = 0.34 (0.21, 0.47), p = < 0.0001). Other variables found to influence accuracy of clinical decision-making included causal assumptions/causal theories and representativeness.

Conclusions: Findings suggest MHPs often make inaccurate clinical decisions influenced by a number of dynamic and context variables. Heuristics and biases may also influence such decisions.

Practitioner Points

1. It would be beneficial to provide empirically guided feedback to students and trainees as to what information to attend to during assessment and treatment.
2. Results relating to diagnostic decision-making have serious consequences on mental disorder diagnosis. When biases occur, this happens because clinicians are attending to stereotypes rather than base rates.
3. A key limitation to the present review is the lack of follow up studies and the use of valid measures and experimental designs that intervene (and train better) decision making amongst clinicians.
4. Much of the research identified in this review relates to diagnosis. This limits the ability to generalise the findings to other mental health professions. Greater understanding is required as to the influence of bias on treatment allocation and progression decisions.

Introduction

Clinicians routinely make many important decisions in their practice which direct the care pathway and type of treatment delivered. Wrong decisions (e.g. the wrong diagnosis) will impact on clinical outcomes and the experience of care. Research suggests that clinicians frequently deviate from recommendations for evidence-based practice (Garb, 2005). Dumont and Lecomte (1987) argue that clinicians engage in work that is highly inferential in nature and which "offers continuous opportunity for error in matters that have the most profound consequences in the lives of their clients" (p.434).

For the last 65 years the subject of clinical judgement and decision-making has been an area of particular interest to researchers. Meehl (1954) introduced the “statistical versus clinical controversy” and since then reviewers and researchers alike have put forward the argument that clinical judgement is often flawed, and that actuarial (i.e. statistical) prediction tends to be more reliable. Actuarial prediction refers to any prediction of behaviour founded solely on statistical information rather than subjective judgement. Hannan et al. (2005) found that clinicians are often poor at identifying those clients who are unlikely to benefit from therapy. Clinical prediction was compared to an algorithm designed to identify clients at risk of treatment failure. At the end of each session therapists were asked to estimate if the patient was worse now compared to the start of therapy and patient outcome by the end of treatment. Three of 550 clients were predicted to deteriorate, one of whom actually did. The therapists failed to predict 39 additional clients who deteriorated during treatment. Conversely the empirical prediction method identified 77% of the patients who went onto deteriorate but also generated numerous false-positive results. The accuracy of clinical versus mechanical (formal, statistical) data-combination techniques has also been demonstrated in several meta-analytic studies. Grove et al. (2000) found that mechanical prediction greatly outperformed clinical prediction in 33%-47% of the 136 studies investigated. Ægisdóttir et al. (2006) obtained 92 effect sizes from 67 studies. There was improved statistical accuracy relative to clinical methods to highlight a 13% improvement in accuracy using statistical over clinical

prediction techniques (overall reported effect size: -.12, 95%, confidence interval did not cross zero). The meta analytic evidence mirrors previous narrative reviews of clinical and statistical prediction (e.g., Dawes et al., 1989; Grove & Meehl, 1996; Meehl, 1954; Sawyer, 1966).

Research regarding accuracy of decision making has evolved into identifying the patient and clinician factors that may adversely influence accuracy of clinical judgement and decision-making. Patient variables such as race (Yamamoto, James, Bloombaum, & Hattem, 1967) and social class (Haase, 1964) were some of the first to be examined. Therapists with lower ethnocentricity were found to treat more patients from ethnic-minorities than those where it was higher. Mental illness was reported more often when a client's background was described as 'lower-class' than 'middle or upper-class'. Since then a significant amount of research has shown that informal observations are difficult to learn from. For example, meta-analyses have shown that reaching accurate conclusions from nonverbal behaviour is often problematic (Ambady & Rosenthal, 1992). One explanation as to why cognitive processes are prone to error is that clinicians do not receive accurate feedback regarding the validity of their judgments (Chapman & Chapman 1969; Garb 1989, 1998). Therefore, learning from clinical experience often does not occur.

To date, there is a dearth of reviews that evaluate clinical prediction employing a systematic review approach. Garb (2005) offers an in-depth review of the clinical judgment and decision-making literature. However, this review was conducted fourteen years ago, is now therefore dated and also the search strategy was not developed using best practice systematic review guidelines (Centre for Reviews and Dissemination, 2009). This is a significant limitation and presents a considerable gap in the literature on the accuracy of clinical prediction, since non-systematic reviews can be limited by selection bias.

Researchers and theorists alike have attempted to understand the processes involved in clinical judgment and decision-making in order to understand the subsequent mistakes which then occur. Garb (2005) argued for the development of research that includes the study of heuristics and biases. The term *bias* commonly denotes a prejudgment or prejudice. Within the context of clinical

prediction, it is often used to refer to *error* or *inaccuracy* in clinical judgments (Lopez, 1989). Heuristics refer to mental shortcuts that ease the cognitive load when making a decision (Myers, 2010). Arkes (1981) considered the influence of preconceived notions, lack of awareness of one's own judgmental processes, overconfidence, and the role of hindsight bias on inaccurate clinical decision-making. Faust (1986) suggested that bad habits (e.g. underuse of base rates) and cognitive limitations (e.g. inability to process multiple-cue tasks) are two reasons for cognitively based judgment errors. Snyder (1981) suggested that clinicians will often seek to gather confirmatory information when deciding, a process often referred to as confirmatory bias.

Tversky and Kahneman's (1974) theory of the role of heuristics and bias are frequently cited in demonstrating how clinicians make inaccurate judgments (e.g., Arkes 1981, Dawes 1986, Garb 1998, Kayne & Alloy 1988, Turk & Salovey 1988, Wedding & Faust 1989). Heuristics and biases research had a strong influence on the development of prospect theory (Kahneman & Tversky, 1979). This predicts risk-aversive behaviour when decisions are framed in terms of possible gains but risk-taking when decisions are framed in terms of losses (Fiedler, & von Sydow, 2015). Therefore, a fruitful avenue of research is to investigate the manner in which heuristics and biases have been applied as a way to understand the processes involved in clinical decision-making.

Aims

The main objective of this review was to conduct a systematic literature search to identify all relevant studies reporting the cognitive processes of clinicians, the validity of their judgments, and utility of their decisions related to patient care (Garb, 2005). As Garb's (2005) review already sets out the scope of the clinical-judgement and decision-making literature a systematic narrative review and, where studies reported appropriate outcomes, selective meta-analysis was planned. Investigations that specifically explored the beliefs, attitudes, and subjective norms potentially influencing clinical decision-making were of particular interest. The primary outcome was the accuracy of clinician's causal judgments, behavioural predictions and treatment decisions. Studies were included where a measure of decision/judgment accuracy was discussed (e.g. patient outcomes,

evidence base, clinical guidelines, diagnostic criteria). Research might explore diagnostic decision-making, behavioural prediction (e.g. estimating risk), and decisions relating to treatment plans (e.g. the advantages of standardized rather than tailored treatment plans).

Methodology

Study Protocol

The systematic review protocol was registered and published in the International Prospective Register of Systematic Reviews (PROSPERO) ahead of conducting the review. (Protocol ID: PROSPERO 2019: CRD42018109651).

Search Strategy

Three research databases (Medline, PsycInfo, and Scopus) were systematically searched on 20th January 2019. To ensure a comprehensive search synonyms were included in the search terms, mapped onto relevant subject headings (when available) and 'exploded' to include other related subject headings. No restrictions were applied in terms of date of publication. Reverse and forwards citation searches took place after screening articles according to title and abstract for eligibility. All literature published up to the date of the search were considered for inclusion. Additional references were also sought by contacting authors of previous reviews and meta-analyses cited in the introduction. Further details as to the search strategy can be found in Appendix A.

Inclusion and Exclusion Criteria

Eligibility for this review was ascertained through a process of title inspection and full article inspection. Studies were included where researchers defined participants as health professionals who work with mental health problems. In addition to variations on the general term ‘mental health professional’ the literature review also included a core lists of health professionals (Table 1), as defined by the International Standard Classification of Occupations (ISCO; International Labour Organisation, 2016). Studies that recruited mental health professionals (MHPs) ‘in training’ as participants were also included. Studies were excluded that included Community Health Workers,

Social Work Associate Professionals, or Medical Health Professionals. This was so that, specifically, the decision-making accuracy of MHPs relative to the regulatory standards and guidelines they are required to follow (e.g. Health and Care Professions Council, 2019; Royal College of Psychiatrists, 2019) could be considered.

Articles that prioritised exploring the cognitive processes of clinicians, validity of judgments, and utility of decisions were of interest (Garb, 2005). Any article that specifically applied a formal method (e.g. mechanical, algorithmic) over clinical judgement to reach the decision was excluded. Studies were included that explored clinical decisions MHPs made related to patient care and where a measure of decision/judgment accuracy was included. Clinical decision-making outcomes could include: (i) interrater reliability, (ii) results from a single test compared with judgments based on an interview and history information, (iii) judgments compared with results from behaviour record forms to measure daily activities (e.g. Wu & Clark, 2003), and (iv) principal components analysis (e.g. Kraemer et al., 2003; the measurement of characteristics, context, perspective, and error of measurement). Other means of measuring decision/judgment accuracy also included those that consulted treatment outcomes, the evidence base, clinical guidelines, and diagnostic criteria. Studies were excluded if they were not in English or were not empirical studies (e.g., a narrative review) and/or multiple articles by an author that utilised the same data set. Editorials, newspaper articles and other forms of popular media were also excluded, as were papers not published in a peer-reviewed journal.

Table 1.

List of mental health related occupations included in review as defined by the International Standard Classification of Occupations (ISCO)

Occupation	Classification of Occupation
Psychiatrists (ISCO	(ISCO-08 minor group 221, unit 2212)
Clinical Psychologists (ISCO	(ISCO-08 minor group 263, unit 2634)
Social Work and Counselling Professionals (ISCO	(ISCO-08 minor group 263, unit 2635)

Quality Assessment

The methodological quality of selected studies was assessed. As the studies included in this review were methodologically diverse, an adapted risk of bias tool was used for both randomised and non-randomised studies (Downs & Black, 1998). Similar to previous studies (e.g. Larson, Vos, & Fernandez, 2013) item 27 was simplified. A score of 1 was given where studies reported that power was achieved, and 0 was given where no sample size calculation was stated. The wording of some questions was altered to better relate to the papers included in this review (e.g., 'patients' changed to 'participants' and ‘intervention’ changed to ‘condition’). Reference to case-control studies was omitted as this was not included in the design of any of the papers included in this review. Questions 13 and 19 were omitted from the checklist as these were not relevant to any of the papers included. The adapted checklist can be found in Appendix B. The total possible range on the adapted critical appraisal tool was 0-24. Due to different study designs not all items were applicable meaning differences in scores between studies was likely. Therefore, the qualitative descriptors were adapted as follows(Hooper, Jutai, Strong, & Russell-Minda, 2008): excellent (26-28); good (20-25); fair (15-19); and poor (≤14). A percentage was calculated from the total score divided by the number of items included to provide a method for comparison (Table 2).

Table 2.

Percentage calculated from Downs and Black (1998) total score

Rating	Points score from	Points score to	% from	% to
Excellent	26	28	93%	100%
Good	20	25	71%	89%
Fair	15	19	54%	68%
Poor	14	0	50%	0%

A second reviewer acted as an independent assessor, repeating the quality appraisal on a random 25% of the included studies. Kappa scores indicated excellent inter-rater reliability, *ICC* = .92 (Koo & Li, 2016). Any disagreements in ratings were resolved through discussion. No papers were

removed based on quality assessment score to ensure that the review reflected the breadth of the literature base.

Data Extraction and Synthesis Method

Included studies were examined for variables highlighted as influencing clinical decision-making and these were grouped into twelve categories (see results section). For many statistical analysis was not appropriate due to the diversity in methodology, outcomes assessed, and measures used. Therefore, for these studies data were synthesized in a narrative summary. Where meta-analysis was possible a statistical synthesis of quantitative data was implemented employing random effects meta-analysis using the R package Meta-Analysis via Shiny (MAVIS; Hamilton, 2011). In accordance with standard recommendations (Borenstein, Hedges, Higgins & Rothstein, 2009; Valentine, Pigott, & Rothstein, 2010) meta-analysis was limited to predictors investigated across two or more studies and where related significance tests were reported. To assist in the analysis relevant inferential statistics (e.g., odds ratios, chi-square tests, t-tests) were transformed into correlation coefficients (r) to achieve standardisation (Borenstein et al., 2009).

Where effect sizes were not comparable or reported and where means and SDs were not available authors were contacted via email to request further information. Only 1 author replied. Subsequently, where reported effect sizes were not comparable these studies were included as part of the narrative synthesis. Potential publication bias was examined using regression and rank-correlation tests and, where possible, visual inspection of funnel plot asymmetry. Additionally, the Fail-safe N calculation was performed using Rosenthal's method (Orwin, 1983). The small number of studies (≤3) entered into meta-analysis meant that more detailed subgroup or moderator analyses was not possible.

Results

This review utilised PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Moher et al., 2015). Figure 2 illustrates the process of selecting papers for inclusion. Database searches returned 1797 articles, from which 522 duplicate results were removed. Titles and abstracts of the remaining 1275 citations were screened against inclusion criteria and a further 1174 were excluded. This meant 101 full-text articles were assessed for eligibility using the

inclusion criteria. From these papers a further 25 full-text articles were added after reverse and forwards citation. One-hundred and twenty-six full-text articles were assessed for eligibility using the inclusion criteria. Of these, 9 were review articles, 5 articles did not discuss a measure of decision/judgment accuracy, 48 articles prioritised the formal method (e.g. mechanical, algorithmic) over clinical judgment to reach the decision, 4 articles were unavailable and 36 did not meet the inclusion criteria of the current review. Therefore, these 102 papers were excluded. Twenty-four papers were included in the final review and a summary of key aspects of these papers is included in Table 3, of which N= 4 were appropriate for meta-analysis.

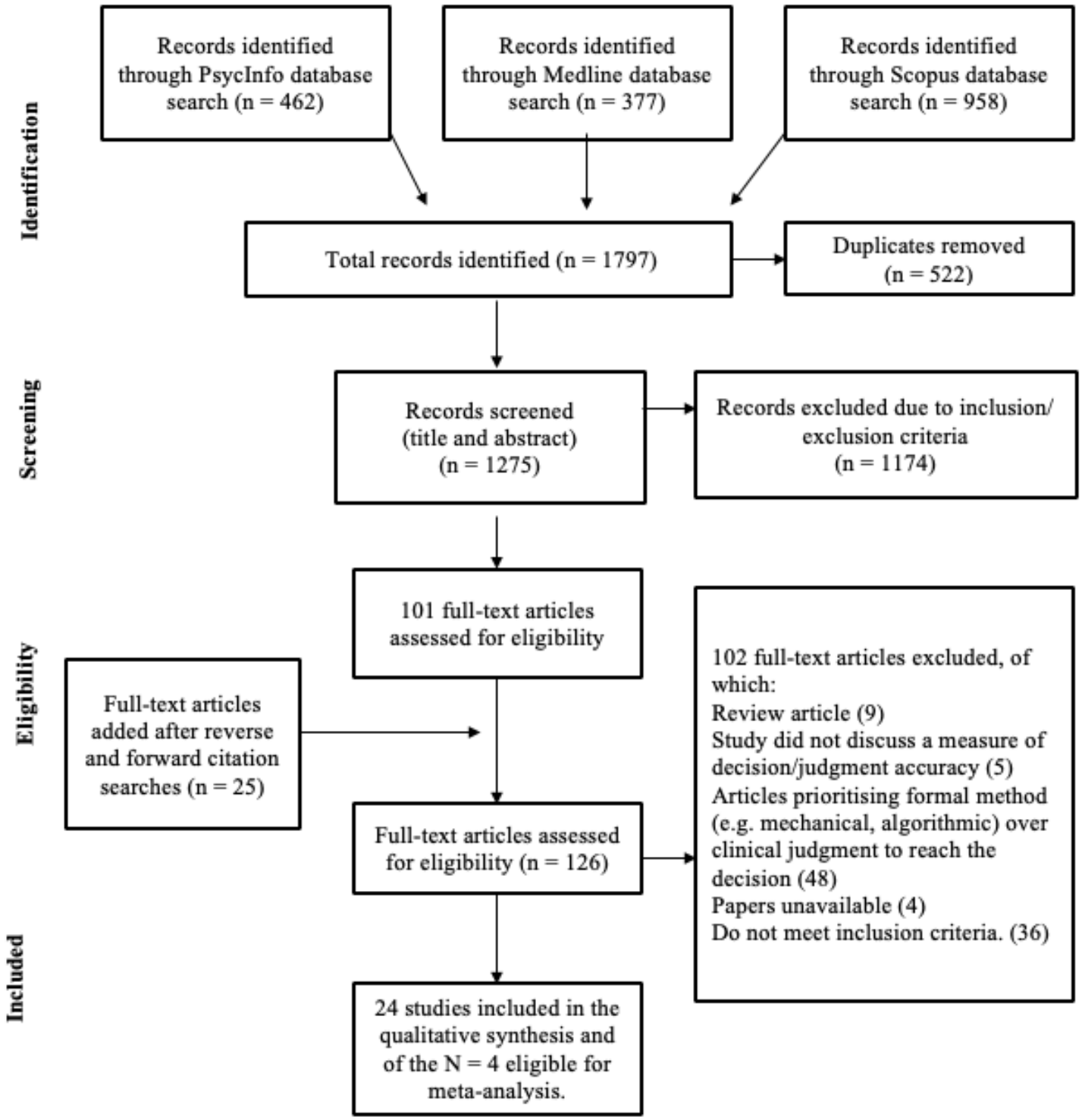


Figure 2**.** Flow diagram summarising the article selection process for the systematic review

Table 3. *Summary of the methodologies of included studies*

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Aarts, Witteman, Souren, & Egger (2012) The Netherlands	Cross-sectional Survey	46	Diagnostic decision-making	Clinical Psychologists Trainees & Clinical Psychologists	Implicit/rapid/automatic thinking processes vs slow/consciously monitored / deliberately controlled	-	Diagnostic classification accuracy according to the DSM IV-casebook.	Rational-experiential inventory (REI)	Significant differences found in diagnostic accuracy according to rationality score ($F = 4.356$, $p = 0.019$). Higher psychologists' rationality & the more they thought about a prototypical client = poorer diagnostic classification accuracy.	69
Berman, Tung, Matheny, Glenn Cohen & Wilhelm (2016) USA	Experimental Study	262	Clinical decision-making regarding suicide risk	Mental health clinicians	How patient age and clinician demographics and training factors moderate clinicians' perception of risk.	Two vignette conditions, patients' age manipulated.	AAS evidence-base re. difference in suicide rates according to age.	Two questions relating to likelihood of patient suicide and decision to hospitalize the patient.	Clinician age may reveal a "similarity" bias. Clinicians perceive those who are different (i.e., older or younger) to be at greater risk. $F(3, 254)=4.28$, $p <.01$, $R^2=.05$.	73
Blashfield, Sprock, Pinkston& Hodgin (1985) USA	Cross-sectional Survey	20	Diagnostic decision-making	Psychiatrists (faculty and resident) and Clinical Psychologists (faculty and graduate students)	Diagnostic decision-making relative to the classification of PD.	-	DSM-III Case Book, psychiatric textbooks, journal articles, and summaries of real cases. Inter-clinician reliability.	Calculation of a disagreement statistic relative to diagnosis suggested by source of case.	Prototypic cases discovered for 8/11 PDs. Using reaction time & agreement data produced no diff between professions or experience. Experience had sig effect on diagnostic speed ($F = 49.95$, $P < .0001$), but not on agreement. Future research can apply prototype model.	60

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Bruchmüller & Meyer (2009) Germany	Quasi-experimental Study	185	Diagnostic decision-making	Psychotherapists	Subjective causal assumptions of the therapists.	Five vignette conditions. Participants received either a unipolar vignette condition, a bipolar vignette condition where information relevant to sleep manipulated, or causal explanation manipulated.	Pretest including four experts asked to make diagnosis according to ICD-10 and DSM-IV.	Questionnaire listing ICD-10 F-codes related to diagnosis.	Therapists don't make diagnoses as DSM-IV and ICD-10 requires. They discount bipolar symptoms if a rational and understandable explanation is provided. E.g., BD diagnosis higher when additional information of reduced sleep compared to normal sleep provided (73% & 38%). But significant interaction = sleep and relationship (OR=0.16, $p < .05$) showed if one piece of additional information pointed away from BD this influenced diagnosis.	59
Bruchmüller, Margraf & Schneider (2012) Germany	Experimental Study	463	Diagnostic decision-making	Psychologists, Psychiatrists & Social Workers	Presence of the representativeness heuristic in therapists in regard to diagnosing ADHD.	Four vignette conditions. patients' gender manipulated for each.	Two pretests including experienced researchers and trained diagnosticians using DSM–IV/ICD–10 criteria.	Questionnaire listing ICD-10 F-codes related to ADHD diagnosis.	Diagnostic manuals are not strictly followed by therapists. Over diagnosis of ADHD occurs and is influenced by patient's gender (OR = 2.66, $p < .034$). Boys significantly more likely to be incorrectly diagnosed with ADHD than incorrectly given another diagnosis ($\chi 2(1, N = 226) = 7.12$, $p < .008$.)	91

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Cwik & Margraf (2017) Germany	Experimental Study	120	Diagnostic decision-making	Psychiatrists, Clinical Psychologists & Trainee Clinical Psychologists	Presence of information order effects in diagnosis related decision-making.	Pretreatment report or no pretreatment report. Core symptoms at the beginning of the vignettes or at the end	One pretest including 7 experienced clinicians using DSM‑5 and/or ICD‑10 diagnostic criteria.	Diagnostic assessment of vignette choosing up to 3 diagnoses out of 19 listed. Participants also asked to indicate whether chosen disorder was present with clinical or subclinical intensity.	Results suggest that the accuracy of diagnostic decisions was predicted by order of symptom descriptors with a recency effect initiating more fully correct diagnostic decisions where diagnostic information was presented last (GAD: OR = 2.89, P = .017; PD: OR = 2.65, P = .024). Receiving incongruent pretreatment reports was not predictive for diagnostic errors.	78
De Los Reyes & Marsh (2011) USA	Cross-sectional Survey	45	Diagnostic decision-making	Psychologists, Counselors and Social Workers	Presence of contextual information upon clinician impression of conduct disorder symptomology.	-	Presentation of conduct disorder symptoms according to DSM–IV (1 in each vignette). Presence of empirically tested contextual risks factors relating to conduct disorder in ‘consistent context’ condition.	A likelihood rating (0-100) for each vignette worded as, ‘‘How likely would a youth with the given life factors be found to have Conduct Disorder if a full clinical evaluation was given.’’	Contextual information highly impacted clinician judgments when consistent with conduct disorder compared with when not ($F(1,44) = 120.1$, $p<.001$, $\eta 2 = .73$). Variation across symptom agreement between clinicians, however. Authors claim these findings are of great consequence to clinical science and practice.	72

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
DeRoma, Hansen, Tishelm, & D'Amico (1997) USA	Experimental Study	40	Decisions related to behavioural prediction	Social Workers, Clinical Psychologists & Trainee Clinical Psychologists	The influence of access to information on evaluative responding.	Three vignette conditions randomised.for (i) presence/absence of behavior problems, (ii) presence/ absence of abuse, (iii) gender of child.	Evidence-base re. effects of abuse across the lifetime for abused/maltreated children.	Ratings along five treatment-related dimensions and four scales related to social functioning.	A history of maltreatment influenced professional judgments . E.g. vignettes with no abuse history or behavior problems rated significantly higher regarding predicted stability (F(1, 19) -- 13.27, p < .002), and treatment referral. Children buffered from negative effects of abuse may be overlooked. Inaccurate judgments may be directed toward maltreated children.	64
Evans, Herbert, Nelson Gray, Gaudiano (2002) USA	Experimental Study	32	Diagnostic decision-making	Clinical & Counseling Psychologists	Determinants of diagnostic prototypicality judgments relative to PD diagnoses.	12 profiles of hypothetical patients whereby 3 factors varied (high vs low category number, high vs medium typicality, high vs low dominance)	3 factors relevant to PD according to DSM-III-R criteria, Inter-clinician reliability.	1-7 Typicality likert rating scale according to 11 DSM-III-R PD.	Typicality and dominance showed strong effects ($F(1,222) = 31.52$, $p < .0001$; $F(1,222) = 13.14$, $p < .0001$) whereas no effects were found for number. Authors conclude cases more prototypic of a specific PD contain highly typical features and those predominantly associated with the diagnostic category.	73

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Ford & Widiger (1989) USA	Experimental Study	354	Diagnostic decision-making	Psychologists	Sex bias in the diagnosis of HPD & APD.	Nine case histories randomised for (i) variation in ambiguity of information relevant to APD & HPD diagnoses (ii) gender of client specified/unspecified. Order of individual behaviours relative to DSM III HPD or APD criterion randomized.	Diagnostic classification accuracy according to presentation method of the DSM-III criteria relevant to HPD & APD.	7-point rating scale of extent to which patient appeared to have each of four Axis I and five Axis II disorders.	Base rate differences in HPD &APD are not only reason for sex differences in diagnosis. Clinicians consider base rates when case history information is ambiguous. When it's less ambiguous males significantly less likely to be diagnosed HPD than females ($\chi 2$ (2, N = 93) = 6.9, p < =.05). APD significantly more often failed to be diagnosed in females than males ($\chi 2$ (2, N = 95) = 8.8, p < =.05). Antisocial female patients significantly more likely to be diagnosed with HPD than with APD ($\chi 2$ (2, N = 95) = 12.6, p < =.01).	61
Fuller & Cowan (1999) United Kingdom	Cross-sectional Survey	-	Clinical decision-making regarding risk predictions	Mental Health Professionals	Consensus judgments relating to the prediction of patient-related risks.	-	Aggregate risk score matched with patient outcome data.	Team consensus risk predictions	CP accuracy of patient-related risks similar to AB studies across comparable time frames. (CP = 0.71, AB = 0.76).	62

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Garb (1996) USA	*Study 1:* Cross-sectional survey *Study 2:* Quasi-experimental *Study 3:* Cross-sectional survey (x2)	*Study 1:* 67 *Study 2:* 59 *Study 3:* a: 107 b: 74	*Study 1 and 2:* Diagnostic decision-making *Study 3:* Decisions related to behavioral predictions	Psychologists and psychology interns from certified psychology internship training programs.	*Study 1 and 2* were interested to predominantly explore the presence of the representativeness heuristic. *Study 3* explored the representativeness heuristic and the past-behavior heuristic.	*Study 1:* - *Study 2:* Ethnic origin of patient in case history manipulated (African-American or White) *Study 3:* -	General statements about importance in attending to criteria contained in DSM-IV.	*Study 1:* Likelihood, similarity and confidence ratings related to three different disorders. *Study 2:* Likelihood and similarity ratings related to three different disorders. *Study 3:* Behavioral predictions, similarity ratings, base rate estimates related to DSM III-R criteria for alcohol abuse.	Results from studies 1 & 2 indicate the representativeness heuristic explains how diagnoses are reached. In study 3 it is claimed this was due to the past-behavior heuristic. Results help understand problems in psychodiagnosis (e.g. race and gender bias relative to clinician stereotypes).	*Study 1:* 46 *Study 2:* 44 *Study 3:* 38
Kerr, Walker, Warner & McNeill (2004) USA	Experimental Study	157	Decision-making related to conceptualization of client problem, diagnosis, psychopathology assessment	Counseling and Clinical Psychology Trainees	Influence of client sexual orientation upon conceptualisation of client problem, diagnosis, and assessment of overall level of psychopathology.	Random allocation to one of three groups relating to client sexual orientation.	DSM-IV criteria regarding dysthymic disorder.	The Assessment and Diagnostic Inventory.	Participants judged lesbian client's problems likely due to sexuality. (Dysthymic disorder vignette: $F = 13.006$ (2, 153), $p < .000$). However, diagnosis/degree of psychopathology not related to client sexual orientation. Addressing clients' sexuality not always beneficial.	70

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Kim & Ahn (2002) USA	Cross-sectional Survey	Experiment 1 (E1): 21 Experiment 2 (E2): 20 Experiment 4 (E4): 19	Diagnostic decision-making	Clinical Psychologists, (E1,2,4) Clinical Psychology Graduate Students (E1, 4), and Clinical Psychology Interns (E2,4)	Clinicians' use of causal theories of disorders in clinical reasoning.	-	Whether clinicians give equal credence to all DSM-IV symptoms presented. Or, contrary to this, is there evidence that clinicians display a causal status effect in their symptom classification reasoning?	Familiarity-rating tasks, disorder-defining tasks, theory-drawing tasks, conceptual centrality tasks, hypothetical patient diagnosis tasks, everyday categories theory-drawing tasks, free-recall tasks.	Hypothetical clients who had causally central symptoms rather than causally peripheral had increased probability of receiving mental health diagnosis. (E1: $F(2, 38) = 27.5$, $MSE = 157.90$; $p < .01$; $\eta 2 = .59$; E2: $F(2, 36) = 15.66$, $MSE = 172.84$; $p < .01$; $\eta 2 = .47$; E4: $F(2, 34) = 6.74$, $MSE = 269.97$; $p < .01$;$\eta 2 = .28$). Despite decades of atheoretical DSM guidelines clinicans make diagnoses by forming causal theories.	E1,E2,E4 : 56
Kirk, Wakefield, Hsieh & Pottick (1999) USA	Cross-sectional Survey	250	Diagnostic decision-making	MSW students	Presence of ideological biases in social work assessment relating to conduct disorder.	-	Descriptions of youths included in all 9 variants of case vignettes met DSM-IV criteria for conduct disorder.	Respondents' judgments indicated by response to item on Likert scale - ''strongly agree'' (scored 1) to ''strongly disagree'' (scored 6).	Overall p's correctly differentiated between disordered/non-disordered youth when contextual info in vignettes. V1: $F(2, 246) = 6.81$; p , .01; V2: $F(2, 246) = 86.10$; $p < .01$; V3: $F(2, 244) = 105.19$; $p < .01$. Social workers not ideologically biased.	69

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Loring & Powell (1988) USA	Experimental Study	290	Diagnostic decision-making	Psychiatrists	The effects of client and psychiatrist gender and race on diagnostic judgment.	Two case studies, client sex & race constant. One fifth either: white male/black male/white female/black female/client race not disclosed.	DSM-III-guided diagnosis given to actual clients who feature in vignettes according to their psychiatrist. Inter-rater reliability also revealed a modal response amongst participants that was in agreement with original diagnosis for both vignettes.	Questions relating to DSM-III (Axis 1 and 2) disorders.	Undifferentiated Schizophrenic Disorder vs. Other: client sex (male) OR = -.639 (.250), $p < .05$; client race (black) OR = -.431 (.200), $p < .05$). Similarity of client & psychiatrist: sex (male): OR = -1.304 (.295), $p < .01$; race (black): OR = -.471 (.208), $p < .01$. Client sex and race known = incorrect diagnosis. Client sex and race the same = correct diagnosis.	68
Mendel, Traut-Mattausch, Jonas, Leucht, Kane, Maino, Kissling & Hamann (2011) Germany	Cross-sectional Survey	150	Diagnostic decision-making	Psychiatrists (n=75) Medical Students (n=75)	The influence of confirmation-bias relative to diagnostic decision-making	-	Pretest including six experts on dementia/depression in order to confirm case vignette content compatible with a diagnosis of Alzheimer's disease according to ICD-10 criteria.	Preliminary diagnosis of either 'Alzheimer's disease ' or 'severe depressive episode'. Information before reaching final diagnosis.	Confirmation-bias present in 13% of psychiatrists and 25% of students. Poorer diagnostic accuracy when confirmation-bias is present in information search (OR 7.3, 95% CI 2.53–21.22, p<0.001 ; OR 3.2, 95% CI 1.23–8.56, p=0.02). Psychiatrists should be instructed in techniques to reduce bias.	67

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Mikton & Grounds (2007) United Kingdom	Experimental Study	220	Diagnostic decision-making	Forensic Psychiatrists	Cross-cultural clinical judgment bias in diagnosis of PD.	Each participant randomly allocated to one of two conditions; Condition 1: vignette 1, African Caribbean, vignette 2 Caucasian. Condition 2: vignette 1, Caucasian, vignette 2, African Carribean.	Vignettes included features at the threshold of meeting DSM-IV criteria for either BPD or ASPD.	Clinicians asked to indicate what individual diagnosis according to DSM-IV were probably present in the vignette.	Caucasians 2.8 times more likely to receive a PD than African Caribbean's (OR 2.8, 95% CI 1.6–5.0, p<0.001). Also, variation in PD diagnosis according to clinician ethnicity. (OR 2.2 , 95% CI 1.1–4.6, p<0.04). No cross-cultural bias present with BPD diagnosis. Forensic Psychiatrists underdiagnose ASPD in African-Caribbean men. Results have implications for race equality & policy issues in mental health.	68
Payne (2012) USA	Experimental Study	239	Diagnostic decision-making	Licensed clinical social workers, licensed marriage & family therapists	Influence of race and symptom expression on diagnostic judgments.	Random assignment to view 1 of 4 videos.	DSM-IV-TR criteria for MDD	Computer-based questionnaire asking what general class of DSM-IV-TR disorder on Axis I would client's problems fall under.	Clinicians under diagnosed MDD more often when clients of either race displayed culturally expressed depression symptoms. (Pearson v^2[df=3] = 44.06, p = .001, Fisher's exact = 0.000)	91

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Pottick, Kirk, Hsieh & Tian (2007) USA	Experimental Study	1401	Diagnostic decision-making	Psychologists, Psychiatrists & Social Workers	Clinician's perceptions of mental disorder based upon clinician, client and contextual characteristics.	Race/ethnicity and context of problem behaviors manipulated in case vignettes.	Vignettes included problematic behaviours meeting the DSM-IV criteria for conduct disorder. Additionally, contextual information suggesting either disorder or nondisorder also presented according to DSM–IV guidelines.	Respondent's judgment about whether the adolescent described in vignette has mental disorder: "According to my own view, this youth has a mental/ psychiatric disorder." Scored on a 6-point Likert scale ranging from 1 (strongly disagree) to 6 (strongly agree).	P's decisions relating to presence or not of mental disorder mainly reliant on contextual info ($\chi 2$ (13, N = 1,401) = 518.04, $p < .001$). Associations also found re. race of young person in vignette (e.g. OR = 0.59 for Black vs. White, $p = .002$; OR = 0.60 for Hispanic vs. White, $p = .003$), clinician occupation, (e.g. psychologists, compared with psychiatrists, OR = 4.62/2.24 = 2.06, 95% CI = 1.38, 3.07, p = .001), theoretical orientation, OR = 0.64, p = .04), & age (OR = 0.80, $p = .03$). Professional ID might influence judgments.	86
Pottick, Tian, Kirk & Hsieh (2017) USA	Experimental Study	1540	Treatment related decision-making	Psychologists, Psychiatrists & Social Workers	Impact of social context and ethnicity upon clinicians' judgments relating to treatment effectiveness.	Contextual information (disorder or non-disorder) and ethnicity (White, Black or Hispanic youth) manipulated.	DSM-IV inclusion and exclusion criteria for mental disorder. Research literature relating to effective and ineffective treatments for youths with conduct disorder symptomology.	Respondent's judgment about effectiveness of 14 intervention approaches often used to treat antisocially behaving youth.	13/14 treatments sig associations context of behavior/diff in effectiveness judgments. White youths in internal dysfunction context judged as gaining most benefit from interventions. Context may trigger implicit racial assumptions.	82

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Spengler & Strohmer (1994) USA	Experimental Study	119	Decision-making related to diagnosis and treatment	Counseling Psychologists	Level of cognitive complexity linked to unambiguous clinical bias.	Random assignment to one of two IQ conditions.	DSM-III-R criteria for schizophrenia diagnosis.	Bieri et al's (1966) repertory grid technique	Counselor cognitive complexity moderated the diagnostic overshadowing bias. (Interaction between IQ condition and counselor complexity on aggregate clinical judgement scores, $F(1, 113) = 4.72$, R^2 change = .04, $p = .032$.) The authors conclude that their results have implications for clinical judgment research and counselor education and practice.	82
Stewart (2004) USA	Experimental Study	308	Judgements about the client. Decisions related to treatment prognosis.	Counseling Psychologists & Doctorate & Masters level Counselors	Influence of the representativeness heuristic relative to likely prognosis in counseling.	Random allocation to vignette describing one of four birth positions (i.e., first, middle, youngest, or only child).	Study author cites a lack of empirical evidence for any substantial relationships between birth order and psychological variables. Therefore, judgment accuracy measured according to accepting/rejecting of study hypotheses.	PBOI	Different impressions developed about vignette client and their family experiences that corresponded with the prototypical descriptions of individuals from 1 of 4 birth orders ($F(9, 873) = 8.83$, $p < .0001$). Prognostic ratings also differed according to client birth order ($F(3, 206) = 10.29$, $p < .0001$, $\eta 2 = .13$). Birth-order effects can influence professional judgments.	77

Authors (year) location of study	Study design	Sample size	Context	Clinician group	Cognitive process explored	Experimental manipulation	Accuracy index	Decision/judgment Measure	Key findings/ comments/conclusion	Critical appraisal score (%)
Trierweiler, Muroff, Jackson, Neighbors & Munday (2005) USA	Quasi-experimental Study	11	Diagnostic decision-making	Psychiatrists (3rd- and 4th-year psychiatric residents)	The usage of situational information in diagnostic decision-making.	Ethnic origin of patient and clinician manipulated (African American or Non-African American)	Research relating to impact of negative situational factors upon risk of major depression.	Questionnaire for clinicians exploring diagnostic decision-making process. Questionnaire consisted of nine open-ended questions.	Situational information employed more by African American than non-African American clinicians. Diagnostic standard differs according to clinician race. E.g., non-African American clinicians associated situation variables stability or change in psychiatric condition (OR = 3.70, $p < .05$) and aggressive behavior directed toward the self (OR = 4.60, $p < .01$) with mood disorder. African American clinicians did not.	61

Notes: Data extracted from Kim and Ahn (2002) relates to 3 major experiments in a study including 5 experiments in total. Study also included 2 minor studies not relevant to current review. Trierweiler et al. (2005) included patients as well as clinicians as participants. Patients were 292 adult inpatients at 2 hospitals. Non-African clinicians completed 144 interviews with patients and African American clinicians completed 148 interviews. AAS = American Association of Suicidology (2019), AB = Actuarily based, ADHD = Attention deficit hyperactivity disorder, APD = Antisocial Personality Disorder, BPD = Borderline Personality Disorder, CP = Consensus Predictions, DSM III = Diagnostic and Statistical Manual of Mental Disorders, third edition (American Psychiatric Association, 1980), DSM-III-R = Diagnostic and Statistical Manual of Mental Disorders, third edition, revised (American Psychiatric Association, 1987), DSM IV = Diagnostic and Statistical Manual of Mental Disorders, fourth edition (American Psychiatric Association, 1994), DSM-IV-TR = Diagnostic and Statistical Manual of Mental Disorders, text revision (American Psychiatric Association, 2000), DSM V = Diagnostic and Statistical Manual of Mental Disorders, fifth edition (American Psychiatric Association, 2013), GAD = Generalised anxiety disorder, HPD = Histrionic Personality Disorder, ICD-10 = International Statistical Classification of Diseases and Related Health Problems, tenth version (World Health Organization, 1994), MDD = Major Depressive Disorder, MSE = Mean square of the error, MSW = Master of Social Work, OR = Odds ratio, PBOI = White-Campbell Psychological Birth Order Inventory (Campbell, White, & Stewart, 1991), Personality Disorder = PD, V= Vignette.

Quality Assessment Results

Details relating to the quality assessment can be found in Appendix B. Table 4 presents mean quality percentage scores. The specific quality of the studies varied, aims were clearly described in all except for Garb (1996). Here three separate studies were reported but the aims of only the second were discussed. Main outcomes were clearly described in all studies. Authors typically clearly described their main findings (n=24; 100 %). Many studies used unstandardized outcome measures (N=20; 83.3%). Where this occurred, it was apparent measures were based upon DSM diagnostic criteria and/or theory relative to clinical judgment and decision-making. Estimates of random variability were presented in the main outcome data of N=20 (83.3%) studies, but eleven studies failed to report actual probability values. Samples were only well defined in half of the included studies (N=14; 50%).

In studies that employed an experimental study design (N=14; 50%) all participants were blind to condition. In those studies where participants were allocated to different groups and then compared, all but N=7 studies provided a list or partial list of principal confounders. When randomisation was viable, all but two studies (Bruchmüller & Meyer, 2009; Garb, 1996) ensured they employed random allocation. Randomised condition assignment was concealed from participants in all experimental study designs where randomization occurred. Two studies (Garb, 1996; Kim & Ahn, 2002) included a follow-up.

The statistical tests used to assess the main outcomes were appropriate in all N=28 (100%) studies. Each paper was clear where any analysis was unplanned. All studies took some consideration of confounders (where identified) in their analyses (e.g., clinician gender or profession). Four (14.3%) studies reported sample size calculation analysis.

Table 4.

N, Means % and (SDs) in each of the Downs & Black (1998) quality categories

Rating	**N**	**Mean %**	**(SD[1])**
Excellent	0	0	0
Good/Excellent	2	91	0
Good	8	77.9	4.78
Fair/Good	3	69.3	0.47
Fair	12	61.5	4.29
Poor/Fair	0	0	0
Poor	3	42.7	3.4

[1] (*SD) = standard deviation*

Study Design and Outcomes

The majority of studies (N=23; 82.1%) employed a case vignette study design. Key outcomes are summarised in Table 3. Six differing contexts were identified in which MHPs make clinical decisions. Figure 3 shows the number of papers relative to each context. By far the most common was diagnostic decision-making (N=20; 83.3%). Despite overlaps occurring in a small number of studies (e.g. where studies included multiple variables) variables highlighted as influencing clinical decision-making were grouped into twelve categories (Table 5 and described below).

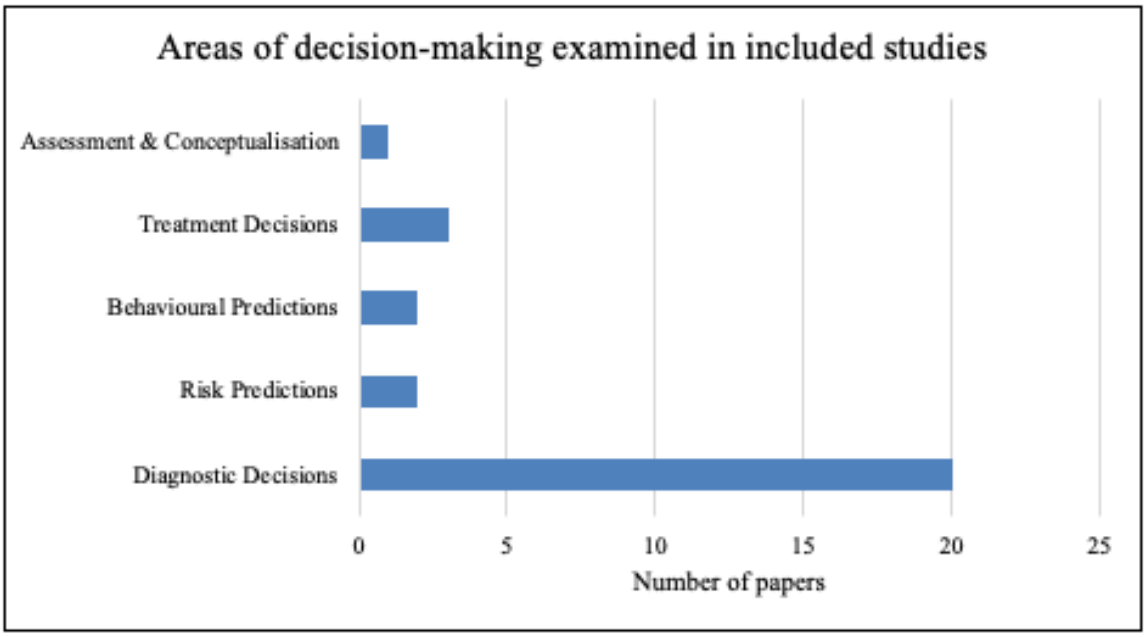


Figure 3. The number of papers relative to each context

Clinician thinking style/cognitive ability. Two studies investigated cognitive processing and its impact on clinical judgement. Aarts et al. (2012) examined the influence of different thinking styles on the accuracy of diagnostic judgements. Participants were asked to answer questions relating to their diagnostic decision-making. Accuracy was measured according to the DSM IV-casebook. A more rational thinking style was associated with poorer diagnostic classification accuracy. This study was rated as fair/good quality in comparison to other papers in this review, but the sample size was low compromising the representativeness and accuracy of the findings. Also, non-parametric analysis was employed meaning effect size estimates are likely to have been adversely affected by departures from normality and heterogeneity of variances.

Spengler & Strohmer (1994) explored whether levels of cognitive complexity might be linked to unambiguous clinical bias. Cognitive complexity is an individual difference defined as "the capacity to construe social behaviour in a multidimensional way" (Bieri et al., 1966). This study found that individual differences in cognitive complexity moderated the likelihood participants' clinical judgements would be impacted by the diagnostic overshadowing bias. Findings revealed psychologists demonstrating lower cognitive complexity were less inclined to

diagnose and treat a fictional client with a learning disability compared to when the client's intellectual ability was manipulated to be in the average range. This paper was rated as 'good', one of fourteen experimental studies in the review, and based within the context of diagnostic decision-making.

Patient and clinician age. One study (Berman et al., 2016) examined the interaction between patient and clinician characteristics. Linear regressions indicated clinician age moderated the relationship between patient's age and ratings of suicide risk. Willingness to hospitalize the patient was also impacted by patient age compared with that of the clinician. The study found clinicians perceive those who are different (i.e., older or younger than themselves) to be at greater risk. The accuracy of clinician's suicide rating was measured by examining the American Association of Suicidology evidence-base relating to difference in suicide rates according to age. This was a good quality paper despite the fact the measure of suicide risk was not an existing measure and had not undergone any extensive validity and/or reliability testing. This study provides some preliminary evidence that clinician demographics and training factors might moderate clinicians' perception of risk.

Prototypicality judgements. Two studies investigated whether the use of prototypic judgements influence the diagnostic process. Despite similar study designs, results could not be synthesised using meta-analytic methods because the outcomes were distinctly different (Appendix D). Blashfield et al. (1985) examined inter-rater reliability in defining a prototype and establishing distinctiveness from other categories to define prototypicality amongst cases. Whereas Evans et al. (2002) explored and evaluated the impact of three existing factors highlighted as key determinants in judging the prototypicality of personality profiles when diagnosing personality disorders (PD; typicality, dominance, number). Evans et al. (2002) found typicality and dominance showed strong effects, whereas no effects were found for number. The authors conclude that cases more prototypic of a specific PD contain highly typical features and those predominantly associated with a diagnostic category. They suggest diagnostic decision-making regarding a specific PD is based on a

prototype-based model rather than the classic category classification system. This was a good quality paper but its lower than average sample size means it is likely to have been underpowered. Despite these limitations, this study provides convincing evidence that clinicians identify factors suggestive as important determinants when diagnosing PD diagnoses.

Blashfield et al. (1985) explored the applicability of the prototype model relative to PD diagnosis by measuring reaction time and inter-clinician agreement data. Differences were explored as a function of profession or experience. As well as examining inter-rater reliability in defining a prototype the distinctiveness from other categories was investigated as a method of defining prototypicality amongst cases. Prototypic cases were discovered for 8 of the 11 PDs. Using reaction time and agreement data produced no differences between professions. Experience had a significant effect on diagnostic speed, but not on agreement. This study was rated as 'fair', as once again a low sample size meant the probability of a type II error increased and therefore there may have been other differences the study failed to identify. Despite these limitations the study provides convincing evidence to suggest how the prototype model might influence diagnostic decision-making relative to the classification of PDs.

Causal assumptions/causal theories. Two papers examined the way clinicians reason about disorders and attribute diagnosis. Kim and Ahn (2002) investigated what they term the 'causal status effect'. This refers to clinicians' use of causal theories of disorders related to their clinical reasoning. Five separate experiments were conducted, three met inclusion criteria for the current review. In these experiments, decision/judgement accuracy was determined by examining whether clinicians give equal credence to all DSM-IV symptoms presented. Or, contrary to this, do clinicians display a causal status effect in their symptom classification reasoning? Employing a complex methodology Kim and Ahn (2002) investigated cause-and-effect relations between symptoms that participants felt were causally connected. They also estimated the strength of the proposed causal relations. Kim and Ahn (2002) showed that when clinicians use the DSM, they do not weigh each criterion equally. All three experiments included in this review received the same

critical appraisal score relating to each question on the critical appraisal tool. Overall this paper was rated as 'fair'. Across all three experiments participant characteristics such as age, gender and theoretical background were not described. These may have acted as confounding variables and impacted the results. Again, the sample size across all experiments was very low and it was not made clear how participants were recruited. This increases the risk of sampling and self-selection bias. Despite these limitations, given the large effects reported where significant results emerged (all $\eta 2 = > 0.14$), Kim and Ahn's (2002) suggestion that clinicians falsely recognise symptoms if they are causally central to their own theories of a disorder merits further investigation.

Bruchmüller & Meyer (2009) examined how clinicians' reason about disorders by exploring the subjective causal assumptions they make related to diagnosis. Clinicians were asked to determine whether they would attribute a diagnosis of bipolar disorder (BD) after reading a case vignette where criteria to diagnose either unipolar or bipolar was manipulated. The study showed that clinicians did not diagnose BD where there is a rational and understandable explanation. This quasi-experimental study was rated as being of fair quality. It had a relatively large sample; power analysis was described but it was unclear whether this had been attained. The theoretical approach and sociodemographic information were obtained and included as part of the analysis and this is a strength to the study design. Limitations include a lack of detail on missing data and on participant characteristics that may have confounded the results. Despite this the study raises some valid points relating to the way in which causal assumptions may impact diagnostic decision-making.

Representativeness. Two papers (Garb, 1996; Stewart, 2004) explored how the representativeness heuristic influences clinical decision-making. Garb (1996) investigated this within a diagnostic context. Stewart (2004) investigated the potential for biases and heuristic thinking in prognostic rating related to client birth order. This study found that once the client was viewed as exemplifying a particular birth order the prognostic ratings differed as a statistically significant medium effect emerged ($\eta 2 = .13$). This experimental study was rated overall as 'good' and had a reasonable sample size. Ironically perhaps, the representativeness of the findings was

unclear, as although participants were randomly drawn the proportion of those asked who agreed was not stated.

Garb (1996) investigated the representativeness heuristic within the context of diagnostic decision-making. This was investigated over three experiments, all documented within a single report. Studies one and two asked participants to read a case history and make likelihood and similarity ratings linked to certain diagnoses. In study two Garb (1996) included a case history taken from a study by Loring and Powell (1988; described later). Garb (1996) concludes that diagnoses are often reached by comparing patients to typical patients with a certain diagnosis and therefore demonstrating the representative heuristic. In study three the past-behaviour heuristic was also investigated. Results suggest clinicians use past behaviour as the best predictor of future behaviour. This paper was found to have the lowest quality in the review as all three experiments were rated 'poor'. The experiments conducted are based upon extensive heuristics/biases literature. Garb (1996) claims his study is the first to empirically explore the potential impact of the representativeness heuristic on clinical judgement. Therefore, despite poor design and methodology, his findings are worth considering.

Gender: Two studies examined the extent gender bias influences diagnostic decision-making. Bruchmüller et al. (2012) investigated whether diagnosis of ADHD was influenced by patient's gender. Ford and Widiger, (1989) explored the prevalence of the histrionic and antisocial PDs amongst men and women. Despite both studies measuring a similar outcome a summary effect could not be calculated using meta-analytic methods. This was because the results from Ford and Widiger (1989) could not be transformed into a correlation coefficient (r). In this study results revealed $\chi 2$ tests with two degrees of freedom, whereas effect size conversion is only correct for $\chi 2$ tests with one (Rosenthal & DiMatteo, 2001, p. 71.)

Bruchmüller et al. (2012) found that over diagnosis of ADHD occurs and is influenced by patient's gender (OR = 2.66). Also, boys were significantly more likely to be incorrectly diagnosed with ADHD than incorrectly given another diagnosis ($\chi 2 = 7.12$). This paper was one of two

receiving the highest overall rating in the review (91%) as it was rated good/excellent. Both vignettes included in the study were constructed on the basis of the DSM–IV and ICD–10 criteria of ADHD thus improving their ecological validity. An experimental study design was also employed. Ford and Widiger (1989) showed that when case history information is varied in the ambiguity of the diagnoses sex biases were seen for diagnoses but not for individual diagnostic criteria. This paper received an overall rating of 'fair'. Despite a relatively good sample size it may not have been representative. It was unclear whether participants in assigned conditions were recruited over the same period of time or if the study was sufficiently powered. Taken together however both papers provide some support that gender influences diagnosis.

Information Order. One paper (Cwik & Margraf, 2017) investigated whether the order of diagnosis-relevant information can predict diagnostic errors. Employing a between-subjects experimental design with random assignment authors found that order of symptom descriptions significantly predicted the correctness of diagnostic decisions. More fully correct diagnostic decisions occurred (producing a medium effect in both GAD and PD case vignettes) where diagnostic information was accessible at the end. The authors suggest this indicates a recency effect. This paper was rated as 'good', one of fourteen experimental studies in the review and based within the context of diagnostic decision-making.

Contextual information. Four studies examined whether contextual information about patients' clinical presentations affected clinicians' judgments of conduct disorder diagnosis or not (De Los Reyes & Marsh, 2011; Kirk et al., 1999; Pottick et al., 2007; Pottick et al., 2017). All four were rated 'fair' or 'good' and reported statistically significant results.

Meta-analysis was only possible that included the studies by De Los Reyes & Marsh (2011) and Pottick et al. (2007). Results for these studies are displayed in Figure 4. This is because the available information in the papers by Kirk et al. (1999) and Pottick et al. (2017) was insufficient to confidently carry out a quantitative synthesis using meta-analysis. Also, the study by Pottick et al.

(2017) was interested in treatment rather than diagnosis and was derived from the sample reported by Pottick et al. (2007).

The pooled effect was statistically significant; r = 0.41 (95% CI 0.37, 0.45), p = < 0.0001. There was no significant evidence of heterogeneity; Q =0.78, df = 1, *p* = 0.376; I2 = 0.0%. The rank correlation test (Kendall's tau = -1.0000, *p* = 1.00) for funnel plot asymmetry was not significant indicating no evidence of likely publication bias. The regression test for funnel plot asymmetry could not be calculated. According to the fail-safe N, 143 null studies would be needed to overturn this meta-analytic result.

Further to meta-analytic results Kirk et al. (1999) found significant differences after manipulating case vignettes to suggest either internal dysfunction (i.e., disorder) or a normal response to a difficult environment (i.e., nondisorder) as the cause of antisocial behaviour in antisocial youths. Here however the authors conclude that social workers are not ideologically biased. This claim is grounded in the finding that participants correctly distinguished between disordered and nondisordered youth based on the contextual information presented.

Pottick et al. (2007) reported additional findings to those included in the meta-analyses. They found that as well as contextual information client race influenced participant's decisions relating to presence or not of mental disorder. Pottick et al., (2017) found that for 13 of 14 treatments there were significant associations between the context of the behaviour and differences in effectiveness judgments. White youths in the internal dysfunction context gained most benefit from interventions than black or Hispanic youths.

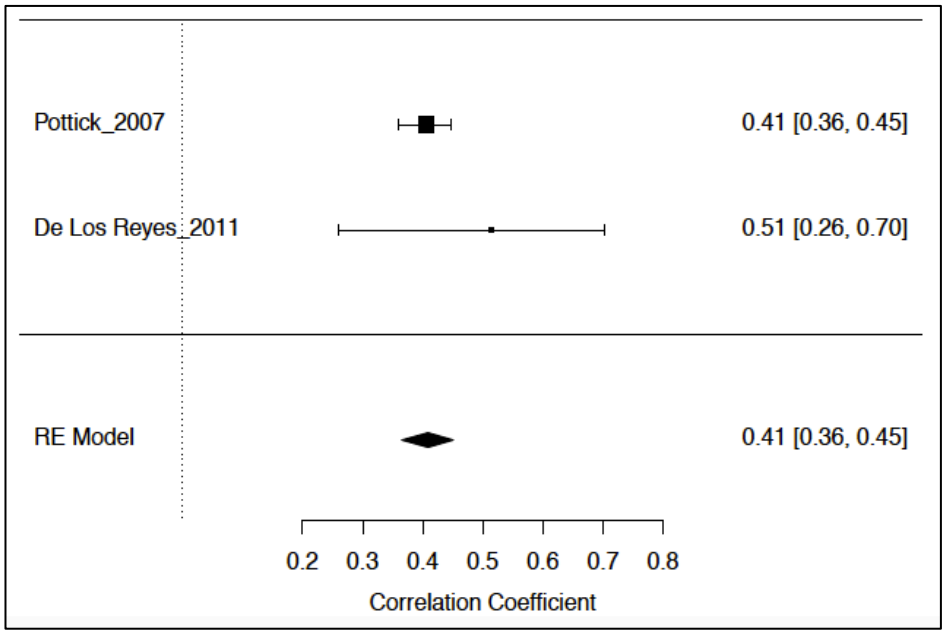


Figure 4. Random effects meta-analysis forest plot: correlations between contextual information and diagnostic decision-making.

Consensus judgements. One study (Fuller & Cowan., 1999) examined consensus judgments relating to the prediction of patient-related risks. Its accuracy of these was found to be similar to actuarially based studies across comparable time frames. Receiver operating characteristics (ROC) analysis suggested the team's aggregate risk score predicted non-specific crisis events at an intermediate level of accuracy (0.71). This finding could be likened to "the wisdom of the crowds" phenomenon, something Tetlock (2016) writes about in his book, 'Superforecasting'. This suggests that individual judgments are error-prone but combining the judgments of several judges considerably improves accuracy. This paper was rated, 'fair'. One of two studies in the current review based within the context of risk-predictions and following a cross-sectional survey design.

Client sexual orientation. One study (Kerr et al., 2004) explored the influence of client sexual orientation on decision-making related to conceptualization of client problem, diagnosis, and psychopathology assessment. Using an experimental design a significant effect was found relating to conceptualization of dysthymic disorder. Diagnosis and degree of psychopathology was not related to client sexual orientation. This study was of fair/good quality (70%) but did not identify

the source population or demonstrate that the distribution of the main confounding factors was the same in the study sample and the source population. This raises questions as to the representativeness of the sample. Despite these limitations the study provides preliminary evidence regarding the influence of sexual orientation in MHP decision-making.

Race/culture. Four studies examined the impact of race upon diagnostic judgments (Loring and Powell, 1988, Mikton and Grounds, 2007, Payne, 2012, Trierweiler et al., 2005). Papers were rated fair/good except Payne (2012) who received the joint highest rating of good/excellent. All studies reported significant results. For all but Payne (2012) findings indicate clinicians were more likely to incorrectly diagnose when they learnt the client was black. Furthermore, non-white clinicians tended to incorrectly diagnose according to client race similarly to that of their white colleague's. Meta-analysis results for these studies are displayed in Figure 5, excluding Payne (2012) and Trierweiler et al. (2005). This was because rather than examine the impact of race upon diagnosis Payne (2012) examined culturally expressed depression symptoms and how clinicians perceive these. Similarly, Trierweiler et al. (2005) investigated situational attributions and how these are interpreted by clinicians in terms of the diagnostic judgments they make. The pooled effect was statistically significant, denoting a medium correlation between client race and diagnosis; $r = 0.34$ (0.21, 0.47), $p = < 0.0001$. There was no significant evidence of heterogeneity; Q =2.78, df = 1, $p = 0.095$; I2 = 64.1% (0.0%; 91.8%). The rank correlation test (Kendall's tau = -1.0000, $p = 1.00$) for funnel plot asymmetry was not significant indicating no evidence of likely publication bias. The regression test for funnel plot asymmetry could not be calculated. According to the fail-safe N, 47 null studies would be needed to overturn this meta-analytic result.

Further to meta-analytic results both Loring and Powell (1988) and Mikton and Grounds (2007) report additional findings. Loring and Powell (1988) also found that despite the presence of diagnostic criteria client sex and psychiatrist sex/race impacts diagnosis. The similarity of client and psychiatrist sex/race also impacts this. Similarly, Mikton and Grounds (2007) found a variation in diagnosis according to clinician ethnicity. Of those studies not included in the meta-analysis, Payne

(2012) found clinicians under diagnosed Major Depressive Disorder more often when clients of either race displayed culturally expressed depression symptoms. Trierweiler et al. (2005) reported situational information is employed more by African American than non-African American clinicians. Further to this, diagnostic standards differ according to the race of the clinician.

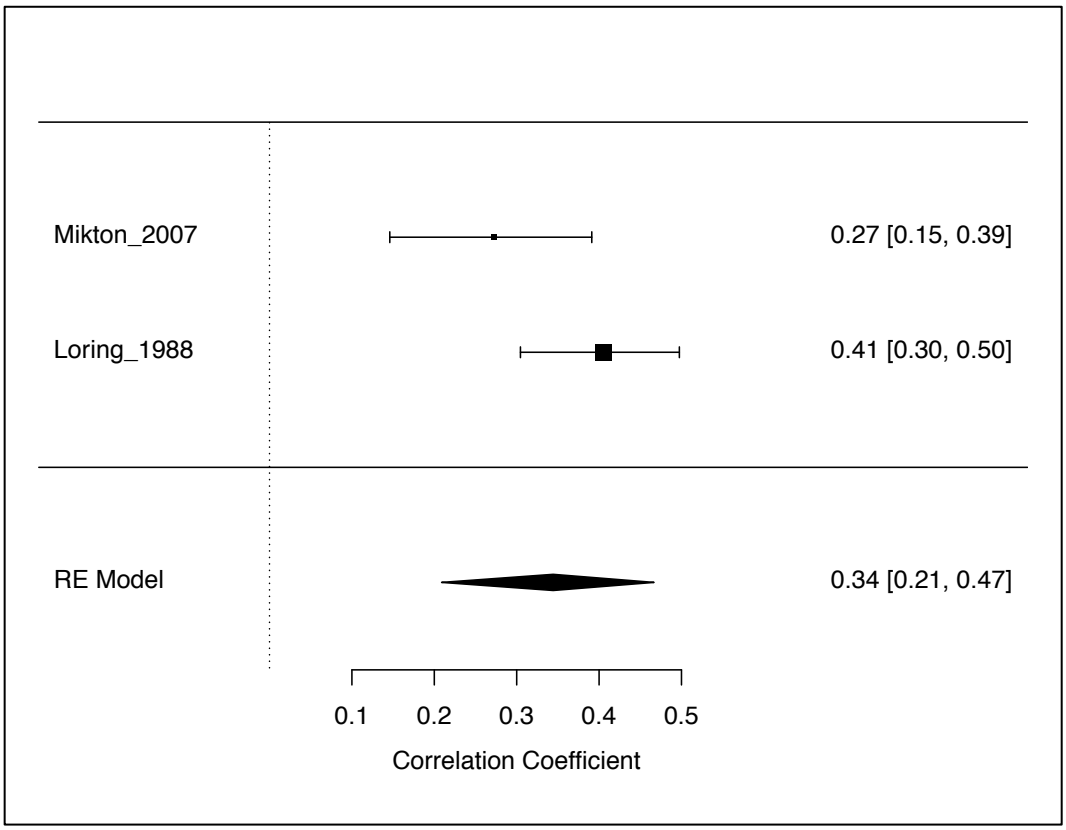


Figure 5. Random effects meta-analysis forest plot: correlations between client race and diagnostic decision-making

Access to information. Two papers examined how access to certain information impacts clinical judgement and decision-making. DeRoma et al. (1997) focussed on decisions related to behavioural prediction. This was amalgamated with Mendel et al. (2011) regarding confirmation bias, as the two concepts seemed analogous. DeRoma et al. (1997) investigated the influence of access to information on evaluative responding relating to a history of physical maltreatment. Randomising three vignette conditions according to presence/absence of behaviour problems, abuse and gender of child, the study found that a history of maltreatment influenced professional judgments. In this study, vignettes with no abuse history or behaviour problems were rated

significantly higher regarding predicted stability and treatment referral. This paper was rated as ‘fair’, one of fourteen experimental studies included in the review, and includes comprehensive research evidence relating to the impact of information about childhood abuse.

Mendel et al (2011) examined the degree to which psychiatrists and medical students are susceptible to confirmation bias when seeking new information after reaching an initial diagnostic decision. They were also interested to discover whether confirmation bias in the information search adversely effects the value of the diagnostic decision and consequent treatment recommendations. Authors found that confirmation-bias was present in 13% of psychiatrists and 25% of student’s information search. Furthermore, poorer diagnostic accuracy was observed when confirmation-bias was present. This paper was ‘fair’ in quality compared to other papers in the review. However, questions remain as to the external validity of the results as the participants were not representative of the entire source population as random sampling did not occur.

Table 5. *Summary of key findings indicating variables that have been found to influence the accuracy of clinical decision-making*

Studies	**Variables**											
	Clinician thinking-style/cognitive ability	Patient and clinician age	Prototypicality judgements	Causal assumptions/theories	Representativeness	Gender	Information Order	Contextual information	Consensus judgements	Client sexual orientation	Race/culture	Access to information
Aarts et al. (2012)	■											
Berman et al. (2016)		■										
Blashfield, et al (1985)			■									
Bruchmüller & Meyer (2009)				■								
Bruchmüller et al. (2012)						■						
Cwik & Margraf (2017)							■					
De Los Reyes & Marsh (2011)								■				
DeRoma et al. (1997)												■
Evans et al. (2002)			■									
Ford & Widiger (1989)						■						
Fuller & Cowan (1999)									■			
Garb (1996) Study 1					■							
Garb (1996) Study 2					■							
Garb (1996) Study 3					■							
Kerr et al. (2004)										■		
Kim & Ahn (2002) Exp 1				■								
Kim & Ahn (2002) Exp 2				■								
Kim & Ahn (2002) Exp 4				■								
Kirk et al. (1999)								■				
Loring & Powell (1988)											■	
Mendel et al. (2011)												■
Mikton & Grounds (2007)											■	
Payne (2012)											■	
Pottick et al. (2007)								■				
Pottick et al. (2017)								■				
Spengler & Strohmer (1994)	■											
Stewart (2004)					■							
Trierweiler et al. (2005)											■	
Total number of studies	2	1	2	4	4	2	1	4	1	1	4	2

Summary of results

Narrative synthesis and selective meta-analysis revealed the variables most commonly highlighted as influencing clinical decision-making. These were grouped into twelve categories (Table 5). Clinicians were more likely to make diagnostic decisions and/or judgements that were incongruent with DSM criteria. This occurred when attributing diagnoses (e.g. Aarts et al., 2012; Kim & Ahn, 2002), defining prototypicality amongst cases (e.g. Blashfield et al., 1985) and providing prognostic ratings (Stewart, 2004). Clinicians also provide estimates of risk contrary to that of the evidence-base (e.g. Berman et al., 2016). They conceptualised client problems according to the sexual orientation of the client without any substantial empirical evidence to support this (e.g. Kerr et al., 2004). Furthermore, they predicted client behaviour based on specific information from the client's history rather than the client's current presentation (DeRoma et al., 1997). Taken together these results suggest that clinician's causal judgments, behavioural predictions and treatment decisions are often prone to error.

Discussion

This review aimed to investigate and synthesise research examining the cognitive processes of clinicians, the accuracy of their judgments, and factors that might influence this. Beliefs, attitudes, and subjective norms of clinicians were of particular interest.

Summary of findings

A total of 24 papers (including 28 studies) were examined. Study quality ranged from poor to good/excellent. Most studies employed a case vignette experimental study design. Given that many studies employed convenience sampling techniques response rates and any confounding factors of the source population remained unknown. This increases the chance of self-selection bias. Statistical analysis techniques varied across studies but largely investigated differences in clinical judgement and decision-making by manipulating variables related to the client, the clinician and/or the context. Clinical diagnosis was by far the most common area of decision-making. Several variables were identified across studies as significantly influencing clinical judgement and decision-

making and these were grouped into twelve categories. Those most frequently found to influence the clinician include causal assumptions/causal theories, representativeness, contextual information, and race/culture. These variables were each investigated in 4 of 28 studies. The remaining variables (Table 5) were examined in no more than 1 of 2 of the studies. There was substantial methodological variability across studies as the independent/dependent variables and outcomes of interest varied considerably. Therefore, even where multiple studies investigated the same variables, studies were not directly comparable and so lacked replicated findings. Two variables were examined across multiple studies however reporting sufficient data to enable meta-analysis. These variables were contextual information and race/culture. Findings should be interpreted with caution however given the small number of studies included in each meta-analysis (≤2).

Studies included in the review assert that contextual information highly impacts clinician judgments. Since the pooled correlation between contextual information and clinician judgement was significant (moderate correlation, $r = 0.41$) this assertion is supported. Furthermore, there was no evidence of potential publication bias or significant heterogeneity, and the according to the fail-safe N, 143 null studies would be needed to overturn this meta-analytic result. It is worth noting Kirk et al's. (1999) findings, however, in that although case vignettes described youths meeting DSM-IV criteria for conduct disorder diagnosis social workers correctly avoided mechanically applying DSM-IV diagnostic criteria for this diagnosis. Rather they appropriately took into account environmental context in their diagnostic decision-making. Other clinical or methodological variables may also influence the relationship between contextual information and clinician judgement given the additional findings reported by Pottick et al. (2007).

Meta-analytic results indicated a significant and moderate correlation ($r = 0.34$) between client race and diagnostic judgments. Furthermore, this review found no evidence of potential publication bias, and there was no significant evidence of heterogeneity ($I^2 = 64.1\%$). The failsafe N calculations for the clinician race (47) indicated that numerous studies with null findings would be needed to overturn these results. This meta-analytic result supports the assertion that client race

impacts diagnostic judgements. Clinicians are more likely to incorrectly diagnose when the race of the client is known.

How do findings relate to the wider literature base?

Despite the substantial variability across studies not eligible for meta-analysis the reviewed findings still suggest that, overall, mental health professionals'(MHPs) decisions may be prone to error and are likely to be influenced by a number of dynamic and context variables. These findings support and expand upon previous reviews that report the variability of clinical judgement and decision-making suggesting heuristics/biases are a likely reason why MHPs judgements are often inaccurate (e.g. Garb, 2005).

Several heuristics/biases were cited to explain why certain variables influenced the decisions made by MHPs (Table 6). Spengler & Strohmer (1994) discuss the diagnostic overshadowing bias. They found counsellors with lower cognitive complexity were three times more likely to overshadow when it came to diagnosing and treating clients with a learning disability. Another study showed that clinician age may reveal a "similarity" bias. Clinicians perceive those who are different (i.e., older/younger) to be at greater risk of suicide (Berman et al., 2016). Two papers investigated the influence of the representativeness heuristic (Garb, 1996; Stewart, 2004). Garb (1996) investigated this across two studies. Results indicate the representativeness heuristic might explain how diagnoses are reached. In his third study however, Garb found this may be due to the past-behaviour heuristic. Stewart (2004) also found evidence to support the influence of the representative heuristic relative to likely prognosis in counselling. This was related specifically to birth-order effects and how these can influence professional judgments regarding client personality. Presence of the representativeness heuristic was also found when therapists were diagnosing ADHD (Bruchmüller et al., 2012). Here researchers linked representativeness and gender bias by showing that not only do clinicians diagnose ADHD if a patient resembles their concept of a prototypical ADHD child but overdiagnosis of ADHD also occurs in boys more than girls. Gender bias has also been implicated in the diagnosis of histrionic and antisocial personality disorders (HPD and HPD;

Ford & Widiger, 1989). The authors found that when case history information is less ambiguous males are significantly less likely to be diagnosed HPD than females. Similarly, antisocial female patients significantly more likely to be diagnosed with HPD than with APD. These results suggest client gender might impact diagnostic decision-making.

Evidence suggests cross-cultural clinical judgment bias may influence diagnostic judgement and decision-making. Regarding the diagnosis of personality disorder (PD), Caucasians were 2.8 times more likely to receive a PD diagnosis than African Caribbean's and there was also variation in diagnosis according to clinician ethnicity (Mikton & Grounds, 2007). Effects of client and psychiatrist race were found in diagnostic judgment related to undifferentiated schizophrenic disorder (Loring & Powell, 1988). Similar to their white colleagues black clinicians evaluated the white case studies as having either undifferentiated schizophrenia or a less severe disorder rather than a paranoid schizophrenic disorder diagnosis, more commonly attributed to black males.

Taken together, these findings suggest there has been some progress in the study of heuristics/biases and how this specifically relates to the research concerning clinical judgment and decision-making.

Table 6.

Heuristics and biases described in review

Heuristic/Bias Identified	**Definition**
Cross-cultural clinical judgment bias	The differential patterns of decision-making, largely in relation to mental disorder diagnosis, based upon client race. For example, overdiagnosis of black people (or underdiagnosis of white people) in such categories as schizophrenia and underdiagnosis of black people (or overdiagnosis of white people) in other categories such as personality disorders (Loring & Powell, 1988).
Gender bias	Differential treatment and/or representation of males and females based on stereotypes and not on real differences.

Overshadowing bias	The unwillingness of mental health professionals to recognise mental health problems in people with intellectual disabilities, and the propensity to assume they are essentially part of the intellectual disability itself (Reiss, Levitan, & Szysko, 1982)
Past-behaviour heuristic	Making predictions of future behaviour based upon past behaviour (Garb, 1996).
Representativeness heuristic	Descriptive of a person's cognitive processes when making a judgment about an object or person by comparing that to another object or person (Tversky & Kahneman, 1973)
Similarity bias	When a person makes a judgement about another person who they perceive as being like them based on specific traits (e.g. age, gender, geographical location).

Critique

Findings in this review must be interpreted alongside three main limitations. The studies included were taken from peer-reviewed articles published in English, possibly not reflecting all the available literature in terms of the accuracy of MHPs decisions. The results may well be biased towards more favourable conclusions increasing the likelihood of publication bias (Sterne, Gavaghan, & Egger, 2000). The decision not to search the grey literature was because studies included are likely to be of higher quality. Furthermore, several studies (N = 25) were identified through reverse and forwards citation, which might imply an improved set of search terms and a wider set of databases is required.

A second limitation is that only 5 of the 24 papers were included in the meta-analyses undertaken in the review because there was too much variance regarding the independent and dependent variables under investigation in the remaining studies. Interpretation of the included research was undertaken primarily by the author so still remains vulnerable to some subjectivity.

A third limitation refers to the generalizability of the findings as many of the selected studies recruited psychologists and psychotherapists as participants. In the United Kingdom diagnosis is largely the responsibility of a General Medical Council (GMC) registered Psychiatrist. This raises questions as to the ecological validity of the findings and might also influence the accuracy of diagnosis in the included studies.

Further limitations include that samples were small and often poorly defined. Studies were also at risk of self-selection bias given many employed opportunity and snowball sampling methods. Despite this samples often included a variety of MHPs increasing generalizability. Rather than using measures validated from previous studies many were grounded in retrospective or prospective self-report. This could mean participants were potentially at risk of responding in what they perceived as the most socially desirable way. It is also of note that this review only included 9 studies published in the last 10 years.

The decision to take a systematic review approach was considered against whether a scoping review might have been a more appropriate format. In their guidance Munn et al. (2018) suggest that a scoping review should be considered to clarify available evidence and key concepts/definitions and to identify and analyse knowledge gaps in the literature. Garb (2005) has already clearly set out key concepts/definitions when providing a map of the evidence regarding the validity of clinical judgement and decision-making. This includes his appraisal of the literature examining the validity of descriptions of personality and psychopathology, the cognitive processes of clinicians, the validity of clinical judgments and the utility of treatment decisions. Garb (2005) also suggests that progress made in studying heuristics and biases is likely to inform research on clinical judgment and therefore further research is needed. One reason why undertaking a further scoping review might have been considered is that Garb's (2005) review was conducted fourteen years ago, and the search strategy was not developed using best practice review guidelines.

Munn et al. (2018) stress that the most important consideration when deciding between a systematic review and a scoping review approach is whether the results of the review answer a clinically meaningful question or provide evidence to inform practice. The present review was

interested in studies that measured decision/judgment accuracy, undoubtedly a clinically meaningful endeavour and one that potentially provides evidence to inform practice. Therefore, on balance taking a systematic review approach was deemed more appropriate in the present study than that of a scoping review approach.

Clinical Implications

Diagnostic and treatment decisions should be based on the evidence-base but current research suggests this is often not the case. Findings suggests diagnostic and treatment related decisions may be at particular risk of heuristics/biases and may well be inaccurate. Normal practice should include asking clinicians in clinical supervision how they reached decisions. Taping sessions to identify and validate decisions made could also be beneficial. This review emphasises the value in using algorithmic/mechanical/statistical methods to aid decisions and reduce the risk of bias.

Results also suggest that improved understanding and a greater awareness of the cognitive processes related to clinical decision-making is required. As Garb (1996) points out, results relating to diagnostic decision-making have serious consequences on mental disorder diagnosis. Despite the reliability of diagnostic decisions being relatively fair (Grove, 1987; Matarazzo, 1983), this will be much lower if clinicians attend to prototypes and this varies from clinician to clinician (e.g. Blashfield & Haymaker, 1988; Livesley et al., 1987; McFall et al., 1991). Results also suggest that when biases occur in diagnosis this happens because clinicians are attending to stereotypes rather than base rates (e.g., Ford & Widiger, 1989; Loring & Powell, 1988). This review found evidence to suggest that racial bias is especially prevalent in diagnostic decision-making. The Royal College of Psychiatrists (RCPsych) also acknowledge unconscious racial bias exists within psychiatry (RCPsych, 2018)

Future Research

Clinical judgement and decision-making should be assessed in more focussed samples. More follow up studies examining clinical decision-making are also required. Valid measures and experimental designs that intervene (and train better) decision making amongst clinicians are also

required. More research focussed upon treatment and assessment decisions, risk and behavioural predictions would be beneficial. Future research might look to assess the influence of bias on decisions regarding treatment allocation and progression. This review confirms that the case vignette method is the most commonly used approach. Hyler, Williams, and Spitzer (1982) suggest that a live interview approach could be more valid and reliable, however as they allow "more complete information" for the clinician to make use of. Using an innovative ‘real time’ scenario-based approach could be one way of overcoming this issue and improving the ecological validity of such studies. Strengths to this approach include tighter control of the variables of interest and also allowing a substantial number of individuals nationally to evaluate the same case (Loring & Powell, 1988).

Conclusions

This review suggests that MHPs make clinical judgements and decisions that are often prone to inaccuracies and may be to the detriment of patient care. Heuristics and biases are one potential cause for this. Variables likely to influence decisions include but are not limited to causal assumptions/causal theories, representativeness, contextual information and race/culture. Inaccuracy occurs within several different clinical contexts where decisions are made directly relating to patient care. This includes diagnostic decision-making, risk predictions, behavioural predictions, treatment decisions, psychopathology assessment and conceptualisation of client problems. Future innovative research addressing methodological flaws in previous research as well as looking to replicate and build upon previous findings would be useful.

References

Aarts, A. A., Witteman, C. L., Souren, P. M., & Egger, J. I. (2012). Associations between psychologists' thinking styles and accuracy on a diagnostic classification task. *Synthese, 189*, 119-130. https://doi.org/10.1007/s11229-012-0081-3

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin, 111*, 256.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49*, 323. http://dx.doi.org/10.1037/0022-006X.49.3.323

Berman, N. C., Tung, E. S., Matheny, N., Cohen, I. G., & Wilhelm, S. (2016). Clinical decision making regarding suicide risk: Effect of patient and clinician age. *Death studies, 40*, 269-274. https://doi.org/10.1080/07481187.2015.1128498

Blashfield, R., Sprock, J., Pinkston, K., & Hodgin, J. (1985). Exemplar prototypes of personality disorder diagnoses. *Comprehensive psychiatry, 26*, 11-21. https://doi.org/10.1016/0010-440X(85)90045-8

Blashfield, R. K., & Haymaker, D. (1988). A prototype analysis of the diagnostic criteria for DSM-III-R personality disorders. *Journal of Personality Disorders*, *2*, 272-280. https://doi.org/10.1521/pedi.1988.2.3.272

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis.* John Wiley & Sons.

Bruchmüller, K., & Meyer, T. D. (2009). Diagnostically irrelevant information can affect the likelihood of a diagnosis of bipolar disorder. *Journal of Affective Disorders, 116*, 148-151. https://doi.org/10.1016/j.jad.2008.11.018

Bruchmüller, K., Margraf, J., & Schneider, S. (2012). Is ADHD diagnosed in accord with diagnostic criteria? Overdiagnosis and influence of client gender on diagnosis. *Journal of consulting and clinical psychology, 80*, 128.

Centre for Reviews and Dissemination. (2009). *CRD's guidance for undertaking reviews in healthcare*. York Publishing Services.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 271. http://dx.doi.org/10.1037/h0027592

Cwik, J. C., & Margraf, J. (2017). Information order effects in clinical psychological diagnoses. *Clinical psychology & psychotherapy, 24*, 1142-1154. https://doi.org/10.1002/cpp.2080

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. Science, 243(4899), 1668-1674.

De Los Reyes, A., & Marsh, J. K. (2011). Patients' contexts and their effects on clinicians' impressions of conduct disorder symptoms. *Journal of Clinical Child & Adolescent Psychology, 40*, 479-485. https://doi.org/10.1080/15374416.2011.563471

DeRoma, V. M., Hansen, D. J., Tishelman, A. C., & D'Amico, P. (1997). Influence of information related to child physical abuse on professional ratings of adjustment and prognosis. *Child abuse & neglect, 21*, 295-308. https://doi.org/10.1016/S0145-2134(96)00155-X

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomized and non-randomized studies of health care interventions. *Journal of Epidemiology and Community Health, 52*, 377-384. https://doi.org/10.1136/jech.52.6.377

Dumont, F., & Lecomte, C. (1987). Inferential processes in clinical work: Inquiry into logical errors that affect diagnostic judgments. *Professional Psychology: Research and Practice, 18*, 433.

Evans, D. L., Herbert, J. D., Nelson-Gray, R. O., & Gaudiano, B. A. (2002). Determinants of diagnostic prototypicality judgments of the personality disorders. *Journal of personality disorders, 16*, 95-106. https://doi.org/10.1521/pedi.16.1.95.22554

Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R.S., Nichols, C.N., Lampropoulos, G.K., Walker, B.S., Cohen, G. & Rush, J.D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341-382. https://doi.org/10.1177/0011000005285875

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology: Research and Practice, 17*, 420-430. http://dx.doi.org/10.1037/0735-7028.17.5.420

Fiedler, K., & von Sydow, M. (2015). Heuristics and biases: Beyond Tversky and Kahneman's (1974) judgment under uncertainty. *Cognitive psychology: Revisiting the classical studies*, 146-161.

Ford, M. R., & Widiger, T. A. (1989). Sex bias in the diagnosis of histrionic and antisocial personality disorders. *Journal of Consulting and Clinical Psychology, 57*, 301. http://dx.doi.org/10.1037/0022-006X.57.2.301

Fuller, J., & Cowan, J. (1999). Risk assessment in a multi-disciplinary forensic setting: Clinical judgement revisited. The Journal of Forensic Psychiatry, 10(2), 276-289. https://doi.org/10.1080/09585189908403681

Garb, H. N. (1996). The representativeness and past-behavior heuristics in clinical judgment. *Professional Psychology: Research and Practice, 27*, 272. http://dx.doi.org/10.1037/0735-7028.27.3.272

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. Psychological bulletin, 105(3), 387. http://dx.doi.org/10.1037/0033-2909.105.3.387

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. American Psychological Association.

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology 2005, 1*, 67-89. https://doi.org/10.1146/annurev.clinpsy.1.102803.143810

Grove, W. M. (1987). The reliability of psychiatric diagnosis. In *Issues in diagnostic research* (99-119). Springer, Boston, MA.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*, 293.http://dx.doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, *12,* 19-30.

Haase, W. (1964). The Role of socio-economic class in examiner bias. In Riesman, F. et al (eds). *The Mental Health of the Poor*. New York, Free Press

Hamilton, W. (2011). Package 'MAVIS' (1st ed.). https://doi.org/10.13140/RG.2.1.3316.7205

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61*, 155-163. https://doi.org/10.1002/jclp.20108

Health and Care Professions Council (2019). *Standards of conduct, performance and ethics.* Retrieved from: https://www.hcpc-uk.org/standards/standards-of-conduct-performance-and-ethics/

Hooper, P., Jutai, J. W., Strong, G., & Russell-Minda, E. (2008). Age-related macular degeneration and low-vision rehabilitation: a systematic review. *Canadian Journal of Ophthalmology, 43*, 180-187.

Hyler, S. E., Williams, J. B., & Spitzer, R. L. (1982). Reliability in the DSM-III field trials: Interview v case summary. *Archives of General Psychiatry, 39*, 1275-1278. http://dx.doi.org/10.1001/archpsyc.1982.04290110035006

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291. https://doi.org/10.2307/1914185.

Kayne, N. T., & Alloy, L. B. (1988). Clinician and patient as aberrant actuaries: Expectation-based distortions in assessment of covariation. In L. Y. Abramson (Ed.). *Social cognition and clinical psychology: A synthesis* (295-365). New York, NY, US: Guilford Press.

Kerr, S. K., Walker, W. R., Warner, D. A., & McNeill, B. W. (2004). Counselor trainees' assessment and diagnosis of lesbian clients with dysthymic disorder. *Journal of Psychology & Human Sexuality, 15*, 11-26. https://doi.org/10.1300/J056v15n02_02

Kim, N. S., & Ahn, W. K. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. Journal of Experimental Psychology: General, 131(4), 451. http://dx.doi.org/10.1037/0096-3445.131.4.451

Kirk, S. A., Wakefield, J. C., Hsieh, D. K., & Pottick, K. J. (1999). Social context and social workers' judgment of mental disorder. *Social Service Review, 73*, 82-104. https://doi.org/10.1086/515798

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155-163.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry, 160,* 1566-1577. https://doi.org/10.1176/appi.ajp.160.9.1566

Larson, J. L., Vos, C. M., & Fernandez, D. (2013). Interventions to Increase Physical Activity in People With COPD. *Annual Review of Nursing Research, Volume 31, 2013: Exercise in Health and Disease*, 297.

Lenhard, W., & Lenhard, A. (2016). *Calculation of effect sizes.* Psychometrica.

Livesley, W. J., Reiffer, L. I., Sheldon, A. E., & West, M. (1987). Prototypicality ratings of DSM-III criteria for personality disorders. Journal of Nervous and Mental Disease, 175(7), 395-401. http://dx.doi.org/10.1097/00005053-198707000-00002

Lopez, S. R. (1989). Patient variable biases in clinical judgment: Conceptual overview and methodological considerations. *Psychological Bulletin, 106*, 184. http://dx.doi.org/10.1037/0033-2909.106.2.184

Loring, M., & Powell, B. (1988). Gender, race, and DSM-III: A study of the objectivity of psychiatric diagnostic behavior. *Journal of health and social behavior*, 1-22. https://doi.org/10.2307/2137177

Matarazzo, J. D. (1983). The reliability of psychiatric and psychological diagnosis. *Clinical Psychology Review, 3*, 103-145. https://doi.org/10.1016/0272-7358(83)90008-9

McFall, M. E., Murburg, M. M., Smith, D. E., & Jensen, C. F. (1991). An analysis of criteria used by VA clinicians to diagnose combat-related PTSD. *Journal of Traumatic Stress, 4*, 123-136. https://doi.org/10.1002/jts.2490040110

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. http://dx.doi.org/10.1037/11281-000

Mendel, R., Traut-Mattausch, E., Jonas, E., Leucht, S., Kane, J. M., Maino, K., ... & Hamann, J. (2011). Confirmation bias: Why psychiatrists stick to wrong preliminary diagnoses. *Psychological Medicine, 41*, 2651-2659. https://doi.org/10.1017/S0033291711000808

Mikton, C., & Grounds, A. (2007). Cross-cultural clinical judgment bias in personality disorder diagnosis by forensic psychiatrists in the UK: A case-vignette study. *Journal of personality disorders, 21*, 400-417. https://doi.org/10.1521/pedi.2007.21.4.400

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine, 6*, 1-6. https://doi.org/10.1371/journal.pmed.1000097

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology*, *18,* 143. https://doi.org/10.1186/s12874-018-0611-x

Myers, David G., (2010). *Social psychology* (Tenth ed.). New York, NY.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157-159. https://doi.org/10.3102/10769986008002157

Payne, J. S. (2012). Influence of race and symptom expression on clinicians' depressive disorder identification in African American men. *Journal of the Society for Social Work and Research*, *3*, 162-177. https://doi.org/10.5243/jsswr.2012.11

Pottick, K. J., Kirk, S. A., Hsieh, D. K., & Tian, X. (2007). Judging mental disorder in youths: Effects of client, clinician, and contextual differences. *Journal of Consulting and Clinical Psychology, 75*, 1. http://dx.doi.org/10.1037/0022-006X.75.1.1

Pottick, K. J., Tian, X., Kirk, S. A., & Hsieh, D. K. (2017). Treating the Child or Syndrome: Does Context Matter for Treatment Decisions for Antisocially Behaving Youth? *Journal of Psychopathology and Behavioral Assessment, 39*, 396-411. https://doi.org/10.1007/s10862-017-9599-5

Quintana, D. S. (2015). From pre-registration to publication: a non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in Psychology, 6*, 1549. https://doi.org/10.3389/fpsyg.2015.01549

Reiss, S., Levitan, G. W., & Szyszko, J. (1982). Emotional disturbance and mental retardation: Diagnostic overshadowing. *American Journal of Mental Deficiency, 86*, 567-574.

Royal College of Psychiatrists (2019). Racism and mental health. Retrieved from: https://www.rcpsych.ac.uk/docs/default-source/improving-care/better-mh-policy/position-statements/ps01_18.pdf?sfvrsn=53b60962_4

Royal College of Psychiatrists (2019). What we do and how. Retrieved from: https://www.rcpsych.ac.uk/about-us/what-we-do-and-how

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178. http://dx.doi.org/10.1037/h0023624

Snyder, M. (1981). Seek, and ye shall find: Testing hypotheses about other people: The Ontario Symposium. In *Social cognition: The Ontario symposium*. Lawrence Erlbaum Associates.

Spengler, P. M., & Strohmer, D. C. (1994). Clinical judgmental biases: The moderating roles of counselor cognitive complexity and counselor client preferences. *Journal of Counseling Psychology, 41*, 8. http://dx.doi.org/10.1037/0022-0167.41.1.8

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, 53*, 1119-1129. https://doi.org/10.1016/S0895-4356(00)00242-0

Stewart, A. E. (2004). Can knowledge of client birth order bias clinical judgment? *Journal of Counseling & Development, 82*, 167-176. https://doi.org/10.1002/j.1556-6678.2004.tb00298.x

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Turk, D. C., & Salovey, P. E. (1988). *Reasoning, inference, and judgment in clinical psychology*. Free Press.

Trierweiler, S. J., Muroff, J. R., Jackson, J. S., Neighbors, H. W., & Munday, C. (2005). Clinician race, situational attributions, and diagnoses of mood versus schizophrenia disorders. *Cultural Diversity and Ethnic Minority Psychology, 11*, 351. http://dx.doi.org/10.1037/1099-9809.11.4.351

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131. http://dx.doi.org/10.1126/science.185.4157.1124

Wedding, D., & Faust, D. (1989). Clinical judgment and decision making in neuropsychology. *Archives of Clinical Neuropsychology, 4*, 233-265. https://doi.org/10.1093/arclin/4.3.233

Wu, K. D., & Clark, L. A. (2003). Relations between personality traits and self-reports of daily behavior. *Journal of Research in Personality, 37*, 231-256. https://doi.org/10.1016/S0092-6566(02)00539-1

Yamamoto, J., James, Q. C., Bloombaum, M., & Hattem, J. (1967). Racial factors in patient selection. *American Journal of Psychiatry, 124*, 630-636. https://doi.org/10.1176/ajp.124.5.630

Appendix A: Full list of search terms used to search papers, abstracts and key-terms

Medline via OvidSP			
Concept	Terms	Search	Exact search term used
Mental Health Professionals	Mental Health Practitioner	1	"mental adj1 health adj1 practitioner*".mp.
		2	"mental adj1 health adj1 professional*".mp
		3	"mental adj1 health adj1 clinician*".mp
	Psychological Therapist	4	"clinical adj1 psycholog*".mp
		5	"psychological therap*".mp.
		6	"psychotherapy/ or psychotherap*".mp
		7	"cognitive therap*".mp.
	Counsellor	8	"university adj1 counsel*".mp
	Psychiatrist	9	psychiatr*.mp.
Decision Making and Clinical Judgement		10	exp "clinical decision-making"/
		11	"clinical decision making".mp.
		12	decision*.mp.
		13	"decision adj1 making".mp.
		14	"decision making"/
		15	"clinical adj1 decision*".mp.
		16	"clinical adj1 judgement*".mp.
		17	heuristics/
		18	heuristics.mp.
		19	"treatment adj1 decision*".mp.
Accuracy		20	accuracy.mp.
		21	accurate.mp.
Combination		22	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9
		23	10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19

	24	20 or 21
	25	22 and 23 and 24

.mp = title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms.

PsycInfo via OvidSP			
Concept	Terms	Search	Exact search term used
Mental Health Professionals	Mental Health Practitioner	1	“mental adj1 health adj1 practitioner*”.mp.
		2	“mental adj1 health adj1 professional*”.mp
		3	“mental adj1 health adj1 clinician*“.mp
	Psychological Therapist	4	“clinical adj1 psycholog*”.mp
		5	“psychological therap*”.mp.
		6	exp psychotherapists/
		7	“psychotherapy/ or psychotherap*”.mp
		8	"cognitive therap*".mp.
	Counsellor	9	“university adj1 counsel*”.mp
	Psychiatrist	10	exp psychiatrists/
		11	psychiatr*.mp.
Decision Making and Clinical Judgement		12	exp “Decision Making/ or exp Clinical Judgment (Not Diagnosis)”/
		13	decision*.mp.
		14	“decision adj1 making“.mp.
		15	“clinical adj1 decision*“.mp.
		16	“clinical adj1 judgement*”.mp.
		17	exp heuristics/
		18	heuristics.mp.
		19	“treatment adj1 decision*”.mp.

Accuracy	20	“accuracy”.mp.
	21	“accurate”.mp.
Combination	22	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11
	23	12 or 13 or 14 or 15 or 16 or 17 or 18 or 19
	24	20 or 21
	25	22 and 23 and 24

.mp = title, abstract, heading word, table of contents, key concepts, original title, tests & measures, mesh

Scopus							
	Mental Health Professionals	AND	*Decision Making and Clinical Judgement*	AND	*Accuracy*	AND NOT	*Database*
OR	"mental health professional*"		"treatment decision"		"accuracy"		Medline
	"mental health practitioner*"		"clinical decision"		"accurate"		PsycInfo
	"mental health clinician*"		"decision"				
	"cognitive therap*"		"clinical judgement"				
	“psychological therap*”		"clinical decision making"				
	"clinical psycholog*"		"decision making"				
	"psychotherap*"		"heuristics"				
	"university counsel*"						
	"psychiatr*"						

** Title, abstract, keywords search used for all search terms.*

Appendix B: Adapted Downs and Black's Critical Appraisal Tool

Appendix B: Adapted Downs and Black's Critical Appraisal Tool

Appendix B: Adapted Downs and Black's Critical Appraisal Tool

Appendix B: Adapted Downs and Black's Critical Appraisal Tool

Appendix B: Adapted Downs and Black's Critical Appraisal Tool

Appendix B: Adapted Downs and Black's Critical Appraisal Tool

Appendix B: Adapted Downs and Black's Critical Appraisal Tool

Appendix C: Table Showing Critical Appraisal for Included Studies

Items:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Total score (%)
Aarts et al. (2012)	1	1	1			1	1	0		1	0	0				1		1		1							0	69
Berman et al. (2016)	1	1	1	1	0	1	1	0		1	0	0		1	0	1		1		1	1	1	1	1	1		0	73
Blashfield et al. (1985)	1	1	0	1		1	1	0		0	0	0				1		1		1	1						0	60
Bruchmüller & Meyer (2009)	1	1	0	1	1	1	1	0		0	1	0		1	0	1		1		1	1	0	0	0	1		0	59
Bruchmüller et al. (2012)	1	1	1	1	2	1	1	0		1	1	1		1	0	1		1		1	1	1	1	1	1		1	91
Cwik & Margraf (2017)	1	1	1	1	2	1	1	0		1	1	0		1	0	1		1		1	1	0	1	1	1		0	78
De Los Reyes & Marsh (2011)	1	1	1	1		1	1	0		1	0	0		1	0	1		1		1	1	1					0	72
DeRoma et al. (1997)	1	1	1	1	0	1	1	0		1	0	0		1	0	1		1		1	1	0	1	1	0		0	64
Evans et al. (2002)	1	1	1	1	0	1	1	0		1	1	0		1	0	1		1		1	1	0	1	1	1		0	73
Ford & Widiger (1989)	1	1	1	1	1	1	0	0		0	1	0		1	0	0		1		1	1	0	1	1	1		0	61
Fuller & Cowan (1999)	1	1	0			1	1	0		1	0	0				1		1		1							0	62
Garb(1996) Study 1	0	1	0			1	0	0		0	1	0				1		1		1							0	46
Garb (1996) Study 2	1	1	0	1	0	1	0	0	1	0	1	0		1	0	1	0	1		1	0	0	0	0	1	0	0	44
Garb (1996) Study 3	0	1	0			1	0	0		0	1	0				1		1		0							0	38
Kerr et al. (2004)	1	1	1	1	1	1	1	0		1	0	0		1	0	1		1		1	1	0	1	1	1		0	70
Kim & Ahn (2002) Study 1	1	1	0			1	1	0	1	0	0	0				1	1	1		1						0	0	56
Kim & Ahn (2002) Study 2	1	1	0			1	1	0	1	0	0	0				1	1	1		1						0	0	56

Kim & Ahn (2002) Study 4	1	1	0			1	1	0	1	0	0	0				1	1	1		1						0	0	56
Items:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Total score (%)
Kirk et al. (1999)	1	1	1	1		1	1	0		0	0	0		1	0	1		1		1							0	69
Loring & Powell (1988)	1	1	0	1	1	1	1	0		0	1	0		1	0	1		1		1	1	0	1	1	1		0	68
Mendel et al. (2011)	1	1	1	1		1	1	0		1	0	0				1		1		1	0						0	67
Mikton & Grounds (2007)	1	1	0	1	1	1	1	0		1	0	0		1	0	1		1		1	1	0	1	1	1		0	68
Payne (2012)	1	1	1	1	2	1	1	0		1	1	0		1	0	1		1		1	1	1	1	1	1		1	91
Pottick et al. (2007)	1	1	1	1	2	1	1	0		1	1	0		1	0	1		1		1	1	1	1	1	1		0	86
Pottick et al. (2017)	1	1	1	1	1	1	1	0		1	1	1		1	0	1		1		1	1	1	1	1	0		0	82
Spengler & Strohmer (1994)	1	1	1	1	1	1	1	0		1	1	0		1	0	1		1		1	1	0	1	1	1		1	82
Stewart (2004)	1	0	1	1	2	1	1	0		1	1	0		1	0	1		1		1	1	0	1	1	1		0	77
Trierweiler et al. (2005)	1	0	0	1	1	1	1	0		1	0	0				1		1		1	1	0			1		0	61

Appendix D. Review of eligibility for inclusion in meta-analysis for those studies where meta-analysis potentially feasible.

Subject Area	Study	Independent Variables	Dependent Variable	Study Design	Outcome	Report means and SDs?	Measure of effect comparable to other studies?
Conduct Disorder	Kirk et al., (1999)	Social context surrounding antisocial behaviors.	Respondent's judgments about whether the adolescents described in the vignettes had a mental disorder.	Cross-sectional Survey	Contextual influences upon decision-making related to applying disorder or non-disorder diagnosis.	Means only	Yes, because stats test used (ANOVA) contained internal dysfunction, environmental reaction and neutral as variables.
	De Los Reyes & Marsh, (2011)	Contextual information to suggest likelihood of either diagnosis or non-diagnosis of conduct disorder	Clinicians judgement about the likelihood of patients meeting conduct disorder criteria or not.	Cross-sectional Survey	Whether contextual information about patients' clinical presentations affected clinicians' judgments of conduct disorder symptoms.	Yes	Yes, because stats test used (ANOVA) contained consistent context, inconsistent context and noncontextualized judgments context. Can be compared to Kirk et al (1999). Cohen's d reported so can also be converted to *r* and compared with Pottick et al (2007).

Subject Area	Study	Independent Variables	Dependent Variable	Study Design	Outcome	Report means and SDs?	Measure of effect comparable to other studies?
Conduct Disorder	Pottick et al., 2007	Contextual information suggesting either disorder or nondisorder Race/ethnicity of client	Respondent's judgment about whether the adolescent described in the vignette has a mental disorder.	Experimental Study	How clients' race/ethnicity and clinicians' professional and social characteristics affect their judgment of mental disorder among antisocially behaving youths.	No	Yes. Still possible despite measure of effect (odds ratio) comparing different contexts to no context rather than to each other.
	Pottick et al., 2017 (Same data set as Pottick et al., 2007)	The social context surrounding antisocial behaviors.	Respondent's judgment about the effectiveness of 14 intervention approaches that are often used to treat antisocially behaving youth	Experimental Study	Clinician judgments about treatment for antisocially behaving youth based on the symptom's social context (e.g., life circumstances) and the youth's race or ethnicity.	No	No because stats tests used were to compare judgement of treatment effectiveness rather than about disorder/non-disorder.

Subject Area	Study	Independent Variables	Dependent Variable	Study Design	Outcome	Report means and SDs?	Measure of effect comparable to other studies?
Client and clinician race	Mikton & Grounds (2007)	Patient race and diagnosis. Clinician race.	Respondent's judgement regarding which PD diagnosis to client.	Experimental Study	Influence of client and clinician ethnicity upon clinician's judgments regarding PD diagnosis.	No	Yes. Chi-square odds ratio reports allocation of no diagnosis or any PD diagnosis based upon whether client black or white. Also difference between black and white clinicians to attribute a diagnosis of any PD.
	Loring & Powell (1988)	Sex and race of client and psychiatrist. Similarity of client and psychiatrist sex and race. Sex of client by race of client.	Respondent's judgement regarding which diagnosis to assign to client.	Experimental Study	Influence of sex and race of client and psychiatrist upon diagnosis.	No	No. Parameter estimates (log linear analysis) reports effects of client and psychiatrist being black or white upon client receiving undifferentiated schizophrenia diagnosis. Effects size conversion not possible.

Subject Area	Study	Independent Variables	Dependent Variable	Study Design	Outcome	Report means and SDs?	Measure of effect comparable to other studies?
Client and clinician race	Garb (1996) Study 2	Race of Client.	Difference in likelihood ratings of developing schizophrenia, major depression, brief reactive psychosis.	Quasi-experimental Study	Impact of client race in the differential diagnosis of schizophrenia and brief reactive psychosis.	No	No. effect size of race of client (black or white) upon receiving a likelihood rating for brief reactive psychosis not reported.
Prototype Cases	Blashfield, et al (1985)	Differences explored among the participants as a function of profession or experience	1. Disagreement statistic. 2. Frequency of diagnostic label use. Frequency of specific diagnoses. 3. Reaction time.	Cross-sectional survey	Inter-rater reliability in defining a prototype. Distinctiveness from other categories also explored to define prototypicality amongst cases.	Means only	1. No. Mean difference of a disagreement statistic not statistically significant. 2. Chi square odds ratio reported. 3. A 2x2 ANOVA. F-values reported
	Evans et al. (2002)	Factors suggestive as important determinants when diagnosing personality disorders varied factorially.	Prototypicality rating.	Experimental Study	Impact of three factors upon the diagnostic process pertaining to personality disorders.	No	No. Repeated ANOVA F-values reported.

Subject Area	Study	Independent Variables	Dependent Variable	Study Design	Outcome	Report means and SDs?	Measure of effect comparable to other studies?
Sex bias	Bruchmüller et al. (2012)	Sex of the child.	ADHD diagnosis or non-diagnosis	Quasi-experimental Study	To assess whether ADHD would be diagnosed more frequently in the boy vignettes than in the girl vignettes.	No	No, different outcome to Ford & Widiger (1989). Chi square odds ratio reported.
	Ford & Widiger (1989)	Sex of the client.	Histrionic personality disorder and antisocial diagnoses.	Experimental	To assess difference in histrionic and antisocial personality disorder diagnoses amongst men and Women.	No	No, different outcome to Bruchmüller et al. (2012). Chi square odds ratio reported.

Part Two: Research Report

Clinical decision-making in stepped-care; testing the influence of heuristics and biases on the decisions made by Psychological Wellbeing Practitioners in the IAPT programme

Abstract

Objective. The manner in which heuristics and biases influence the decisions of mental health workers has not been fully investigated and the methods previously used have been rudimentary. Two studies were conducted to design and test a trial-based methodology to assess the influence of bias on decisions regarding treatment allocation and progression.

Method. Using qualitative analysis an innovative 'real time' scenario-based approach was developed in the first study (referred to as a dynamic measure). The second study employed quantitative analysis to test the dynamic measure's ability to identify differences in decision-making between Psychological Wellbeing Practitioners (PWPs). A sample (N= 133) of PWPs completed two decision-making tasks. Decisions when encountering a particularly challenging scenario were compared with when treatment was relatively straightforward. Participants also completed validated static measures of decision-making style, reflective capacity and personality.

Results. Cumulatively dynamic measure score was not predicted by decision-making task or the static measures. When treatment fidelity and decisions to prolong or conclude treatment were examined in isolation variability in the responses to these scenarios were not better explained by chance. This differed relative to which case vignette participants received in the experimental condition.

Conclusions. PWPs may vary in the decisions they make regarding treatment delivery and this has implications clinically for patients seen in the early stages of the stepped-care model and organizationally. The degree of treatment fidelity demonstrated by PWPs, and reasons why they might sometimes prolong or conclude treatment may be due to an interaction between the PWP and the context.

Practitioner Points

1. PWPs vary in the decisions they make regarding treatment allocation and delivery. Decisions could be affected by heuristics and biases which may adversely influence patient outcomes.
2. PWPs may benefit from supervisors providing feedback on what particular biases are likely to be activated in certain situations. Especially when making decisions regarding 'complex' clients.
3. The study was limited by the design of the dynamic measure (e.g. lack of variability in the scoring system) thus reducing its ecological validity.
4. Given that the present study was explorative and the convergent validity of the dynamic measure as a test of heuristics and biases was not achieved further research is required.

Introduction

Evidence-based psychological interventions

The Improving Access to Psychological Therapies (IAPT) programme in the UK is a national programme offering access to evidence-based psychological therapies recommended by clinical guidelines for the treatment of anxiety and depressive disorders (NICE, 2011). IAPT services follow a stepped-care model meaning therapy increases in terms of duration, frequency and intensity according to risk, severity and non-responsivity to previous interventions (Bower and Gilbody, 2005). Stepped-care is believed to be an efficient means of delivering psychological services (Haaga, 2000; Bower and Gilbody, 2005; Tolin, Diefenbach and Gilliam, 2011) and is also supported via necessary policy drivers (e.g. The Five Year Forward View for Mental Health; The Independent Mental Health Taskforce, 2016).

The large-scale national implementation of IAPT means that a significant number of patients are seen annually across these services. Over one million patients per year are referred (Clark, 2019) and therefore the assessment skills of practitioners working in the early steps of the model are important clinically (e.g. treating patients suitable for low intensity approaches) and organizationally (e.g. the efficiency of the overall system). Psychological Wellbeing Practitioners (PWPs) work in the earliest step of IAPT and are trained to assess and then deliver brief low intensity psychosocial interventions for depression and anxiety disorders. The role of the PWP arose with the inception of IAPT and though PWPs are an established part of the system relatively little is known of the parameters and competencies of the role (Kellett et al, 2019). Previously, it has been likened to that of a coach rather than a traditional therapist (Turpin, 2010). Clearly, PWPs because they work in the early stages of the stepped-care model are making many decisions about the care pathway for many patients each year.

Therapist Variability in Stepped-Care

In spite of national curricula (UCL, 2015), clinical guidelines (NICE, 2011) and the availability of validated competency frameworks and measures related to assessment and treatment

procedures (Kellett et al, 2019), research suggests that PWPs vary considerably in their effectiveness and efficiency (Firth, Barkham, Kellett & Saxon, 2015). Whilst IAPT suggests the presence of homogeneity of decision-making, treatment allocation and treatment delivery, significant heterogeneity appears to be the norm (Johns, Barkham, Kellett & Saxon, 2019). This variability could be due to the influence of a range of contextual and clinical decision-making factors. For example, patient characteristics, attitudes and preferences may influence the decisions made by healthcare professionals (Visintini, Ubbiali, Donati, Chiorri and Maffei, 2007). Relationships with clients and colleagues, how confident the therapists feel, and perceptions regarding their own abilities influence the decisions clinicians make about treatment (Stavrou, Cape and Barker, 2009; Anthony et al., 2010; Sigel and Leiper, 2004; Pilgrim, Rogers, Clarke and Clark, 1997).

In a qualitative study based in stepped-care Gellatly, (2011) found that scarcity of resources (e.g., low numbers of high intensity therapists in a service) also had a substantial impact upon variability in treatment allocation decisions. Additionally, IAPT workers tended to adopt an overly individualized approach rather than follow standardised procedures and guidelines, especially when it came to decisions about "stepping up" or "holding" [1] patients and offering them lengthy interventions. Gellatly (2011) suggests this may be due to the "caring" values of health professionals conflicting with that of the "economic / public health" perspectives underpinning the stepped-care approach. Using a survey to gather information about IAPT therapists' clinical decision-making, Delgadillo el al. (2015) investigated 'stepping decisions' via principal component analysis, and found four distinct factors that were associated with a greater self-reported tendency to offer lengthy interventions – referred to as "holding patients in therapy"; (i) when the therapist believes there are obstacles in referring the client for further treatment; (ii) if the client is liked by the therapist; (iii) if there is a positive alliance between the patient and the therapist; (iv) if the

[1]Providing or delaying patients' access to more intensive treatments.

therapist is confident that they are capable to accomplish a positive outcome for the client by extending treatment. Delgadillo et al. (2015) concluded that incongruence and inaccuracy in decision-making was due to a complex interplay of beliefs, attitudes, subjective norms and self-efficacy.

The role of bias in decision-making

A highly influential theory regarding decision-making is the heuristics and biases model. This is based on research originally developed by Amos Tversky and Daniel Kahneman in the early 1970s. Kahneman & Tversky (1972) introduced the notion of cognitive biases; these occur unconsciously and may lead to a perceptual distortion regarding judgements made about the world. Biases emerge as manifestations of heuristics: strategies utilised from previous situations used to influence and inform current choices (Tversky & Kahneman, 1974). The role of bias might explain the well-established observation that clinical intuition tends to be inaccurate when determining likely prognosis of individual patients (Grove, 2005; Grove & Meehl, 1996).

"System 1" and "System 2". As his central thesis, Kahneman (2011) describes two modes of thought, based on terms originally proposed by psychologists Keith Stanovich and Richard West, and explores the different ways the brain has evolved and uses these to navigate through life. Kahneman describes "System 1" which is fast, intuitive and emotional, and "System 2" which is slower and more deliberate and logical. Kahneman (2011) claims that rather than generating new patterns linked to each new experience, "System 1" thinking employs the use of heuristics. This involves associating new information with existing prototypes. Some PWPs could be more prone to "System 1" thinking than others when making treatment allocation decisions. This variability may adversely influence outcomes for patients.

Anchoring and adjustment. Tversky and Kahneman (1974) first described "System 1" thinking when discussing 'Anchoring and adjustment'. This is the common human tendency to base too much significance upon the first piece of information proposed (the anchor) when making any decision. Once an anchor is established all other judgements are considered relative to it and

adjusted from it accordingly. Tversky and Kahneman (1974) demonstrated the strength of its effect in an experiment where participants were asked to estimate numerous amounts as a percentage, such as the percentage of African countries in the U.N. Answers were dependent a number, 10 or 65, seen beforehand. The experiment highlighted the idea that different starting values yield different estimates, as the average estimates of those who saw 10 and 65 were 25% and 45%, respectively.

The Halo Effect. Kahneman (2011) also introduces a term known as the halo effect. This describes a form of instant judgment discrepancy, or cognitive bias when a person forming a preliminary valuation of an individual, place, or object will suppose ambiguous information based upon that which is known as concrete information. Kahneman explains how the halo effect is employed as the tendency to like or dislike everything about a person including that which you have 'not observed'. Kahneman describes a classic psychological experiment conducted by Solomon Asch (1964). Participants were presented with two descriptions of individuals and asked to comment on each individual's personality. In the first description the individual's characteristics are presented in the following order: intelligent, industrious, impulsive, critical, stubborn, envious. In the second the same characteristics are presented as so: envious, stubborn, critical, impulsive, industrious, intelligent. In this experiment the majority of participants claimed that their impression of the individual changed based upon the order of the descriptions. Overall the individual described in the first list was viewed much more favorably than that of the individual in the second. Kahneman explains that the stubbornness of an intelligent individual is viewed as defensible, possibly even earning respect. On the other hand, intelligence in those who are envious and stubborn may equate to them being more dangerous. Therefore, the halo effect demonstrates a suppressed ambiguity. Considered relative to decision-making, this example highlights that sequence matters. Despite the order an individual's personal characteristics are observed often being down to chance, the halo effect increases the weight (i.e. influence) of first impressions. Consequently, any subsequent information is mostly wasted.

Rationale for the current research

It appears plausible that both anchoring, and the halo effect might be unconsciously employed by PWPs during assessments and therapy and therefore impact their work-related judgments and decisions. This is important given the number of patients seen for assessment and this influencing what type of treatment happens where. The degree of treatment fidelity demonstrated by PWPs, and reasons why they might sometimes "hold onto" patients rather than "stepping them up" to more intensive treatments may be due to an interaction between the PWP and the context (Delgadillo et al., 2015). PWPs are therefore likely to be susceptible to influence by heuristics and biases as part of their work-related judgments and decisions (Ægisdóttir et al., 2006). Existing research has examined therapist alignment to treatment protocols (e.g. Firth et al., 2015; Lambert, 2010; Thijssen, Albrecht, Muris and de Ruiter, 2017) and issues relating to "stepping up" and "holding" [2] (e.g. Delgadillo et al., 2015; Davison, 2000). Research focussing specifically upon the treatment and assessment decisions of PWPs is sparse especially given that existing clinical decision-making research largely relates to diagnostic decision-making.

Rationale for use of a dynamic measure

The case vignette method remains the most commonly used approach to measure the influence of bias on clinical decision-making (e.g. Spengler & Strohmer, 1994; Garb, 1996; Berman, Tung, Matheny, Glenn Cohen & Wilhelm, 2016). But the reliability and ecological validity of this method has been called into question (e.g. Hare-Mustin, 1983; Hyler, Williams, & Spitzer, 1982). Employing a scenario-based approach measuring clinical decisions in 'real time' might be one way of overcoming such issues. Therefore by employing a dynamic measure in the present study an analogue task was operationalised to track the accuracy of PWP decision-making regarding treatment allocation and progression and whether this is prone to anchoring and halo effects. This approach allows the measurement of a number of dynamic and context variables usually

[2]Providing or delaying patients' access to more intensive treatments.

only seen in experiments employing live interviews due to the "subtle cues that would not appear in a case summary" (Hyler, Williams, & Spitzer, 1982). Unlike a live interview approach a dynamic measure is also able to maintain tight control of the main variables of interest. Two studies were conducted to design and test this trial-based methodology to assess the influence of bias on decisions.

Specific objectives

The aims of the first study (Study A) were:

1. To develop a dynamic measure to:

   a.) Assess clinical judgement and reasoning traits of PWPs

   b.) Assess whether these are prone to anchoring and halo effects.

2. That the design of the dynamic measure will meet the following success criteria. Expert consensus will agree that the measure reflects clinical scenarios that have face validity to PWPs regarding:

   a. Patient suitability for treatment (e.g. allocation to step 2 or 3[3]).

   b. Treatment fidelity (e.g. degree of alignment to treatment protocol).

   c. Holding decisions (e.g. do PWPs choose to "hold", even if a client is not showing reliable improvement by session 4?).

The aims of the main study (Study B) were:

1. (a) For the dynamic measure to demonstrate sufficient levels of internal reliability during its preliminary testing.

   (b) That the dynamic measure will demonstrate sufficient levels of convergent and divergent validity during its testing.

2. (a) Identify differences in decision-making between PWPs.

[3] Step 2 refers to the low intensity treatments that are available in a stepped care model and typically this is a starting point for patients with mild-to-moderate conditions. Step 3 refers to the high intensity treatment pathway and is intended for those individuals who have not benefitted from Step 2 or whose mental health difficulties are somewhat more complex and severe.

(b) Use the dynamic measure to explore the influence of cognitive biases and heuristics on PWP decision-making.

(c) Profile the thinking styles of PWPs and examine the way these impact their clinical judgement and decision-making.

To measure the convergent validity of the dynamic measure and different thinking styles amongst PWPs the "Cognitive Reflection Test" (CRT; Frederick, 2005) and the Rational and Intuitive Decision Styles Scale (DSS: Hamilton, Shih & Mohammed, 2016) were employed. The Mini- International Personality Item Pool (Mini-IPIP; Donnellan et al., 2006) was also utilised to discover whether personality traits were affecting decision-making and if so, to what extent.

Hypotheses

Relative to the first aim of the main study (Study B), the hypothesis was as follows:

1. (a) Results will not differ according to whether participants receive case vignette 1 or 2 in the experimental condition.

   (b) Overall dynamic measure scores will positively correlate with the CRT (Frederick, 2005) for both case vignettes.

   (c) The dynamic measure total score will not significantly correlate with extraversion or neuroticism in either case vignette.

In relation to the secondary aims, the hypotheses were as follows:

2. a) PWPs with a higher score on the CRT (Frederick, 2005) will also have a more rational decision-making style on the DSS (Hamilton, Shih & Mohammed, 2016).

   b) When PWPs complete the experimental condition of the dynamic measure PWPs will tend to follow a counter-normative decision-making style and therefore achieve a lower dynamic measure score.

   c) A lower CRT score and a higher intuitive decision-making style score on the DSS will be correlated with a larger difference occurring between scores on experimental and control versions of the dynamic measure.

Ethics

Ethical approval for Studies A and B was granted by the University of Sheffield Ethics Committee in February 2018 (Reference 017478; Appendix B).

Study A: Development of the Dynamic Measure

Methodology

Design

A non-systematic review of the cognitive biases and heuristics literature was used to develop a preliminary draft of the dynamic measure in Study A. Previous studies have employed case vignette paradigms prompting respondents to make either "normative" (e.g. logical/expected) or "counter-normative" (intuitive/biased) choices/decisions (e.g. Kahneman & Tversky, 1972; Tversky & Kahneman, 1974) so this approach was used here. An inductive process was undertaken informed by ethnographic decision tree modelling (EDTM; Gladwin, 1989) that includes the 8 steps in the development of a composite group model (Figure 1). Appendix A contains a detailed explanation of this process. This incorporated thematic analysis of a focus group and a pilot study.

Recruitment

A purposive sampling methodology was employed to identify and select the most suitable, but also available, staff members for the most proper utilisation of available resources (Patton, 2002).

Ethnographic Decision Tree Modelling (EDTM; Gladwin, 1989)	
Phase 1: Model Building	
1	Identify the decision to be studied.
2	Specifying the set of decision options
3	The development of the researchers' ethnographic interviewing skills.
4	Participant observation
5	Selecting a sample of decision makers
6	Elicit the decision criteria
7	Develop a decision tree
8	Forming a Group Decision Model

Figure 1. The 8 steps of the EDTM Model Building Phase (Gladwin, 1989)

Participants for focus group. The author liaised with experienced PWP teaching staff from the Sheffield University IAPT Programmes. Subsequently N= 2 members of the course team were recruited.

Participants for pilot study. Teaching staff from the Sheffield University IAPT Programmes were approached via email to take part in a pilot study (Appendix C). Isaac and Michael (1995) suggest including between 10-30 participants in a pilot study. Therefore N =10 staff members were recruited.

Data Collection Procedures

1.) Once the preliminary draft of the dynamic measure was complete the focus group took place. A semi-structured interview document (Appendix D) was developed to guide the focus of the meeting and acquire qualitative data relating to the structure and content of the case vignette. The focus group lasted one hour and aimed to be informal, so participants felt able to speak openly and honestly about the dynamic measure design. Consent was sought to take part and for it to be audio recorded and transcribed (Appendix E).

2.) As anticipated mixed responses relating to the ecological versus the face validity of the dynamic measure emerged and so was discussed with the research supervisors. A 'living document'

(Shanahan, 2015) was developed and went back and forth to facilitate this discussion (Appendix F) until consensus was reached regarding necessary adaptations.

3.) Following the same process, to operationalise an experimental manipulation and test the internal reliability of the dynamic measure a second case vignette was developed (Appendix F1). Vignette content differed to the first so participants were unaware they were being tested for heuristics and biases.

4.) Following completion of the focus group a pilot study took place. Participants were emailed a link to a simulation of the full study (procedure listed in Study B). Participants were required to complete the first page of the survey detailing information relating to the study and consent. Upon completion participants were invited to provide feedback by email.

5.) Feedback from the focus group and pilot study was incorporated into the final survey design.

Analysis Strategies

The focus group and subsequent thematic analysis was conducted by the author. Issues of reflexivity were considered. The author's own bias was likely to influence how participant comments were received given the author is not a PWP and intended the dynamic measure to possess both ecological and face validity.

Data from the focus group was transcribed from audio recordings verbatim by a paid transcriber, managed using NVivo 12 (see Appendix G; Figure 1) and analysed employing thematic analysis (Braun & Clarke, 2006). To enable data immersion the transcript was read repeatedly. An essentialist/realist epistemological position was taken that assumes there is an accurate reality in the data. Initial codes and potential emerging themes were generated. Key themes were decided to describe how interviewees felt about the dynamic measure.

Inter-rater reliability. A secondary analyst, also a 3rd year Trainee Clinical Psychologist, independently reviewed preliminary codes, themes and sub-themes and these were discussed in a

meeting (Appendix G). Two contentious items were recoded resulting in a percentage agreement score of 97.8% and Krippendorff's alpha of 0.79.

Qualitative Results

Focus group

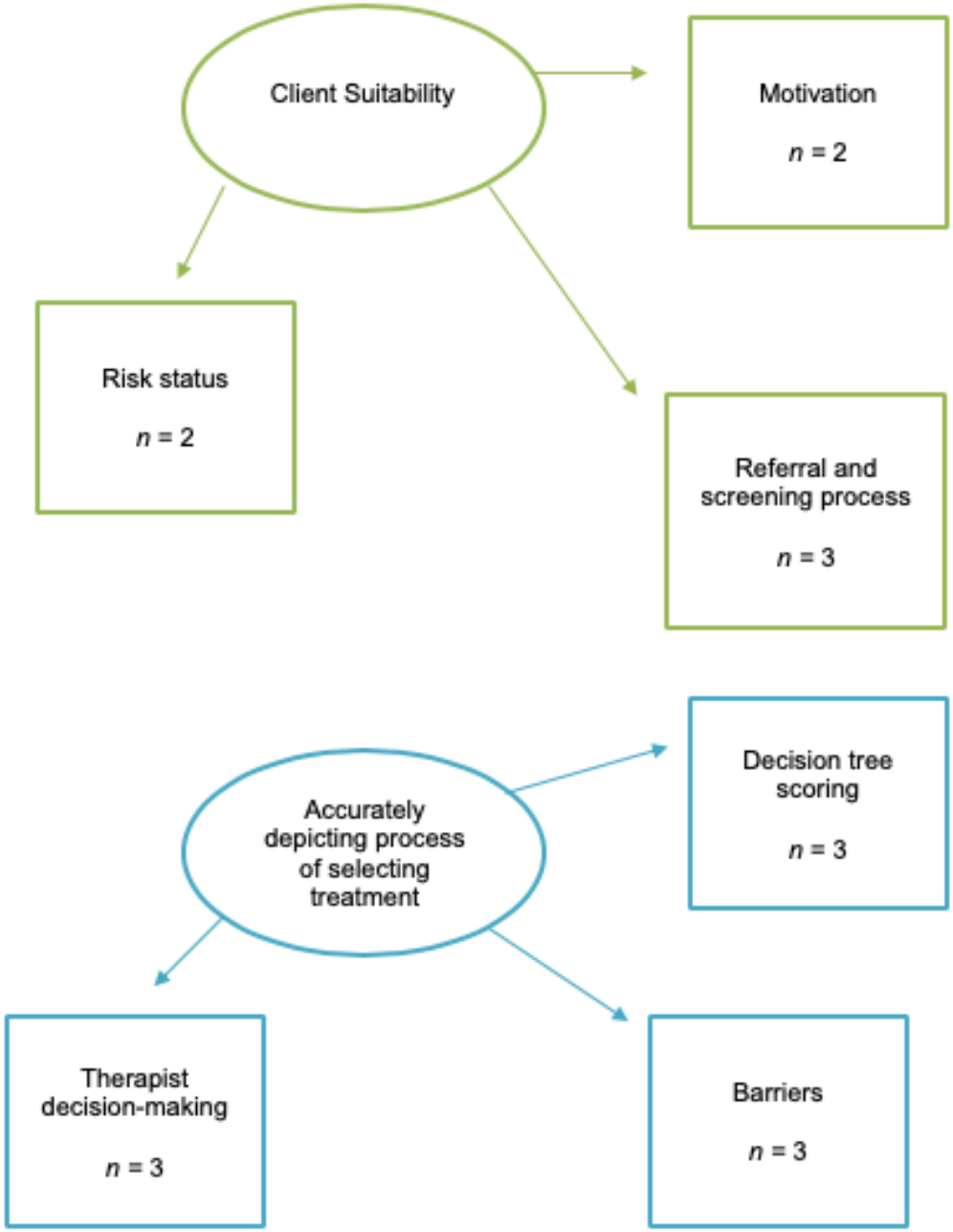
Appendix I shows the in-depth thematic analysis summarized in Table 1. Appendix J describes this process. Figure 3 depicts the themes and sub themes, and the number of times each was coded in the data.

Table 1.
Summary of focus group thematic analysis

Theme	Subtheme	Overview
Client Suitability	Risk Status	PWP teaching staff (Participant 1 and 2) commented that Jack's risk status sounded suitable for a standard patient who would be seen under IAPT (e.g., "[…] but that's (referring to Jack's profile) probably a sort of standard profile […]")
	Referral and screening process.	They explained the usual process when a client is referred to IAPT by their GP and the screening assessment process for treatment suitability in their region. The teaching staff agreed that the vignette accurately depicted a common situation in IAPT at the point of client screening (e.g., "Very true to life.")
	Motivation	One staff member talked about how important it was that the client 'bought into the model' and stated that the 'Jack' vignette gave a sense of this (e.g., "[…] because he's sort of done the 5 areas and the problem statement and the goal, he's done - the narrative that he's responded well to the 5 areas, I guess the – the options are good…")
Accurately portraying a collaborative approach	Setting goals with the client.	Both staff members commented that the chosen intervention (cognitive restructuring) did not appear in line with Jack's goals and suggested an alternative option (e.g., "You'd probably go on to the behavioural activation stuff."). Staff indicated the importance of collaborative goal setting at the point of assessing patient suitability (e.g., " […] you set a collaborative goal together. So is that then assuming that he's then on board with the process or..?")
	Client preconceptions	One staff member commented that in IAPT there was often some expectation that the client might not have been given correct information. This might be why the client is dubious and therefore this would be seen as understandable by the PWP (e.g., "[…] So it is quite natural that you will have someone who is a little bit dubious but will give it a go.")

Theme	Subtheme	Overview
Accurately depicting process of selecting treatment	Therapist decision-making	One staff member suggested why therapists often choose behavioural activation (BA) to begin with (e.g., " And quite often it's easier to see a change with BA, so you might start there with depression.") The current researcher discussed with staff whether it was realistic that in the vignette participants have the option to change their intervention. One staff member advocated for this and spoke about how common it was for therapists to change their minds (e.g., "[…] you do swap around when the pressure gets to you […]")
	Barriers	Staff spoke of some of the barriers that can come up regardless of what treatment a PWP selects for a client (e.g., "Or sometimes people will come back and say they didn't like it, but what they actually mean is they didn't understand it, or they need an explanation, or we've used too much jargon, or they might not be able to read…") Teaching staff suggested that the vignette could include reference to the Com-B approach as a way of overcoming potential barriers to the work (e.g., " […] we talk about using something called Com-B which is looking at someone's ability to […] understand, engage […]")
	Decision-tree scoring	Staff offered advice on how to structure treatment related content so as to make it congruent with both behavioural activation and cognitive restructuring. That way it would still be possible to score normative/counter-normative choices but participants would not suspect whether they had made the 'right' choice or not (e.g., "Or like he's – he brings his diary back, because that diary could be behavioural or cognitive.")

Theme	Subtheme	Overview
Ecological Validity	Realism of vignette	One staff member raised continuity issues in the vignette that meant the text would no longer apply to participants if they did not choose behavioural activation as a treatment option (e.g., "[…] if people select the previous one and go on to this, do they then realise they might have chosen the wrong option or?") Staff members suggested ways of making the content richer and more realistic such as doing homework in the session and adding more information about what got in the way of the client not doing homework (e.g., "So if he's not brought his BA diary, you might talk about how his week's been; if he's not brought his cognitive restructuring, you might do a bit […] you could look at what got in the way.") Staff members also commented upon how common the pull to offer more therapy is for therapists, even when client outcome measure scores show the client is not responding to therapy (e.g., "[…]because there's a pull when – there's a pull from patients when you're going to finish treatment, or when treatment's not working, to stick with them […] they almost like flatter you a little bit or say they like you […]"
	Evidence-base	Teaching staff discussed ideas for the 'hold' option of the vignette that were in line with the research literature but also realistic regarding what a PWP might do when it came to holding a client (e.g., "I don't know if someone would do a relapse prevention for someone that's not basically shown reliable or sort of – and is showing that improvement […] ")
	PWP Characteristics	One staff member suggested that it might be interesting to examine whether the stage of a person's career might impact their decision-making process (e.g., "[…] So, whether people who have been qualified longer, right at the beginning, like 'if you don't want it, you can go for counselling with mind', and newer people 'let's just try'.")



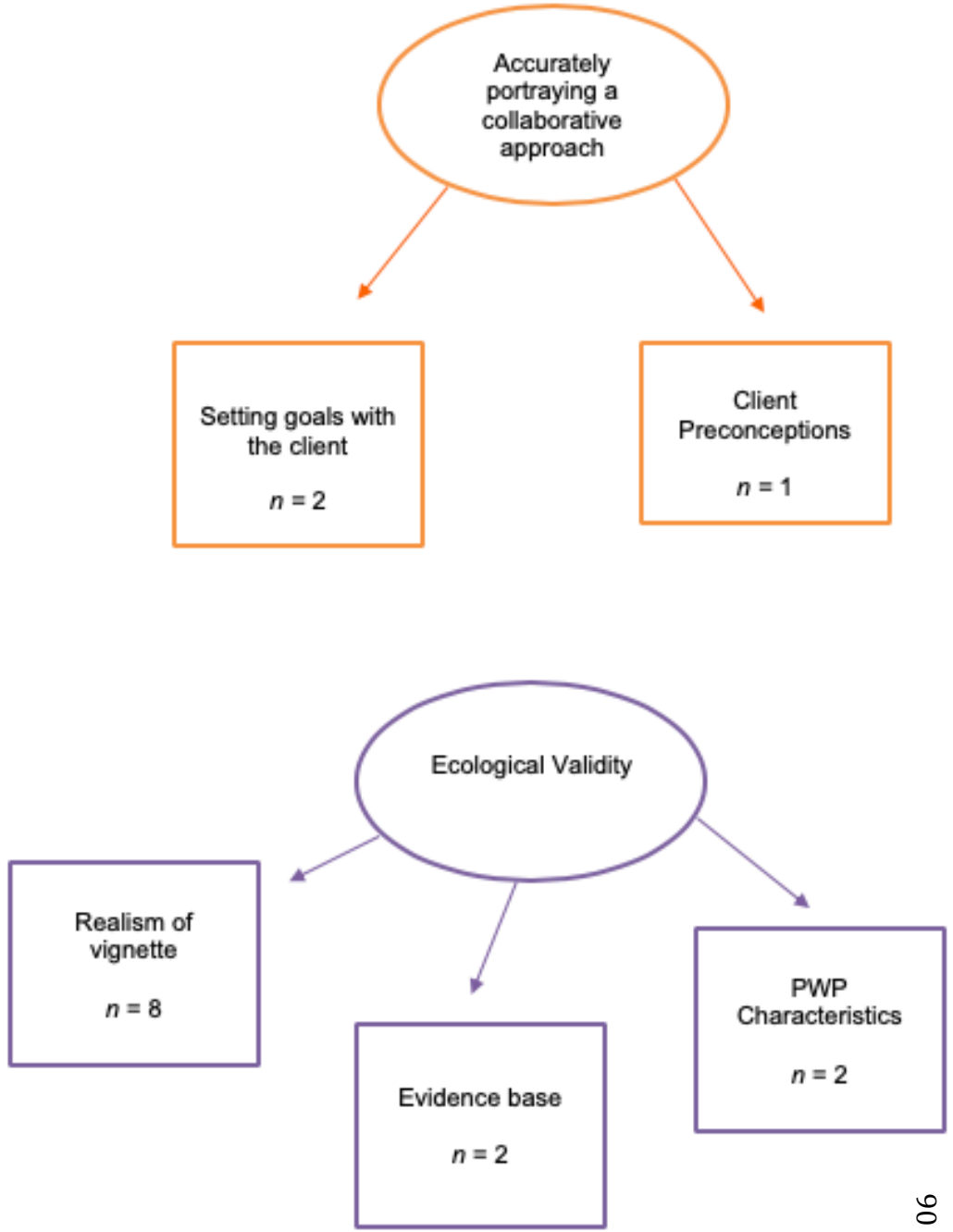


Figure 3. Themes generated from thematic analysis

Summary. Participants felt the case adequately reflected what might be seen in IAPT at step 2 and how PWPs might respond to the client during treatment. Participants suggested ways to resolve continuity issues in the vignette, how to make the content richer and more realistic, and how PWP characteristics might impact decision-making.

Pilot study

Table 2 lists feedback from the pilot study. Following discussion with research supervisors all suggestions were included in the dynamic measure except those relating to 'dynamic measure options/scoring'. This was because these related to specific IAPT service protocols and an option outside the scoring system.

Table 1.
Pilot Study: Views of the Dynamic Measure

Language	"All the questions seem good, in a language that PWP's will understand."
	"I would change the scenarios to 'assessment' session rather than screening session."
Level of detail	"The scenarios are detailed; they give enough information to build up a picture of the patient. It did not take long to fill in, which is a bonus."
Layout	"Also, for Chloe there seems to be a jump to session 4 (after session 2) - does that matter?"
	"The blocks of text for the case examples are quite dense, which might make it difficult for people to read/concentrate on them on screen, so if there is a way of spacing them out a bit more that might be better."
	"The only other bit of feedback I have is around reading the scenarios. This may seem picky, but I would consider putting some paragraphs breaks in to make them easier to read or double spacing."
Relevance to IAPT	"It felt very relevant to IAPT and a typical PWP presentation."
	"The case scenarios are good - true to life as a PWP, common dilemma's, succinct and easy to relate to."
Dynamic measure options and scoring	"I was mindful on one of the vignette questions both the answers would have been correct depending on the discussion in supervision. But I suppose that is the nature of clinical judgement.

> It was the second one about exploring COMb barriers or taking to supervision to discuss step up. I chose step up as I take anyone who's scores haven't reduced to supervision at session 4 but the outcome of supervision may well have been to do the 1st option. Alternatively, if I chose option 1 and it didn't work, I would have taken her after session 5. Of course, it all depends on service protocol (some supervision systems automatically select patients to take at session 4 regardless) and it is just my judgement. It would be interesting to see if that fits with other PWPs and supervisors?"
>
> "The other thing was about the first case example - at the stage where the patient does not really seem to be engaging, the options given are either to carry on working with him or step him up - as a PWP in reality I would probably not have done either of these, I would have been more likely to have a review with him to discuss non-completion of homework and consider whether now is the right time for him to be engaging in therapy/whether a different form of therapy might be more helpful."

Integrating findings

To detect the underlying criteria that underpins PWP decision-making relating to assessment and treatment (requisite face validity) expert consensus regarding each version of the dynamic measure was required. To achieve this success criteria some suggestions from the focus group remained but were simplified. Sections of the content were also generalised into an expanded criterion. Key changes included the number of options in stage two (treatment fidelity) and stage three (hold/step up) being reduced. This was to reduce risk of potential confounders influencing the results and to increase the likelihood vignettes were testing for anchoring and halo effects. It was also important that the experimental/control conditions got exactly the same information bar what was manipulated (internal reliability). Appendix F2 shows the final version of the dynamic measure including experimental/control conditions. Figure 3 illustrates the revised scoring system. The total each participant could score was 3 points.

The resulting changes ensured that Study B (main study) could proceed. The dynamic measure adequately reflected prototypical clinical scenarios encountered by PWPs relating to assessment and treatment decisions whilst also ensuring an acceptable level of internal reliability.

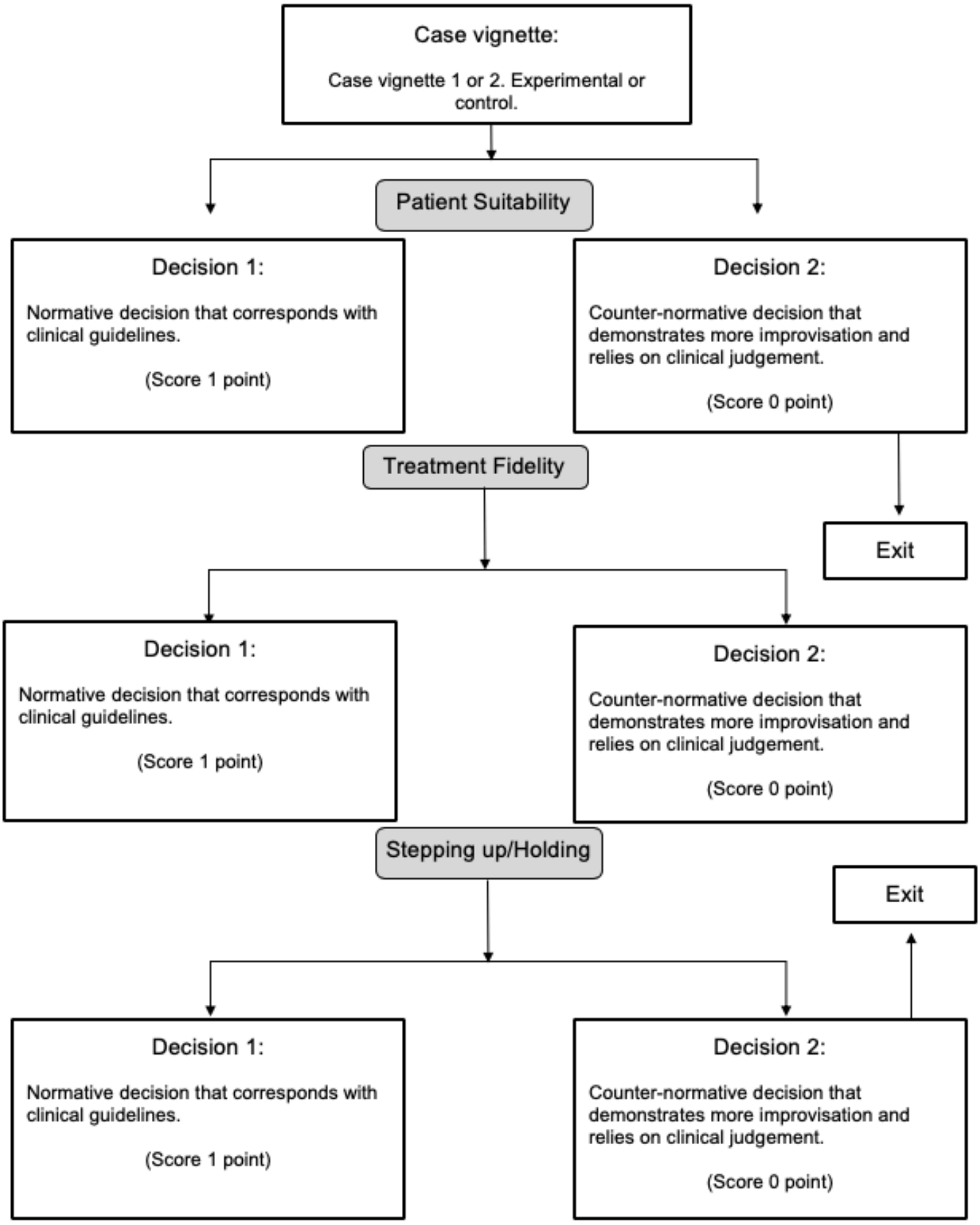


Figure 3. Revised scoring system of dynamic measure.

Study B: Main Study

Methodology

Design

A randomised crossover design was employed in Study B (see Figure 1). This followed an online survey design.

Participants

Power analysis. The Cognitive Reflection Test (CRT) shows an average correlation of r=.49 with performance on heuristic and biases tasks (i.e. decision-making case vignettes; Toplak, West & Stanovich, 2011). According to Cohen (1992), r = .50 is indicative of a large effect. On this basis a large effect was expected in the present study. Cohen's (1992) table was used to calculate the required sample size. Cohen (1977) proposes that 80% power is sufficient. Therefore, to show a large effect of 0.50 with an alpha or significance level of 0.05 and a power of 0.8 the required sample would yield a sample size of 38 Psychological Wellbeing Practitioners (PWPs) per group in an experimental design (total sample, 76) based on a linear multiple regression model (hypotheses 2b/c). The final sample included 133 participants after 57 participants were excluded as they did not complete the dynamic or static measures. Despite attrition all statistical tests remained adequately powered.

Recruitment. Participants were recruited over a four-month period (September 2018 - January 2019) and from a national sample of PWPs working as part of the Improving Access to Psychological Therapies (IAPT) programme in England (Clark et al., 2009). A convenience, snowball sampling method was used. Recruitment took place via email by approaching PWPs via the Psychological Professions Network, Health Education England, The British Psychological Societies (BPS) PWP Training Committee, Course Directors network list (nationally) and liaising with course lead contacts at Sheffield IAPT. The BPS Ethics Guidelines for Internet mediated Research was followed at all times (BPS, 2017). As an incentive to take part a £1 donation was

made to mental health charity, Rethink Mental Illness per participant for the first 50 PWPs completing the study.

Characteristics. Table 1 indicates inclusion/exclusion criteria of the study.

Table 1.

Inclusion and Exclusion Criteria.

Inclusion Criteria	Exclusion Criteria
Be a trainee or qualified PWP.	Therapists not working in IAPT services, in a PWP role.
Have access to a computer, an email and internet access to complete the online survey.	Those participants who did not provide complete data when completing the dynamic measure.

Clinician demographics. Table 2 shows participant demographics and clinical characteristics of those who provided information.

Table 2.

Participant demographics and clinical characteristics

Clinician Demographics/Clinical Characteristics	Frequency %	M (SD); Range
Gender		
Male	13.7%	
Female	86.3%	
Age		32.86 (9.13) 22-56
Ethnic Origin		
White	88.2%	
Mixed	9.8%	
Black or Black British	2%	
Years Qualified as PWP		5.02 (3.37) 1-8
N/A as still training	7.8%	
Less than 1 year	13.7%	
1 – 4 years	60.7%	
5 – 8 years	11.8%	
Ten or more years	5.9%	
Role within IAPT		
PWP Trainee	9.8%	
PWP	54.9%	
Senior PWP	19.6%	
Lead/Deputy Lead PWP	7.8%	
Other	5.9%	

Table 2. (cont.)

Participant demographics and clinical characteristics

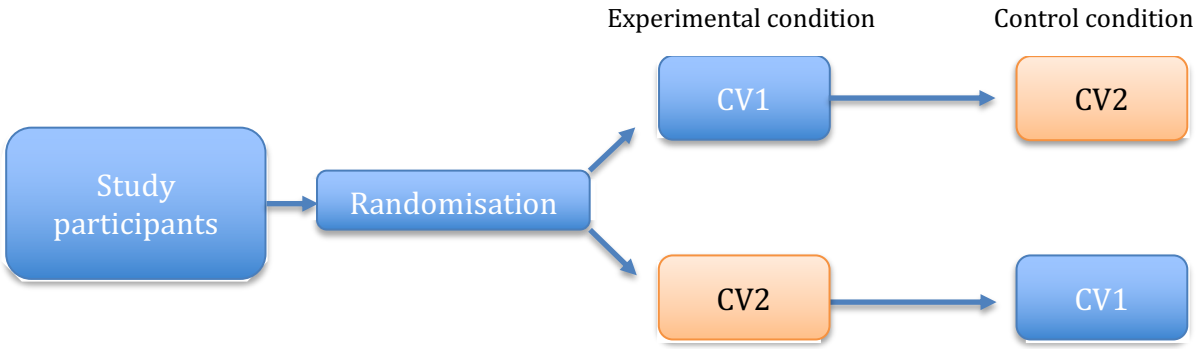
Clinician Demographics/Clinical Characteristics	Frequency %	M (SD); Range
Location		
Buckinghamshire	2%	
Cheshire	5.9%	
City of London	5.9%	
County Durham	7.8%	
Greater London	2%	
Greater Manchester	9.8%	
Lancashire	15.7%	
Lincolnshire	3.9%	
Merseyside	3.9%	
North Yorkshire	7.8%	
Nottinghamshire	3.9%	
Oxfordshire	11.8%	
South Yorkshire	5.9%	
Tyne and Wear	3.9%	
West Midlands	2%	
West Yorkshire	7.8%	

Procedure

Data collection. A hyper-link in the email took PWPs to the survey created in online survey software Qualtrics (2002). The first page included information relating to the study and a consent section (Appendix N). PWPs could only proceed after reading this and clicking to agree to take part. Participants could complete the survey up to one week from commencing.

Experimental manipulation. Participants completed two tasks so decisions when encountering challenging scenarios during low-intensity treatment (experimental) could be compared with when treatment was relatively straightforward (control). Participants were randomised to whether they received the case vignette 1 (CV1) or case vignette 2 (CV2) experiment, both including the relevant unfolding scenario, in the experimental condition. If participants received CV1 then they received CV2 as the control or vice-versa. Figure 1 illustrates the design and analysis of a cross-over trial. Figures 2 and 3 show flow of participants through the case vignettes/unfolding scenarios.

Figure 1. Illustration of the design and analysis of a crossover trial.



Outcome Measures

Dynamic measure. Clinicians were provided with a case vignette and three clinical-scenarios that related to patient suitability, treatment fidelity, and stepping up/holding. For each they were asked to choose the statement closest to what they would do next from two possible options. Options were conceptualised a priori as either "normative" (following clinical guidelines) or "counter-normative" (deviating from clinical guidelines). The experimental version of the vignettes was designed to evoke/prime heuristics and biases, thus increasing the likelihood of counter-normative responding.

Static measures. Clinicians completed validated static measures of decision-making style, reflective capacity and personality. To reduce the risk of order effect these were counterbalanced.

Cognitive Reflection Test (CRT; Frederick 2005). The CRT is a three-item measure that measures the tendency to override an initial "gut" response that is incorrect and engage in further reflection to find a correct answer (Appendix K). Toplak et al. (2011) showed the CRT is a particularly effective measure of 'miserly processing' in that it is a performance measure rather than self-report. Although the CRT substantially correlates with cognitive ability through a series of regression analyses Toplak et al. (2011) showed that it is also a unique predictor of performance on heuristics-and- biases tasks ($r = .49$). It accounted for substantial unique variance (11.2%, $p < .001$) after other measures of individual differences had been statistically controlled.

Rational and Intuitive Decision Styles Scale (DSS; Hamilton, Shih & Mohammed, 2016). This is a 10-item decision style scale capturing a broad range of the rational/intuitive thinking styles construct domains (Appendix L). Test–retest reliability was high for both rational ($r = .79$, $p < .01$) and intuitive ($r = .79$, $p < .01$) dimensions. The DSS was developed from three studies with five samples. Authors claim the resulting evidence shows dimensionality, stability, and validity (convergent/discriminant) of the DSS. The DSS has demonstrated high internal consistency and clear factor structure. Confirmatory factor analyses (CFAs) were employed to verify its two-factor structure. Fit indexes were generally at or above recommended standards across both samples. The 10-item scale correlates across decision-making, individual differences and the International Personality Item Pool (IPIP) Big Five traits.

The Mini- International Personality Item Pool (IPIP; Donnellan et al., 2006). The IPIP is a derivative of the 50-item International Personality Item Pool—Five-Factor Model measure. It has four items per Big Five Trait and is therefore a 20-item short form version of the original Five-Factor Model (Appendix M). It was developed and validated across five studies, all showed consistent and acceptable internal consistencies (alpha at or well above .60). Across intervals of a few weeks and then months the test-retest correlations of the IPIP scales were reported to be moderately comparable to the original Five-Factor Model. Convergent, discriminant, and criterion-related validity with other Big Five measures were also reported. The IPIP was included to allow divergent validity to be established. Hamilton, Shih & Mohammed (2016) found neuroticism and extraversion did not significantly correlate with rational/intuitive styles on the DSS. The same was therefore expected regarding the IPIP and the dynamic measure.

Descriptive analyses. To aid interpretation of the findings Means and SDs of participant's dynamic and static measure scores were calculated (Table 3).

Inferential analysis. Several inferential analyses were conducted. Non-parametric statistics were used to explore potential differences between experimental/control conditions. To test the convergent and divergent validity of the dynamic measure (hypothesis 1a) Pearson Correlations

were conducted between it and the static measures (between-subjects comparisons). Pearson Correlations were also conducted between each of the static measures (within- subjects). Multiple linear and logistic regression analysis were conducted, both mixed research design comparisons, to test interaction effects of continuous and categorical variables on a continuous dependent variable (hypotheses 2a/b/c). In both analyses and according to each stage of the dynamic measure a score of '1' indicated a normative decision corresponding with clinical guidelines. A score of '0' indicated a counter-normative decision demonstrating more improvisation and relying on clinical judgement.

For multiple regression analysis, CRT scores, rational/intuitive decision style scores (all within-subjects), and the Group variable (between-subjects classifying cases according to experimental/control) were entered simultaneously as predictors of the dynamic measure total score. A series of logistic regression analyses were conducted as part of secondary analysis to examine the difference between experimental/control conditions at stages two and three (treatment fidelity, stepping up/holding) of the dynamic measure. Stage one (patient suitability) was excluded from this analysis as all participants chose to see either patient as part of a step 2 intervention. The logistic regression followed the same process as the linear regression, including the same predictors. The dependent variable was the specific answer to the unfolding scenario (stage 2 in the first logistic regression model, stage 3 in the second). In this way, the regression analyses were designed to examine if the randomization to an experimental version of each vignette (group variable) was associated with systematically different responses to the clinical scenarios after controlling for standardized measures of decision-making style and CRT.

In both stages of the dynamic measure for both vignettes the reference category (0) was the control group, and the signal category (1) was the experimental group. The dependent variable was coded "1" for a normative answer and "0" for a counter-normative answer.

Data Analysis

Data were downloaded from Qualtrics and transferred to IBM SPSS V.24 for data analysis. Overall scores on the static and dynamic measures was used in the analysis.

Data Screening

There were some missing data from the static measures, accounted for by excluding cases listwise in the analyses.

Outcome data were screened in relation to the basic assumptions of parametric analysis (Appendix O). Firstly, data relating to the dynamic measure score were screened to assess the distribution of data in the experimental/control conditions for CV1/CV2. The data were found to violate the assumption of normality and so non-parametric statistics were utilized to assess potential differences between these conditions.

Sensitivity analysis was conducted to assess whether the assumptions for Pearson's correlation were met. Scatterplots were inspected to determine whether relationships between the included variables was linear; and to check whether there were any outliers that could be problematic. Histograms and Normal Q-Q Plots were also inspected to determine if the included variables were normally distributed. (Appendix O)

Sensitivity analysis was conducted to assess whether assumptions for multiple linear regression were met for both experiments. Normal distribution curves were inspected on histograms as were results from the Kolmogorov-Smirnov and Shapiro-Wilk tests of normality and scatterplots of the relationship between the outcome variable and predictors (Field, 2013). The relationship between the predictor variables and dependent variable in both experiments was not linear and therefore violated the assumptions of homogeneity of variance and linearity. In CV2 there was independence of residuals, as assessed by a Durbin-Watson statistic, of 1.905. In CV1 there were correlated errors, as assessed by a Durbin-Watson statistic, of 1.187. Dependent variables in either experiment also showed small negative skews (below 1) and were within the range of normal distribution. In both cases the Kolmogorov-Smirnov and the Shapiro-Wilk tests of normality were significant. The scatterplot of standardised predicted values verses standardised residuals showed data relative to CV1 violated the assumptions of homoscedasticity since the variation in the residuals was not constant (Appendix O). For CV2 the scatterplot showed the width of the scatter as

predicted values increased was roughly the same. Therefore, in this instance the assumption of homoscedasticity was met. There was no evidence of multicollinearity in either experiment, as assessed by tolerance values greater than 0.1. There were no studentized deleted residuals greater than ±3 standard deviations (SDs), no leverage values greater than 0.2, and values for Cook's distance above 1.

In light of these results, the dynamic measure score data for both experiments was transformed using the square root transformation. This further reduced the skewness value in the dynamic measure score variable for both from -.48 (SE = .21) to 0.32 (SE = .21) and from -.28 (SE = .21) to .04 (SE = .21) respectively. However, in both cases the Kolmogorov-Smirnov and the Shapiro-Wilk tests of normality remained significant. Log-linear transformation was attempted to account for the skewness issues in the dependent variable data, but the data remained skewed in both experiments.

As many of the assumptions of multiple linear regression were violated this provided a statistical rationale for performing logistic regression analysis relative to CV1/CV2 as part of the secondary analysis of the data. All logistical regression assumptions were met relative to both experiments other than linearity of the continuous independent variables with respect to the logit of the dependent variable. Therefore, the Box-Tidwell (1962) procedure to determine whether the continuous independent variables were linearly related to the logit of the dependent variable was carried out.

Results

Participant flow through dynamic measure

Figures 2 and 3 summarise the number of participants completing CV1/CV2 of the dynamic measure, that were in the experimental/control group, whether experimental manipulation influenced their normative/counter normative decision-making, and if this was different at each stage of the dynamic measure.

Distribution of data relating to experimental/control conditions

Kruskal-Wallis Tests were conducted for CV1/CV2 (between-subjects) to examine whether there were any total dynamic measure score overall differences between experimental/control conditions. No significant differences emerged in either experiment.

Correlational Analysis

Data from the static and dynamic measures was scored and calculated. The assumption of normality was only satisfied for the personality measures extraversion, neuroticism and intellect as assessed by visual inspection of their histograms and Normal Q-Q Plots (Appendix O). Preliminary analyses showed no linear or systematic-relationships between any of the dynamic and static measure scores (Appendix P).

Table 3.

Descriptive analysis of overall static and dynamic measure scores

Variables	Mean	N	SD
1. Dynamic Measure CV1	2.49	133	.55
2. Dynamic Measure CV2	2.18	133	.72
3. CRT Measure	1.28	133	1.28
4. DSS Rational Subscale	20.45	132	2.56
5. DSS Intuitive Subscale	13.49	132	3.01
6. Extraversion Total	12.18	133	3.62
7. Agreeableness Total	16.04	131	1.34
8. Neuroticism Total	15.12	133	3.03
9. Conscientiousness Total	11.56	133	2.97

10. Intellect Total	14.53	133	3.01

The only exception were relationships between agreeableness and extraversion and agreeableness and intellect. Monotonic relationships were lacking between these variables as the value of one did not increase as the other increased/decreased (Field, 2013). Subsequently, there was little value in attempting Spearman's non-parametric rank-order correlations to measure the strength and direction of the relationship between two continuous variables when relationships are non-linear. Therefore, results from the correlational analysis were not interpreted as they were invalid (Field, 2013), but are reported in Appendix Q.

Dynamic Measure Validation. As the dynamic measure failed assumptions of linearity in either experiment its convergent validity was not established. To establish its divergent validity, it was expected the total score should not significantly correlate with extraversion/neuroticism. As no relationships were observed with either personality subscale in either experiment the divergent validity of the dynamic measure was established. Therefore, Hypothesis 1a was partially supported.

Multiple linear regression analysis. A series of multiple linear regression analyses were calculated to predict dynamic measure scores based on baseline measures of decision-making style (CRT, DSS rational/intuitive subscales) and according to vignette group (experimental/control) for CV1/CV2. Regression coefficients and standard errors are reported in Table 4.

CV1. The multiple regression model did not significantly predict the dynamic measure score. The adjusted R-squared value was .04 meaning that 4% of the variation in dynamic measure score could be explained by the model. DSS Rational subscale score significantly predicted the dynamic measure score, $p < .05$. ($\beta = .19$, $p < .05$). Vignette group approached significance ($p = .052$). DSS intuitive subscale score and CRT score were not significant predictors.

Figure 2. CV1 Experimental Decision Tree Flow Diagram

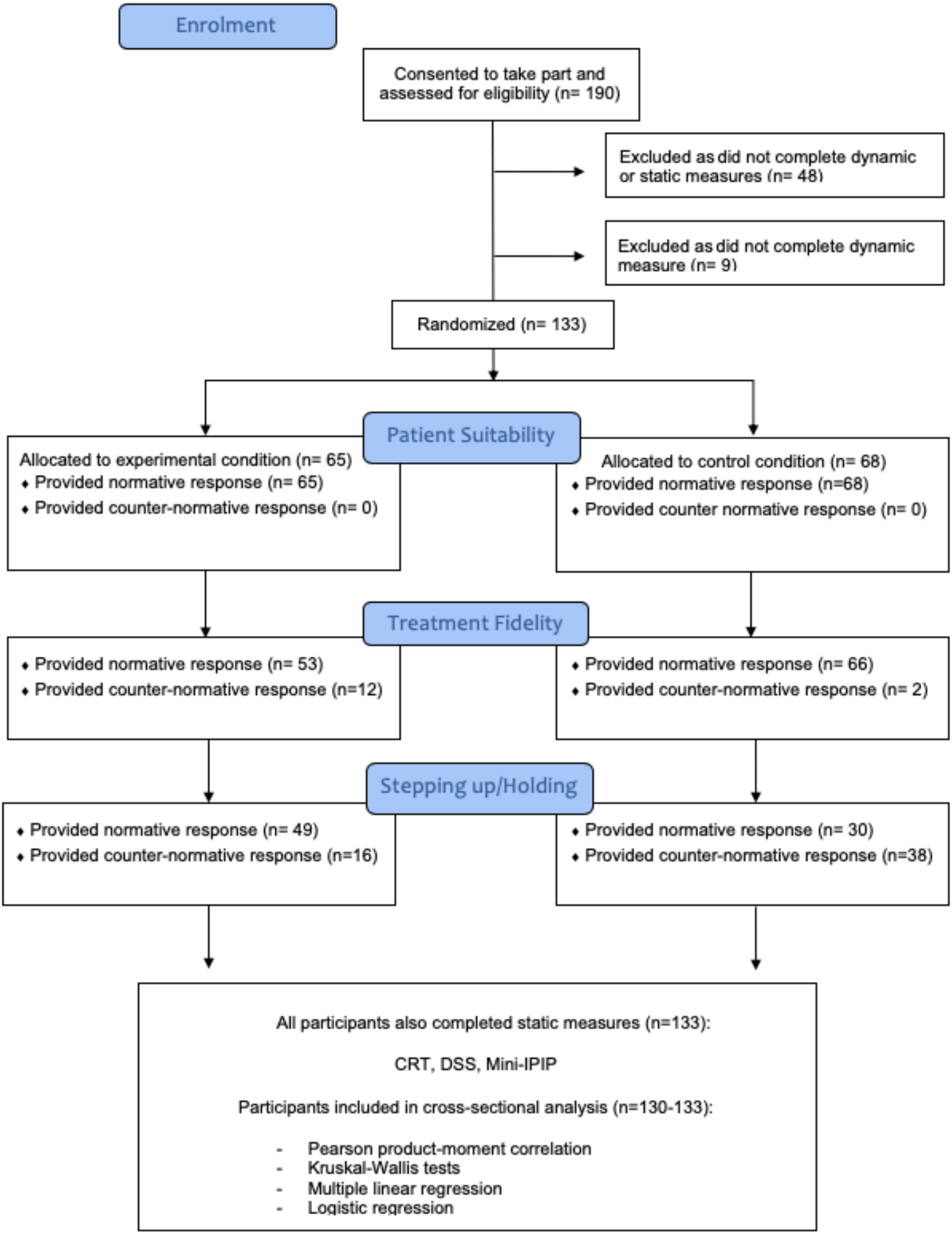
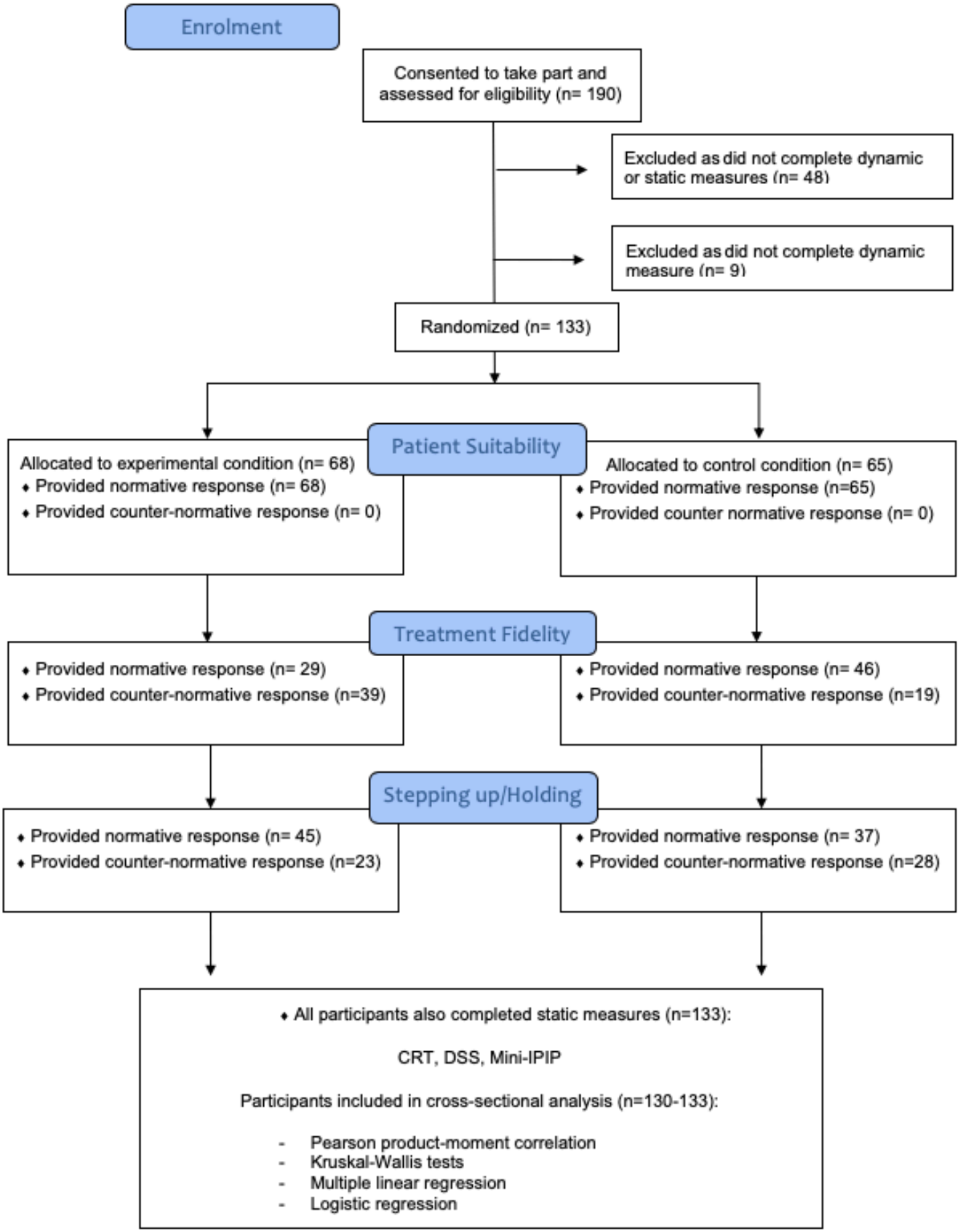


Figure 3. CV2 Experimental Decision Tree Flow Diagram



CV2: Again, the multiple regression model did not significantly predict the dynamic measure score. The adjusted R-squared value was -.001 meaning that 0% of the variation in dynamic measure score was explained by the model. None of the four variables significantly predicted the dynamic measure score.

As many of the assumptions of multiple linear regression were violated relative to both experiments these results could not be interpreted with confidence. In an attempt to overcome this and examine in more detail any relationships between existing measures of decision-making style, vignette group and dynamic measure scores logistic regression analysis was employed.

Logistic regression analysis. A logistic regression was performed for both experiments. This aimed to ascertain whether normative/counter-normative answers could be predicted at stage 2 (treatment fidelity) in the first logistic regression model, stage 3 (hold/step up) in the second. All continuous independent variables were found to be linearly related to the logit of the dependent variable. Linearity of the continuous variables with respect to the logit of the dependent variable was assessed for the first and second logistic regression models and in respect to CV1/CV2 via the Box-Tidwell (1962) procedure. A Bonferroni correction was applied using all eight terms in the model resulting in statistical significance being accepted when $p < .00625$ (Tabachnick & Fidell, 2014).

Testing for outliers.

CV1. In stages two and three there were no significant outliers in the analysis.

CV2. In stage two there were four standardized residuals with values of -6.957, -5.245, -2.855, -2.522 SDs respectively. After running "reflect and square root", "reflect and logarithmic" and "reflect and inverse" transformations more than one outlier remained. Therefore, data prior to transformation was utilized, the initial four standardized residuals were removed, and the regression-analysis was re-run. One standardized residual with a value of -2.810 SDs remained and this was kept in the analysis. In stage three there was one standardized residual with a value of -2.512 SDs, which was kept in the analysis.

Table 4.

Summary of Multiple Regression Analyses for Variables Predicting Decision Tree Scores in the Case vignette 1 and 2 experiments (N = 131)

	CV1			CV2		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Group (Exp/Con)	-0.08	0.04	-.17	0.08	0.05	.15
CRT Measure	-0.02	0.02	-.11	-0.01	0.02	-.02
DSS Rational Subscale	0.02	0.01	.19*	-0.01	0.01	-.08
DSS Intuitive Subscale	0.01	0.01	.10	-0.01	0.01	-.07
Adjusted R^2		.04			-.00	
F		2.35			0.98	

*$p < .05$; *B* = unstandardized regression coefficient; *SE B* = standard error of the coefficient; *β* = standardized coefficient

CV1.

Stage Two. The logistic regression model was significant, $\chi^2(4) = 20.391$, $p < .0005$. The model explained 37.0% (Nagelkerke R^2) of the variance in the answer to the unfolding scenario question and correctly classified 92.9% of cases. Sensitivity was 100.0%, specificity was .0%, positive predictive value was .0% and negative predictive value was 100.0%. None of the five predictor variables were significant (Table 5). Therefore, hypotheses 2a/b/c were rejected in CV1/stage 2.

Stage Three. The logistic regression model was significant, $\chi^2(4) = 18.679$, $p < .0005$. The model explained 17.9% (Nagelkerke R^2) of the variance in the answer to the unfolding scenario question and correctly classified 67.9% of cases. Sensitivity was 75.3%, specificity was 57.4%, positive predictive value was 71.6% and negative predictive value was 62.0%. Of the five predictor variables the Group variable was significant, and the rational decision style score approached significance (Table 6). The sign of the "group" coefficient was positive in this model suggesting that the effect of the experimental manipulation increased normative responding. Those in the experimental condition had 4.19 times higher odds to give a specific normative answer to the hold/step up unfolding scenario question. At CV1/stage 3 hypothesis 2b was rejected and as the CRT and DSS rational/intuitive subscales were not significant predictors hypothesis 2a/c were also rejected.

CV2.

Stage Two. The logistic regression model was significant, $\chi^2(4) = 11.939$, $p < .0005$. The model explained 11.7% (Nagelkerke R^2) of the variance in the answer to the unfolding scenario question and correctly classified 64.1% of cases. Sensitivity was 60.8%, specificity was 68.4%, positive predictive value was 71.4% and negative predictive value was 57.4%. Of the five predictor variables the Group variable was significant (Table 7). The sign of the "group" coefficient was negative in this model suggesting the effect of the experimental manipulation increased

Table 5.

Logistic Regression Predicting the Treatment Fidelity Question in the CV1 Experiment[4]

	B	SE	Wald	*df*	*p*	Odds Ratio	95% CI for Odds Ratio	
							Lower	Upper
Group (Exp/Cont)	-19.54	.4585.816	.00	1	.997	.000	.000	.
CRT Measure	.693	.380	3.32	1	.068	2.00	.95	4.22
DSS Rational Subscale	-.25	.21	1.40	1	236	.78	.52	1.18
DSS Intuitive Subscale	.22	.16	2.09	1	.149	1.25	.92	1.70
Constant	22.94	4585.82	.00	1	.996	9210021747.15		

Table 6.

Logistic Regression Predicting the Hold/Step Up Question in the CV1 Experiment

	B	SE	Wald	*df*	*p*	Odds Ratio	95% CI for Odds Ratio	
							Lower	Upper
Group (Exp/Cont)	1.43	.39	13.20	1	<.001	4.19	1.93	9.06
CRT Measure	.16	.16	1.04	1	.309	1.17	.86	1.59
DSS Rational Subscale	-.161	.08	3.80	1	.051	.85	.72	1.00
DSS Intuitive Subscale	-.096	.07	1.93	1	.165	.91	.79	1.04
Constant	4.12	2.23	3.41	1	.07	61.45		

[4] In both stages of the dynamic measure (treatment fidelity and hold/step up) for both vignettes (1 and 2) the reference category (0) was the control group, and the signal category (1) was the experimental group. The dependent variable was coded "1" for a normative answer and "0" for a counter-normative answer.

Table 7.

Logistic Regression Predicting the Treatment Fidelity Question in the CV2 Experiment[5]

	B	SE	Wald	*df*	*p*	Odds Ratio	95% CI for Odds Ratio	
							Lower	Upper
Group (Exp/Cont)	-1.22	.37	10.60	1	.001	.30	.14	.62
CRT Measure	.12	.15	.61	1	.436	1.12	.84	1.51
DSS Rational Subscale	.02	.08	.04	1	.839	1.02	.87	1.18
DSS Intuitive Subscale	-.002	.07	.00	1	.978	1.00	.88	1.14
Constant	.476	2.15	.05	1	.825	1.61		

Table 8.

Logistic Regression Predicting the Hold/Step Up Question in the CV2 Experiment

	B	SE	Wald	*df*	*p*	Odds Ratio	95% CI for Odds Ratio	
							Lower	Upper
Group (Exp/Cont)	.474	.36	1.70	1	.194	1.61	.79	3.28
CRT Measure	-.042	.15	.09	1	.770	.96	.72	1.28
DSS Rational Subscale	.07	.08	.77	1	.381	1.07	.92	1.24
DSS Intuitive Subscale	.07	.07	1.11	1	.293	1.07	.94	1.22
Constant	-2.05	2.11	.94	1	.332	.13		

[5] In both stages of the dynamic measure (treatment fidelity and hold/step up) for both vignettes (1 and 2) the reference category (0) was the control group, and the signal category (1) was the experimental group. The dependent variable was coded "1" for a normative answer and "0" for a counter-normative answer.

counter -normative responding. Those in the experimental condition had .30 times higher odds to give a counter-normative answer to the unfolding scenario question. The CRT and DSS rational/intuitive subscales were not significant predictors. Results confirm in CV2/stage 2 that hypothesis 2b was supported as the experimental manipulation significantly increased counter-normative responding. Hypothesis 2a/c were rejected.

Stage Three: Hold/Step Up: The logistic regression model was not statistically significant, $\chi^2(4) = 3.177$, $p > .0005$. The model explained 03.2% (Nagelkerke R^2) of the variance in the answer to the unfolding scenario question and correctly classified 63.4% of cases. Sensitivity was 92.5%, specificity was 17.6%, positive predictive value was 63.8% and negative predictive value was 60.0%. None of the five predictor variables were significant (Table 8). Therefore, hypotheses 2a/b/c were rejected.

Summary of results. To establish convergent and divergent validity of the dynamic measure and examine differences between experimental/control conditions several inferential analyses were conducted. Kruskal-Wallis Tests did not reveal significant differences between experimental/control conditions for either experiment. Pearson correlations were carried out but breached the assumption of linearity and so results were not reported. Subsequently, convergent validity of the dynamic measure was not established but divergent validity was. Multiple regression analysis was undertaken to predict dynamic measure scores based on baseline measures of decision-making style. For CV1/CV2 the model did not significantly predict dynamic measure score. In CV1 DSS Rational subscale score significantly predicted dynamic measure score. Results should be interpreted with caution given many of the assumptions of multiple linear regression were violated. Logistic regression analysis was conducted relative to CV1/CV2 to ascertain whether normative/counter-normative answers could be predicted at stage 2 (treatment fidelity) and stage 3 (hold/step up). This revealed significant results as in CV2 at stage 2 those in the experimental condition had .30 times higher odds to give a specific counter-normative answer to the unfolding

scenario question than those in the control. Similarly, in CV1 at stage 3 those in the experimental condition were significantly more likely to follow a normative decision-making style.

Discussion

This study explored variability in Psychological Wellbeing Practitioners (PWPs) decision-making, whether different thinking styles and certain cognitive biases influence this, and if so, under what clinical contexts. This study sought to replicate the clinical quandaries and dilemmas faced by PWPs working in stepped-care. To operationalise this procedure development and validation of a scenario-based 'dynamic measure'was attempted across two studies.

Summary of findings

Study A. The aims were achieved as the dynamic measure was developed and employed to assess clinical decision-making of PWPs.

Study B.

Dynamic Measure Convergent/Divergent Validity. It was not possible to determine whether the dynamic measure converged with another measure of heuristics and biases, the Cognitive Reflection Test (CRT), as results were not interpretable. In accordance with previous research examining relationships between measures of decision-making and personality (e.g. Hamilton, Shih & Mohammed, 2016) there was no relationship between the dynamic measure and extraversion/neuroticism. Therefore, hypothesis 1a was partly supported.

Cumulative Effect of Multiple Decisions. An overall dynamic measure score represented a cumulative effect of multiple decisions. As many of the assumptions of multiple linear regression were violated across experiments interpretations are made with caution. In both CV1/CV2 the model did not predict dynamic measure score. In CV1 the DSS rational subscale was a significant predictor ($\beta = .19$, $p < .05$) indicating greater rational thinking predicted more normative decisions. None of the four variables were significant predictors in CV2. In either experiment hypothesis 2a was rejected as the CRT score was not associated with the DSS Rational subscale in predicting dynamic measure score.

When clinical decisions were cumulatively tested, and clinical cases became more complex they did not reveal anchoring and halo effects. This was because vignette group in CV1/CV2 was not a significant predictor, although vignette group approached significance in CV1 ($p = .052$). Participant's total dynamic measure score for both experiments was relatively high (CV1 total mean score = 2.49; CV2 total mean score = 2.18). This suggests that, in general, within an IAPT context during treatment, PWPs predominantly make normative responses.

Further indication of this was seen in CV1 as participant's cumulative decisions were associated with a more rational decision style. This is congruent with clinical governance of PWPs supporting the notion that weekly IT-driven case management supervision enables PWPs to think normatively with a strong emphasis on adherence to treatment protocols and clinical guidelines. This could be why it made little difference when experimental/control groups were compared.

In CV2 no associations were found between the dynamic measure and the DSS Rational subscale. Perhaps cumulatively CV2 was not a very accurate test of decision-making.

Context-Specific Decisions. Logistic regression revealed that in specific circumstances certain clinical decisions are impacted by the situation since variability in the normative/counter-normative responses to these scenarios were not better explained by chance. This differed relative to whether participants received CV1/CV2 in the experimental condition.

In CV2 at stage 2 (treatment fidelity) significant results emerged. Those in the experimental condition had .30 times higher odds to give a specific counter-normative answer to the unfolding scenario question. In stage 3 of CV1 (hold/step up) those in the experimental condition were significantly more likely to follow a normative decision-making style. The logistic regression model as a whole was also statistically significant in both CV1/CV2 at stage 2 and 3 respectively.

Despite significant results hypothesis 2b was not supported in CV1 at stage 3 but was supported in CV2 at stage 2. In CV1/CV2 hypothesis 2a/c were also rejected as the static measures were not significant in either model. The sign of the "group" coefficient was negative in CV2 at

stage 2 and positive in CV1 at stage 3. Given these differing sets of results the internal reliability of the dynamic measure could not be established (hypothesis 1a).

Relationship to the existing theory/evidence

The convergent validity of the dynamic measure was not established, and there was only one significant relationship observed between the DSS and the dynamic measure. Therefore, it cannot be assumed it is a valid and reliable test of heuristics and biases or effective in identifying decision-making style of PWPs. Subsequently, results are interpreted with caution.

Results suggest certain heuristic and biases (in this case, anchoring and the halo effect) may only be observable when specific situations are examined in isolation. Experimental manipulation increased the likelihood of "stepping up" the client in CV1 but not CV2. There may have been something specifically about CV1 that triggered this. In the experimental condition indications the client was 'complex' included a history of depression/self-harm, recent unemployment and increased alcohol dependence. In CV2, despite an equally complex history there were no significant differences in terms of normative/counter-normative responding relating to stepping up /holding decisions resulting from experimental manipulation.

Information relating to CV1 being industrious/intelligent could be a factor but was provided in both conditions (experimental/control). Therefore, this alone does not explain the significant difference between them. In the experimental condition of CV1 the client's level of complexity (the anchor) combined with potentially being viewed as more industrious and intelligent (halo effect) may have been what influenced PWPs to "step up". CV1 may have been viewed as 'complex but capable' and therefore suitable for step 3 intervention.

In CV2 at stage 2 when PWPs completed the experimental condition this significantly more likely to follow a counter-normative decision-making style. The same did not occur in CV1 and this might be something to do with the level of detail regarding client complexity included in the experimental condition of CV2. Here we learn more about the predisposing factors that contributed towards the client's current difficulties than in CV1. The CV2 client experienced the loss of her

mother as a teenager, restricted what she ate, was bullied at school and went on to self-harm. This increased level of detail regarding historical information could have been what increased counter-normative responding in the experimental condition and subsequently led to lower therapist alignment to treatment protocol.

Gender-bias might also explain some of the variability between the male/female client in CV1/CV2. Research has found these influences diagnostic decision-making (Bruchmüller et al., 2012; Ford and Widiger, 1989). The same might be said for treatment decisions.

Critique

Participants/recruitment. Recruitment in Study B relied upon convenience, snowball sampling increasing the risk of self-selection bias. This has implications for the generalisability of the findings. Whilst random sampling would have been preferable recruiting enough participants took priority. PWPs also varied in their level of experience which may have impacted clinical decision-making.

Methodological critique. A significant limitation of the dynamic measure was its lack of convergent validity as a test of heuristics and biases. This could be due to adaptations made to the EDTM approach. Also, specific information regarding the complexity of the client might have influenced the decisions participants made regardless where it was located in the dynamic measure.

Preserving the ecological validity of the measure whilst ensuring it was empirically robust was challenging. The final version had a scoring system with a narrow range. This meant variability in scores was limited and may have contributed to many of the assumptions relating to correlation and multiple linear regression being violated.

An analogue approach was employed rather than studying decision-making in a naturalistic setting. Whilst strengths to this approach include tighter control of variables participants might have been inclined to respond in a socially desirable manner (Hare-Mustin, 1983). Also, participants may have felt less connectiveness and empathy towards the clients than in a naturalistic setting and this

may have increased the likelihood of providing more normative responses. Therefore, results may not be a true reflection of how PWPs actually respond in a real-life clinical setting.

Future Research

Decision-making research in IAPT and mental health settings generally is in the early stages of development and testing. Future research should seek to address limitations regarding the design of the dynamic measure such as the lack of variability in the scoring system and potential limitations regarding its ecological validity. Studies comparing decision-making styles of participants according to level of experience would also be beneficial.

The convergent validity of the dynamic measure could not be established, further research should address this issue. One method would be to test for other well documented cognitive biases such as the affect heuristic (Finucane et al., 2000) within the context of clinical decision-making.

Clinical Implications

Results have clinical implications for PWPs working in the early stages of stepped-care. If supervisors were able to provide feedback in case management supervision on what biases are likely to be activated in certain situations this might help develop a greater awareness of cognitive processes during therapy. PWPs and their supervisors should be particularly watchful when making decisions regarding 'complex' clients.

Conclusion

These findings partially support the assertion that psychological therapists are susceptible to influence by heuristics and biases as part of their work-related judgments and decisions (Ægisdóttir et al., 2006). PWPs may vary in the decisions they make regarding treatment delivery; this has implications clinically for patients seen in the early stages of the stepped-care model and organizationally (e.g. the efficiency of the overall system). More research is needed to investigate this both within the early stages of the stepped-care model and in the delivery of other evidenced based psychological therapies.

References

Anthony, J. S., Baik, S. Y., Bowers, B., Tidjani, B., Jacobson, C. J. and Susman, J. (2010). Conditions that influence a primary care clinician's decision to refer patients for depression care. *Rehabilitation Nursing, 35,* 113–122. https://doi.org/10.1002/j.2048-7940.2010.tb00286.x

Asch, S. E. (1964). The process of free recall. *Cognition: Theory, research, promise*, 79-88.

Beck, K. A. (2005). Ethnographic decision tree modeling: A research method for counseling psychology. https://doi.org/10.1037/0022-0167.52.2.243

Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: access, effectiveness and efficiency: narrative literature review. *The British Journal of Psychiatry, 186*, 11-17. https://doi.org/10.1192/bjp.186.1.11

Box, G. E., & Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics, 4,* 531-550.

Bruchmüller, K., Margraf, J., & Schneider, S. (2012). Is ADHD diagnosed in accord with diagnostic criteria? Overdiagnosis and influence of client gender on diagnosis. *Journal of consulting and clinical psychology, 80*, 128. http://dx.doi.org/10.1037/a0026582

Clark, D. M., Layard, R., Smithies, R., Richards, D. A., Suckling, R. & Wright, B. (2009). Improving access to psychological therapy: initial evaluation of two UK demonstration sites. *Behaviour Research and Therapy*, *47*, 910–920. https://doi.org/ 10.1016/j.brat.2009.07.010

Clark, D.M., (2019). *NHS England: IAPT at 10: Achievements and challenges.* Retrieved from: https://www.england.nhs.uk/blog/iapt-at-10-achievements-and-challenges/

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (revised ed.).

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Connelly, L. M. (2008). Pilot studies. *Medsurg Nursing, 17*, 411-2.

Davison, G. C. (2000). Stepped care: doing more with less?. *Journal of consulting and clinical psychology*, *68*, 580. https://doi.org/10.1037/0022-006X.68.4.580

Delgadillo, J., McMillan, D., Lucock, M., Leach, C., Ali, S., & Gilbody, S. (2014). Early changes, attrition and dose-response in low intensity psychological interventions. British Journal of Clinical Psychology, 53(1), 114e130. https://doi.org/10.1111/bjc.12031

Delgadillo, J., Gellatly, J., & Stephenson-Bellwood, S. (2015). Decision making in stepped care: how do therapists decide whether to prolong treatment or not? *Behavioural and cognitive psychotherapy*, *43*, 328-341. https://doi.org/10.1017/S135246581300091X

Delgadillo, J., Moreea, O., & Lutz, W. (2016). Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour research and therapy*, *79*, 15-22. https://doi.org/10.1016/j.brat.2016.02.003

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological assessment*, *18*, 192. https://doi.org/10.1037/1040-3590.18.2.192

Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R.S., Nichols, C.N., Lampropoulos, G.K., Walker, B.S., Cohen, G. & Rush, J.D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341-382. https://doi.org/10.1177/0011000005285875

Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics* (4th Ed.). London, UK: SAGE Publications Ltd.

Fiedler, K., & von Sydow, M. (2015). Heuristics and biases: beyond Tversky and Kahneman's (1974) judgment under uncertainty. *MW Eysenck & D. Groome, Cognitive psychology: revisiting the classical studies*, 146-161.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of behavioral decision making*, *13*, 1. https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S

Firth, N., Barkham, M., Kellett, S., & Saxon, D. (2015). Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: a multilevel modelling analysis. *Behaviour research and therapy*, *69*, 54-62. https://doi.org/10.1016/j.brat.2015.04.001

Fishbein, M., & Ajzen, I. (1975). Belief, attitude, intention, and behavior: An introduction to theory and research. Reading, MA: Addison-Wesley.

Ford, M. R., & Widiger, T. A. (1989). Sex bias in the diagnosis of histrionic and antisocial personality disorders. *Journal of Consulting and Clinical Psychology, 57*, 301. http://dx.doi.org/10.1037/0022-006X.57.2.301

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*, 25-42.

Gellatly, J. (2011). *Decision Making in Stepped Care for Common Mental Health Problems.* PhD Thesis. University of Manchester, School of Nursing, Midwifery and Social Work.

Gladwin, C. H. (1989). *Ethnographic decision tree modelling*. (Vol. 19). Sage.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Grove, W. M. (2005). Clinical versus statistical prediction: The contribution of Paul E. Meehl. *Journal of clinical psychology*, *61*, 1233-1243. https://doi.org/10.1002/jclp.20179

Haaga, D. A. (2000). Introduction to the special section on stepped care models in psychotherapy. *Journal of consulting and clinical psychology*, *68*, 547. https://doi.org/10.1037/0022-006X.68.4.547

Hamilton, K., Shih, S. I., & Mohammed, S. (2016). The development and validation of the rational and intuitive decision styles scale. *Journal of personality assessment*, *98*, 523-535. https://doi.org/10.1080/00223891.2015.1132426

Hare-Mustin, R. T., & Marecek, J. (1988). The meaning of difference: Gender theory, postmodernism, and psychology. *American psychologist, 43*, 455.

Independent Mental Health Taskforce, (2016). *Five Year Forward View for Mental Health for the NHS in England.* Retrieved from: https://www.england.nhs.uk/wp-content/uploads/2016/02/Mental-Health-Taskforce-FYFV-final.pdf

Isaac, S., & Michael, W. B. (1995). *Handbook in research and evaluation*. San Diego, CA: Educational and Industrial Testing Services. https://doi.org/10.1002/1520-6807(198207)19:3<413::AID-PITS2310190328>3.0.CO;2-X

Johns, R., Barkham, M., Kellett, S., & Saxon, D. (2019). A systematic review of therapist effects: A critical narrative update and refinement to Baldwin and Imel's (2013) review. *Clinical Psychology Review, 67*, 78–93. https://doi.org/10.1016/j.cpr.2018.08.004

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*, 430-454. https://doi.org/10.1016/0010-

Kahneman, Daniel; Frederick, Shane (2002), Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin and D. Kahneman (Eds). *Heuristics of Intuitive Judgment: Extensions and Applications.* New York: Cambridge University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Lambert, M. J. (2010). *Prevention of treatment failure: The use of measuring, monitoring, and feedback in clinical practice*. American Psychological Association. https://doi.org/10.1037/12141-000

National Institute for Health and Care Excellence (2011). *Common mental health problems: identification and pathways to care [CG123].* Retrieved from: https://www.nice.org.uk/guidance/CG123/chapter/1-Guidance#stepped-care

Patton, M. Q. (2002). *Qualitative research and evaluation methods (3rd ed.)*. Thousand Oaks, CA: Sage.

Pilgrim, D., Rogers, A., Clarke, S., & Clark, W. (1997). Entering psychological treatment: decision-making factors for GPs and service users. *Journal of Interprofessional care, 11*, 313-323. https://doi.org/10.3109/13561829709034128

Qualtrics (2002). Retrieved from: https://www.qualtrics.com/uk

Shanahan, D. R. (2015). A living document: reincarnating the research article. *Trials, 16*, 151. https://doi.org/10.1186/s13063-015-0666-5

Sigel, P., & Leiper, R. (2004). GP views of their management and referral of psychological problems: a qualitative study. Psychology and Psychotherapy: Theory, Research and Practice, 77(3), 279-295. https://doi.org/10.1348/1476083041839394

Stavrou, S., Cape, J., & Barker, C. (2009). Decisions about referrals for psychological therapies: a matched-patient qualitative study. *Br J Gen Pract, 59*, e289-e298. https://doi.org/10.3399/bjgp09X454089

Stoler, N. (1963). Client likability: A variable in the study of psychotherapy. *Journal of Consulting Psychology, 27*, 175. http://dx.doi.org/10.1037/h0041054

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics*. Harlow.

Thijssen, J., Albrecht, G., Muris, P., & de Ruiter, C. (2017). Treatment Fidelity during Therapist Initial Training is related to Subsequent Effectiveness of Parent Management Training—Oregon Model. *Journal of Child and Family Studies*, 1-9.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of applied psychology*, *4*, 25-29

Tolin, D. F., Diefenbach, G. J., & Gilliam, C. M. (2011). Stepped care versus standard cognitive–behavioral therapy for obsessive–compulsive disorder: A preliminary study of efficacy and costs. *Depression and anxiety, 28*, 314-323. https://doi.org/10.1002/da.20804

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition, 39*, 1275.

Turpin, G. (Ed.) (2010): IAPT *Good Practice Guide to using Self-help Materials*. NMHDU/IAPT, 1-40. https://doi.org/10.3758/s13421-011-0104-1

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185,* 1124–1130. https://doi.org/10.1126/science.185.4157.1124.

UCL (2015). *National Curriculum for the Education of Psychological Well-being Practitioners, 3rd edn*. Available at: https://www.ucl.ac.uk/pals/research/cehp/research-groups/core/pwp-review/docs/PWPREVIE_-curriculum

Visintini, R., Ubbiali, A., Donati, D., Chiorri, C., & Maffei, C. (2007). Referral to group psychotherapy: A retrospective study on patients' personality features associated with clinicians' judgments. *International Journal of Group Psychotherapy, 57*, 515-524. https://doi.org/10.1521/ijgp.2007.57.4.515

Appendices

Appendix A: Ethnographic Decision Tree Modeling Conceptual Framework in Relation to Study A

EDTM was developed to identify the necessary factors required for groups of people to make decisions. This method attempts to both explain and predict group behaviour by detecting the underlying criteria that underpins decision-making. EDTM has been widely utilised in research examining psychological, medical and social phenomena. Beck (2005) explains that researchers have previously applied this approach to study decision-making in the context of substance use treatment, child abuse reporting, and treatment choices in healthcare of patients with cancer. Broadly speaking EDTM involves two phases: model building and model verification. In the model-building phase ethnographic interviews are conducted aimed at identifying key factors that are used in the decision-making process (Beck, 2005). In the model verification phase, it is then tested on a separate, yet similar, group of individuals drawn from the same population (Beck, 2005). Gladwin (1989) provides a complete description of the development of ethnographic decision tree modeling.

Study A was informed by the EDTM model building phase but with some modifications made to the process. Conducting up to 20-30 ethnographic interviews with individuals from the group of interest (e.g. Psychological Wellbeing Practitioners) and as recommended by Beck (2005) would have been time-consuming, resource-heavy, and complicated to achieve. Instead, a focus group and a pilot study were conducted with experienced PWP teaching staff. This was considered as rigorous as a series of interviews given participants level of knowledge and experience relating to the stepped-care approach. Also, thematic analysis is a rigorous, theoretically flexible and widely used qualitative analytic method frequently used within psychology (Braun & Clarke, 2006). How the ETDM model building phase informed the process is listed below. The focus group and pilot study checked that the unfolding scenarios adequately reflected prototypical clinical scenarios

encountered by PWPs. The content of the decisions was also examined to explore whether they reflected the actual decisions teaching staff believed PWPs would make if faced with that situation.

Phase 1: Model Building

Step 1: Identify the decision to be studied. In accordance with step one of the EDTM model building phase an initial preliminary draft of the dynamic measure was developed. This was based on a fictional male client, "Jack". The content of the dynamic measure aimed to reflect the typical clinical decisions made by PWPs in their day-to-day clinical practice.

Step 2: Specifying the set of decision options. Relative to step 2 of the EDTM model building phase figure 1 demonstrates how each decision was situated in the dynamic measure, how the dynamic measure unfolded and how the various decision points were structured. The total number that each participant could score on a decision-total scale was 9 points. The case vignette aimed to test for anchoring (Tversky & Kahneman, 1974) and halo effects (Thorndike, 1920) where we expected biases to be activated and counter normative decisions to be made that don't always correspond with clinical guidelines.

A normative or counter-normative decision-making scoring scale was operationalised. A normative decision was conceptualised as one that corresponds with clinical procedures and guidelines consistent with NICE guidelines for depression and anxiety. A counter-normative decision was defined as one that demonstrates a more idiosyncratic and improvised response and relies more on clinical judgement. A normative decision scored 1 point; a counter normative decision scored 0 points.

Step 3: The development of the researchers' ethnographic interviewing skills. Prior to conducting the focus group and pilot study the author met with research supervisors, both experienced in the delivery of the stepped-care model and IAPT services more generally. Engaging in the process of collaboratively drafting an initial preliminary version of the dynamic measure with supervisors meant that the author was able to further develop ethnographic interviewing skills and

increase knowledge relative to IAPT processes and protocols. It also meant that the author was able to practice interview techniques on someone from the group of interest. Given that the author is not a PWP this increased their awareness of what specific decisions PWPs are required to make during therapy (e.g., decisions about patient suitability based upon assessment of client motivation) especially. This enabled the author to have greater insight and increase curiosity as to what specific questions to ask during the focus group with IAPT teaching staff.

Prior to the focus group taking place the author and supervisors developed a semi-structured interview document (see Appendix D) to guide the focus of the meeting. This could be likened to an invitation for the participant to share his or her decision-making process (Beck, 2005). Prior to commencement of the focus group the author provided an explanation as to the purpose of the focus group. This could be likened to an orientation statement that involves a description of the research study and its rationale (Gladwin, 1989).

Possible issues of reflexivity related to the process of conducting the focus group and analysing the resulting data were also considered (see 'Analysis Strategies' section in Study A).

Step 4: Participant observation. Once the preliminary draft of the dynamic measure was complete the focus group took place integrating step 4 of the EDTM model building phase.

Step 5: Selecting a sample of decision makers. In keeping with step 5 of the EDTM model-building phase (Gladwin, 1989) a sample of decision makers were selected to take part in a series of ethnographic interviews (focus group and pilot study). Those selected were seen as representative of the sample of interest especially given that they were PWPs themselves and had specialist knowledge regarding stepped-care processes and protocols within IAPT.

Step 6: Elicit the decision criteria. Following the focus group, the author, with some assistance from the research supervisors, looked to elicit the decision criteria that the PWP teaching staff used in their decision process (Beck, 2005). To derive decision criteria following the focus group involved creating more general categories to incorporate elements mentioned by teaching staff. For example, the criterion "I would empathize with Jack and introduce the principles of the

COM-B model with him in order to think about barriers" was created to more accurately depict the process of overcoming barriers. The amended decision criteria relating to the dynamic measure following the focus group can be found in Appendix F1.

Step 7: Develop a decision tree. This stage in the development of a decision tree was based upon the indirect method (Gladwin, 1989). This involves developing a group composite model on an ongoing basis after each interview. As the model building phase was based upon a focus group and pilot study rather than a series of ethnographic interviews the group composite model was developed after further discussions with the research supervisors. Rather than utilising a survey to test decision criteria from one decision maker to another the author developed a 'living document' (Shanahan, 2015) to facilitate this discussion. This was based upon the original initial preliminary draft of the dynamic measure. Discussions went back and forth via the use of the 'living document'.

Step 8: Forming a Group Decision Model. The final stage of the model building phase involved combining the results from the focus group and pilot study, along with the recommendations from the research supervisors, to form a group decision model. Eventually the author and research supervisors agreed that the dynamic measure was of a sufficient enough standard that the recruitment phase of Study B could commence.

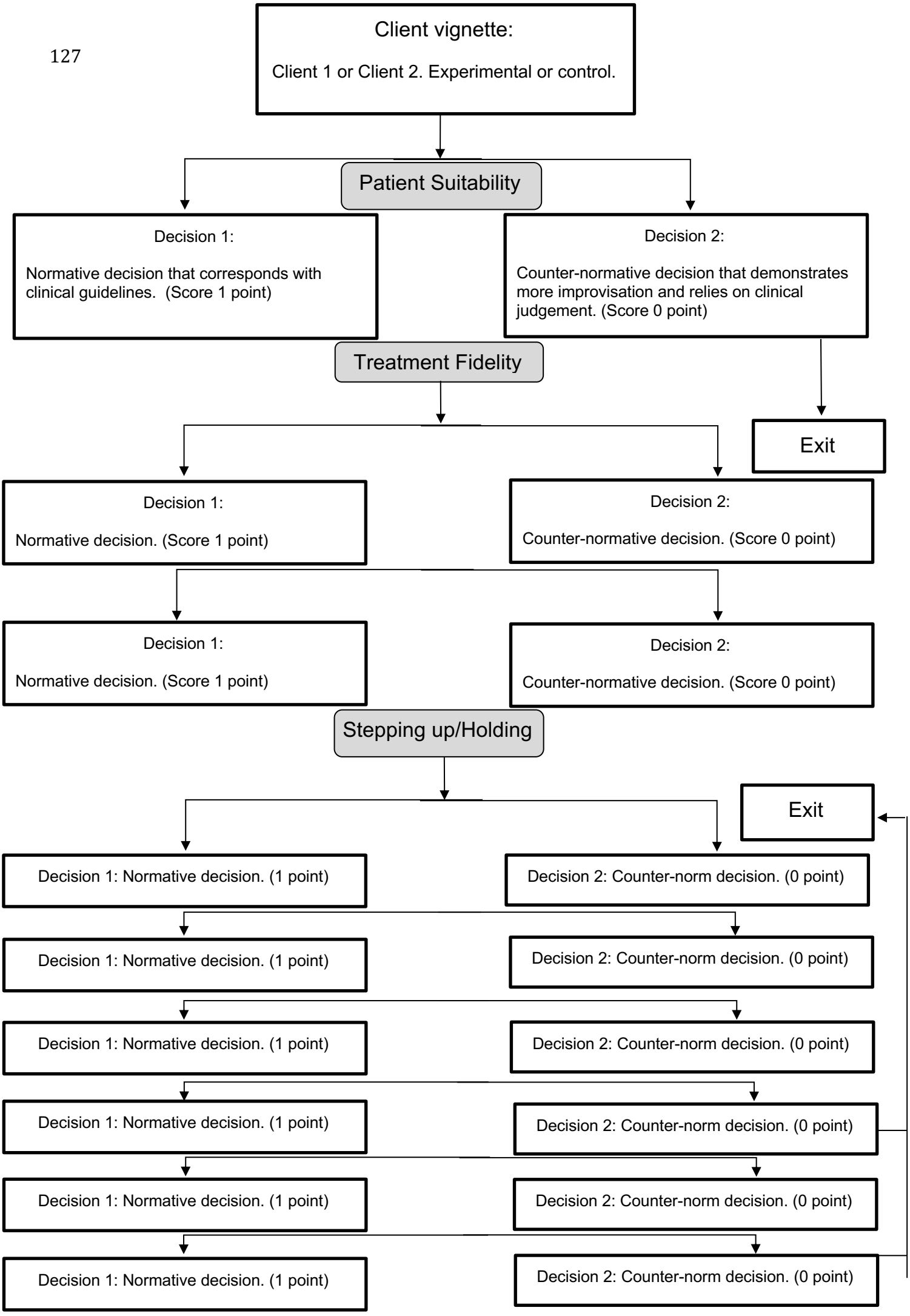


Figure 1. Initial scoring system of dynamic

Appendix B: Ethical approval

Downloaded: 06/02/2018
Approved: 06/02/2018

Benjamin Michael
Registration number: 160124433
Psychology
Programme: Doctorate in Clinical Psychology

Dear Benjamin

PROJECT TITLE: Clinical judgement: An investigation of clinical decision-making
APPLICATION: Reference Number 017478

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 06/02/2018 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 017478 (dated 29/01/2018).
- Participant information sheet 1038802 version 1 (10/01/2018).
- Participant consent form 1038803 version 1 (10/01/2018).

The following optional amendments were suggested:

Please amend the reference to the registrar, as the university no longer has a registrar. Now any complaints should be directed to Glenn Waller, as head of department, instead. Thank you.

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Yours sincerely

Thomas Webb
Ethics Administrator
Psychology

Appendix C: Recruitment email for pilot study

Dear Teachers,

My name is Ben Michael and I am a second year Trainee on the DClinPsy at Sheffield University. I am supervised by Steve Kellett and Jaime Delgadillo for my thesis project.

I am contacting you because I am in the final stages of developing my thesis study. When I was in the first stage PWP teachers, XXXX and XXXX provided incredibly helpful input. I am now hoping to get some more feedback before the experiment goes live. Specifically, I am interested in the face validity of my experiment and also regarding the usability of my online data collection system.

My thesis study will develop and test the utility of a 'real time' scenario-based clinical judgement decision-tree. This will assess therapist variability by evaluating clinical decision-making of PWPs. An effective method for assessing and capturing the process of clinical decision-making remains sparse. The methodology of my study will be an online survey design via a situational decision tree that PWPs will follow. To ensure ecological validity the decision tree will include typical decisions that a PWP is required to make during their on-going clinical practice. This could be useful in order to understand variability in the decisions that PWPs make in routine care.

Your participation would be greatly appreciated. Please click on the link below to access my study:

https://sheffieldpsychology.eu.qualtrics.com/jfe/form/SV_9TVnR9jsWMI4yFf

Following completion of the survey you are invited to provide feedback to me. This might be about the face validity of the situational decision tree, the time it takes to complete the survey online, the look and feel of the system, any technical issues that you may have encountered, or any other points that come to mind.

I look forward to hearing back from you.

Best Wishes,

Ben

Benjamin Michael
Trainee Clinical Psychologist

Appendix D: Semi-structured interview document

1. Is this case vignette example typical enough of a client that might present in IAPT at Step 2 or 3?

2. What would you change or add if critiquing the vignette?

3. What is the process of referral to see a PWP in IAPT at Step 2 or 3 like? Does the vignette accurately portray this?

4. In terms of risk, is the client presented in the vignette typical of what you might see at Step 2 or 3?

5. In terms of the client's history outlined in the vignette is there anything that you might change or add?

6. Does the unfolding scenario accurately depict how a client might engage with the therapist at Step 2 or 3?

7. Is the way that the therapist responding to the client realistic?

8. In terms of how they might work with the client, are the options the therapist has available to them realistic/accurate?

9. Are the decisions that the PWPs will face in the vignette realistic?

10. Are the intervention options (i.e. cognitive restructuring/behavioural activation) accurate?

11. Are the case management supervision conversations accurate?

Appendix E: Informed consent for IAPT teaching staff involved in focus group

Title of Research Project: Clinical judgement: An investigation of clinical decision-making

Name of Researchers: Benjamin Michael, supervised by Dr Stephen Kellett and Dr Jaime Delgadillo.

If you agree, please 'tick' each of the following statements.

- I confirm that I have read and understand the information for participants dated 6th March 2018 explaining the above research project and that I agree to take part in the research. ☐
- I understand that my participation is entirely voluntary, and that I am able to withdraw my participation and consent during the focus group. The responses will be anonymised upon submission; therefore, I understand that I will not be able to withdraw from the study after the focus group has taken place. ☐
- I am aware that the content pertaining to the focus group will be audio recorded and transcribed and I give my consent to this. ☐
- I understand that the information collected during this study will be kept strictly confidential. I give permission for members of the research team to have access to my anonymized responses. ☐
- I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research. ☐

Signed:

……

Date:

……

Appendix F: Excerpts from Living Document illustrating key changes made to dynamic measure

Appendix F1: CV1 (Version 3) and CV2 (Version 1)

Anchoring and adjustment refers to our common human tendency to base too much significance upon the first piece of information proposed (the anchor) when making a decision. Once an anchor is established all other judgements are considered relative to it and adjusted from it accordingly. (Tversky et al, 1974).

The halo effect is where first impressions influence later judgments. For example, a positive cognitive bias might be directed towards an individual who demonstrates a positive characteristic (Thorndike, 1920).

Experimental CV1- Anchoring and Halo Effect

Jack is 47 years old and lives with his wife and two daughters who are aged 11 and 7 years old. Jack is currently unemployed due to suddenly being made redundant from his job as a senior data analyst 6 months ago. He believed he would find a similar role straight away and so he took an analyst job that he feels is beneath him. Recently, he has been feeling really low, struggling with his energy levels and isolating himself from family and friends. He fears his career will come to little after years of striving. This has left Jack feeling hopeless and that his future is bleak in terms of employment opportunities. Jack's wife is a solicitor and Jack refers to her as a 'high flyer'. Jack describes his relationship with his wife as tolerable but that the 'spark' left their marriage years ago. He also reports fleeting thoughts of suicide but has not made any plans and denies any true intent. Jack enjoys maintaining and riding his motorbike, but recently has been using this more as an escape rather than pleasurable activity. Jack was prescribed fluoxetine by his GP 3 months ago and he continues to take this but does not report and particular benefits from taking the medication. This is the first time Jack has been referred for treatment and he has not experienced any significant episodes of low mood prior to his current presentation.

Please try to complete this next section with as little distraction as possible and in one go.
You will be presented with 3 scenarios. Each scenario has different response options for each decision. Please choose the response that is closest to how you would react. The scenarios do not have much information but please try to imagine yourself in the situations.

In this case anchoring and halo effects relate to the initial discovery that Jack has complicating features due to complex elements in his history. This information aims to elicit a first impression that participants will base too much significance upon (Asch, 1964). According to Kahneman (2011) subsequent information will be mostly wasted. It is proposed that participants who make the 'counter-normative' decision to refer Jack to step 3 will base this decision upon their less favourable first impression of Jack that he is 'complex'. This is despite also learning that Jack has read about CBT and the theory that underpins it; aiming to evoke a sense that Jack also has more favourable features (e.g. he is industrious and intelligent). This will be overlooked however due to anchoring and halo effects.

Session 1: Patient suitability for treatment.

You meet Jack for the first time for a screening appointment in order to assess his needs and suitability for a step 2 intervention. He tells you that he has experienced depression on and off for the last 12 years and when he was around 15 years old, he went through a period of self-harm. He explains that he has been drinking more alcohol than usual since losing his job as he now drinks one or two pints of beer daily. He also informs you that in the past he received counseling for his depression. He enjoys the 5-areas assessment and gains insight from this. The problem statement is: "When I am in my crappy new job or when I am alone, I start to feel really down. I think that nothing will change, and I lack the motivation to do anything about it. The impact of this is that I am becoming increasingly isolated and lonely." You collaboratively agree that his goal for the PWP work would be to start to see his friends again. Jack tells you that he has been reading self-help books about CBT prior to the appointment and he has a good understanding of the theory. You complete the PHQ-9 and he obtains a score of 14. You discuss specific treatment options and give him some psychoeducation on depression to read between sessions.

What would you do next?

- (Normative)I would class Jack's mental health problems as mild to moderate depression and plan to see him for 4-6 treatment sessions and take this to case management supervision.
- (Counter-normative) I would class Jack as struggling with complex depression requiring a higher intensity treatment. I would refer him to the step 3 pathway, end my involvement with him and take the decision to case management supervision. ***END OF DECISION TREE TASK.***

It is proposed that participants who make the 'counter-normative' decision to move from behavioural activation to cognitive restructuring are being influenced by the initial discovery (anchoring) that Jack has complicating features (halo effect). This means that their view of Jack's report that he is hopeless, dismissive, and hard to engage is being measured (adjustment) according to the anchor. It is proposed that 'counter-normative' decision makers will subsequently decide that his treatment requires some adaptation. This is despite Jack also demonstrating more favourable characteristics such as diligence and conscientiousness as he has read the psychoeducation material and also attempted the homework.

Session 2: Treatment fidelity.

After case management supervision, you have decided to see Jack fortnightly for 35-minute sessions and commence behavioral activation with him. At the second session you learn that Jack has read the psychoeducation you gave him about depression and has also attempted the homework you set him last week. Jack's PHQ-9 score is now 16 however and he is hopeless and hard to engage in the session. He is a little dismissive of the guided self-help workbook that supports the behavioural activation.

Which statement is closest to how you would react?

- (Counter-normative Response) I would introduce the principles of the COM-B model with him in order to think about barriers but also wonder if another CBT approach might be a better option. I would agree homework in line with cognitive restructuring and encourage him to keep a diary of his thoughts and feelings.
- (Normative Response) I would introduce the principles of the COM-B model with Jack in order to think about barriers. I would then agree homework in line with the principles of behavioural activation and set a task of meeting friends in the pub.

At your next group clinical supervision, you discuss that:

- (Normative) You need to concentrate on session structuring with him in terms of effective agendas.
- (Counter-normative) Jack's presentation is complex.

For those participants who make 'counter-normative' decisions evidence will accumulate gradually (Kahneman, 2011) throughout the 'hold or step up' scenario to continue seeing Jack despite no reliable improvement. These clinicians will feel a pull to continue treatment because Jack is complex (anchor) and they fear that his mental health will significantly decline if he faces long waiting lists for suitable treatments (Delgadillo, Gellatly & Stephenson-Bellwood, 2015). Signs to suggest that there is no reliable improvement include lack of motivation, the request for 'more time', and the fact that his PHQ-9 score remains at 16. It is proposed that the dominating belief that Jack is complex will override any subsequent information. For example, that due to your work he is becoming more psychologically aware but may benefit from a higher intensity treatment.

Session 3: Hold or step up.

Jack attends at session 3; you are trying to deliver the next stage of the intervention. He has not brought any homework to the session. You decide to go through his homework in the session and look at what got in the way and how he could overcome this for the following fortnight. His PHQ-9 score remains at 16. You agree homework of completing what he didn't complete last time and you give him COM-B literature to review in relation to barriers.

Which statement is closest to what you would do next?

- (Normative) I would plan to continue to see Jack but would use case management supervision to consider stepping him up as his scores aren't responding.
- (Counter-normative) I would plan to continue to see Jack and take him to case management supervision.

At your next supervision case management meeting you decide that:

- (Counter-normative) Jack's case only needs to be briefly discussed at the end of supervision.
- (Normative) Jack's case needs to take priority.

Session 4: Hold or step up.

Jack attends at session 4; he has attempted the homework and tells you that he has found this helpful. He notes that the homework and learning about the COM-B model has helped him to become more psychologically aware. When you introduce the next step of the treatment however Jack tells you that he needs more time. He tells you that talking to you is helpful because he feels he can trust you and that you understand. His PHQ-9 score remains at 16, however. You agree homework of continuing with the current treatment plan.

Which statement is closest to what you would do next?

- (Normative) I would plan to continue to see Jack but would use case management supervision to discuss the possibility of stepping him up, as his scores aren't responding.
- (Counter-normative) I would plan to continue to see Jack and take him to case management supervision in order to discuss continuing with the intervention and reviewing at a later date.

At your next supervision case management meeting you decide that:

- (Counter-normative) You will see Jack for two more sessions.
- (Normative) You will step him up after having a collaborative conversation with Jack about this.

END OF DECISION TREE TASK (award 2 points).

Session 5: Hold or step up.

After case management supervision, you have decided to see Jack for two more sessions. However, his PHQ-9 score still remains at 16. Despite this Jack has been regularly completing his homework and tells you that he feels that he is benefiting from your sessions. He wants to carry on seeing you. Which statement is closest to what you would do next?

- (Normative) I would organize another appointment to see Jack but would use case management supervision to discuss the possibility of stepping him up, as his scores still aren't responding.
- (Counter-normative) I would plan to continue to see Jack and take him to case management supervision. You are aware of the long waiting lists when referring him for a more intensive treatment at higher steps in the mental healthcare system. You fear that Jack's mental health will decline during this time.

At your next supervision case management meeting you decide that:

- (Counter-normative) You will see Jack for two more sessions.
- (Normative) You will step him up.

Control CV2

Chloe is 26 years old and lives with her husband and their 6-year-old daughter. Chloe works as an immigration appeals officer and reports persistent worries that she has not done her job properly. She is frequently concerned that someone might get deported as a result of her negligence. She often imagines the worst happening and states that when she worries, she feels sick, has headaches, feels butterflies in her stomach and is aware of her heart pounding. Chloe frequently gets hot and sweaty and says her anxiety makes it difficult to concentrate and do her job properly. Subsequently she often makes mistakes at work. Chloe struggles to play with her daughter due to feeling so anxious and worries about the effect her anxiety is having on her family. This leads her to feel low in mood. Chloe also currently believes that she is not good enough for her husband and that he deserves someone better. Subsequently she has experienced fleeting thoughts of suicide but has not made any plans and denies any true intent. Chloe began experiencing panic attacks 3 month ago, often on Sunday nights before going into work the next day. She went to see her GP who prescribed her sertraline for moderately severe depression and associated panic attacks. The sertraline has been effective in helping to reduce Chloe's panic attacks, but her anxiety and low mood remain. She is otherwise physically fit and well and is not prescribed any other form of medication.

In this case anchoring and halo effects relate to the fact that Chloe presents with difficulties that are 'typically common' for step 2 of primary care. As in the experimental condition, this information aims to elicit a first impression (particularly in those participants who predominantly gave 'counter-normative' responses in the experimental condition) that participants will base too much significance upon (Asch, 1964). It is proposed that regardless as to whether participants have made mainly 'normative' or 'counter-normative' decisions in the experimental condition they will all predominantly make 'normative' decisions in the case of Chloe.

Session 1: Patient suitability for treatment.

You meet Chloe for the first time for a screening appointment in order to assess her needs and suitability for a step 2 intervention. At the start of the consultation Chloe states that she is attending due to problems with worry. After questioning about how things have been for her recently, Chloe discloses she is feeling under considerable stress. Chloe discloses that she has anxiety upon waking which stays with her throughout the day. She feels like her head is going to explode and her heart will jump out of her chest. She feels overwhelmed with fear, cannot work properly and cannot play with her daughter. You introduce the 5-areas assessment and she engages well with this process. The problem statement is: "When I am at work or when I am at home and think about work, I start to feel really worried. In my job despite how hard I try I often think that I have failed my clients. I fear that I am good to nobody and this makes me feel hopeless and low in mood. The impact of this is that I am becoming increasingly low in mood and I believe that my family is suffering as a result." You collaboratively agree that her goal for the PWP work would be to start to worry less about her job. You complete the GAD-7 and she obtains a score of 13. You discuss specific treatment options with Chloe and give her some psychoeducation on anxiety to read between sessions.

What would you do next?

- (Normative) I would class Chloe's mental health problems as moderate anxiety and plan to see her for 4-6 treatment sessions and take this to case management supervision.
- (Counter-normative) I would class Chloe as struggling with complex anxiety and depression requiring a higher intensity treatment. I would refer her to the step 3 pathway, end my involvement with her and take the decision to case management supervision. ***END OF DECISION TREE TASK.***

It is proposed that despite the fact that at session 2 Chloe demonstrates less favourable traits (e.g. she is unenthusiastic and distractible) the majority of participants will make the 'normative' decision to continue with the self-help resource and remain aligned with the treatment protocol for treating GAD. This will also be despite the fact that her GAD-7 score is getting worse.

Session 2: Treatment fidelity.

Chloe's GAD-7 score and her background information point to a diagnosis of generalised anxiety disorder (GAD). After case management supervision, you have decided to see Chloe fortnightly for 35-minute sessions to commence individual guided self-help with her. At the second session you learn that Chloe has not read the psychoeducation you gave her about anxiety and her GAD-7 score is now 15. She is very apologetic and tells you that she recently learnt that one of her clients at work had been deported last week and this has caused Chloe significant distress. Chloe has been off work as a result and tells you in great detail about her own financial difficulties.

Which statement is closest to how you would react?

- (Counter-normative Response) I would empathise with Chloe and offer her the opportunity to speak further about her financial concerns if she needs to. Chloe needs to feel sufficiently validated and understood before I can support her to use the self-help resource. I would agree homework encouraging Chloe to engage in some self-care at home.
- (Normative Response) I would empathise with Chloe but explain that my role is to guide and support Chloe's use of the self-help resource and monitor and review the process and outcome of treatment. I would then introduce the resource and agree homework of identifying unhelpful thoughts.

At your next group clinical supervision, you discuss that:

- (Normative) You need to concentrate on supporting Chloe to find ways to understand, manage or overcome her anxiety using the self-help resource.
- (Counter-normative) Chloe's presentation is complex and may require adaptations to be made to the intervention.

Once again, evidence will accumulate gradually (Kahneman, 2011) throughout scenario 3. This time however it will be in relation to stepping Chloe up at the appropriate point as there is no reliable improvement by session 4 (Delgadillo et al., 2014). It is proposed that the majority of clinicians will not hold Chloe and instead will follow the appropriate procedures in order to step her up. Where the halo effect occurs participants will conclude that Chloe's symptoms are in line with GAD (anchor) but require a higher-intensity psychological intervention (adjustment). This will be despite Chloe's request that you continue working with her. Participants will also overlook that Chloe is demonstrating similar 'complex characteristics' to Jack in the experimental condition. Those participants who are less influenced by heuristics and biases will follow treatment protocol anyway which would also suggest offering Chloe a higher-intensity treatment as her scores are not improving.

Session 3: Hold or step up.

Chloe attends at session 3; you are trying to deliver the next stage of the intervention. She has been able to engage in the homework that you set her during the last session, but her GAD-7 score remains at 15 and she has doubts about her anxiety improving. She discloses that she has experienced anxiety and depression on and off since her mother died when Chloe was 15. Around this time Chloe went through a phase of restricting what she ate. She explains that recently she has been using marijuana occasionally in order to manage her anxiety which has made her forgetful and contributed to her making mistakes at work.

Which statement is closest to what you would do next?

- (Counter-normative) I would plan to continue to see Chloe and use the COM-B model to discuss Chloe's doubts and support her in having a sense of herself as someone who can make the change.
- (Normative) I would plan to continue to see Chloe but would use case management supervision to consider stepping her up, as her scores aren't responding.

At your next supervision case management meeting you decide that:

- (Normative) Chloe's case needs to take priority.
- (Counter -normative) Chloe's case only needs to be briefly discussed at the end of supervision.

Session 4: Hold or step up.

Chloe attends at session 4; she has attempted the homework and tells you that she has found this helpful. When you introduce the next step of the treatment however Chloe becomes tearful and tells you that she needs more time. She tells you that talking to you is helpful because she feels she can trust you and that you understand. His GAD-7 score remains at 15, however. You agree homework of continuing with the current treatment plan.

Which statement is closest to what you would do next?

- (Counter-normative) I would plan to continue to see Chloe and take her to case management supervision in order to discuss continuing with the intervention and reviewing at a later date.
- (Normative) I would plan to continue to see Chloe but would use case management supervision to discuss the possibility of stepping her up, as her scores aren't responding.

At your next supervision case management meeting you decide that:

- (Normative) You will step her up after having a collaborative conversation with Chloe about this.

  END OF DECISION TREE TASK (award 2 points).
- (Counter-normative) You will see Chloe for two more sessions.

Session 5: Hold or step up.

After case management supervision, you have decided to see Chloe for two more sessions. However, her GAD-7 score still remains at 15. Despite this Chloe tells you that she feels that she is benefiting from your sessions and she wants to carry on seeing you.

Which statement is closest to what you would do next?

- (Counter-normative) I would plan to continue to see Chloe and take her to case management supervision.
- (Normative) I would organize another appointment to see Chloe but would use case management supervision to discuss the possibility of stepping her up, as her scores still aren't responding.

At your next supervision case management meeting you decide that:

- (Normative) You will step her up.
- (Counter-normative) You will see Chloe for two more sessions.

Appendix F2: CV1 and CV2 Final Version

Anchoring and adjustment refers to our common human tendency to base too much significance upon the first piece of information proposed (the anchor) when making a decision. Once an anchor is established all other judgements are considered relative to it and adjusted from it accordingly. (Tversky et al, 1974).

The halo effect is where first impressions influence later judgments. For example, a positive cognitive bias might be directed towards an individual who demonstrates a positive characteristic (Thorndike, 1920).

Experimental and Control Case Vignettes - Jack

Jack is 47 years old and lives with his wife and two daughters who are aged 11 and 7 years old. Jack is currently unemployed due to suddenly being made redundant from his job as a senior data analyst 6 months ago. He believed he would find a similar role straight away and so he took an analyst job that he feels is beneath him. Recently, he has been feeling really low, struggling with his energy levels and isolating himself from family and friends. He fears his career will come to little after years of striving. This has left Jack feeling hopeless and that his future is bleak in terms of employment opportunities. Jack's wife is a solicitor and Jack refers to her as a 'high flyer'. Jack describes his relationship with his wife as essentially fine, but that the 'spark' left their marriage years ago.
He also reports fleeting thoughts of suicide but has not made any plans and denies any true intent. Jack enjoys maintaining and riding his motorbike, but recently has been using this more as a distraction rather than pleasurable activity. Jack was prescribed fluoxetine by his GP 3 months ago and he continues to take this but does not report any particular benefits from taking the medication.

Please try to complete this next section with as little distraction as possible and in one go.
You will be presented with 3 scenarios. Each scenario has different response options for each decision. Please choose the response that is closest to how you would react. The scenarios do not have much information but please try to imagine yourself in the situations.

In this case anchoring and halo effects relate to the initial discovery that Jack has complicating features due to complex elements in his history. This information aims to elicit a first impression that participants will base too much significance upon (Asch, 1964). According to Kahneman (2011) subsequent information will be mostly wasted. It is proposed that participants who make the 'counter-normative' decision to 'hold' Jack at step 2 will base this decision upon their first impression of Jack that he is 'complex'. This is despite also learning that Jack has read about CBT and the theory that underpins it; aiming to evoke a sense that Jack also has more favourable features (e.g. he is industrious and intelligent). This will be overlooked however due to anchoring and halo effects.

Experimental CV1 - Anchoring and Halo Effect

Session 1: Patient suitability for treatment.

You meet Jack for the first time for a screening appointment in order to assess his needs and suitability for a step 2 intervention. At the start of the consultation Jack states that he is attending due to problems with low mood and low energy. Jack explains that he has been isolating himself from family and friends. He describes feeling hopeless and that his future feels bleak. He tells you that he has experienced depression on and off for the last 12 years and when he was around 15 years old, he went through a period of self-harm for two years which involved some superficial cutting of his legs and arms. He explains that he has been drinking more alcohol than usual since losing his job as he now drinks one or two pints of beer daily and has an intended binge about once a month. He also informs you that in the past he received counseling for his depression. He enjoys the 5-areas assessment and gains insight from this. The problem statement is: "When I am in my crappy new job or when I am alone, I start to feel really down. I think that nothing will change, and I lack the motivation to do anything about it. The impact of this is that I am becoming increasingly isolated and lonely." You collaboratively agree that his goal for the PWP work would be to start to see his friends again. Jack tells you that he has been reading self-help books about CBT prior to the appointment and he has a good understanding of the theory. You complete the PHQ-9; he scores 14. You discuss specific treatment options and give him some psychoeducation on depression to read between sessions.

What would you do next?

- (Normative) I would plan to see Jack for 4-6 treatment sessions.
- (Counter-normative) I would class Jack as requiring a higher intensity treatment. I would refer him to the step 3 pathway. ***END OF DECISION TREE TASK.***

It is proposed that participants who make the 'counter-normative' decision to change treatment from behavioural activation to cognitive restructuring are being influenced by the initial discovery (anchoring) that Jack has complicating features (halo effect). This means that the suggestion that he is easily annoyed, prickly, and appears irritable is being measured (adjustment) according to the anchor. It is proposed that 'counter-normative' decision makers will subsequently decide that his treatment requires some adaptation. This is despite Jack demonstrating favourable characteristics such as motivation as he has read the psychoeducation material you set him as homework and that there is also a SMART behavioural goal.

Session 2: Treatment fidelity.

After case management supervision, you have decided to commence behavioral activation (BA). At the second session you learn that Jack has read the psychoeducation you gave him about depression last week. Jack's PHQ-9 score has increased to 17. He appears more easily annoyed and somewhat prickly in the session in terms of how guided self-help can make a difference . He appears irritable when you introduce the BA self-help workbook and flicks through it a little dismissively. He says to you "I always ruin everything; this will not work."

Choose your next step:

- (Counter-normative Response) I would explore the barriers to the work and introduce cognitive restructuring.
- (Normative Response) I would explore the barriers to the work and continue with behavioural activation.

For those participants who make 'counter-normative' decisions in the 'hold or step up' scenario evidence will have accumulated gradually (Kahneman, 2011) to continue seeing Jack despite no reliable improvement. These clinicians will feel a pull to continue treatment because Jack is complex (anchor) and they fear that his mental health will significantly decline if he faces long waiting lists for suitable treatments (Delgadillo, Gellatly & Stephenson-Bellwood, 2015). Signs to suggest that there is no reliable improvement include the fact that Jack feels he needs more time and that his PHQ-9 score remains at 16. It is proposed that the dominating belief that Jack is complex will override any subsequent information. For example, that due to the PWPs work he is becoming more psychologically aware but may benefit from a higher intensity treatment.

Session 4: Hold or step up.

Jack attends at session 4. He is easily distractible during the session. He has not really engaged with the homework and states that he is struggling to manage his drinking. He talks about ruminating about his childhood in the week and seems to want you to listen to this. Overtime you have come to like and empathise with Jack and he tells you that talking to you is helpful but that he needs more time to change. His PHQ-9 score remains at 17 (compared to initial score of 14).

Choose your next step:

- (Normative) I would step Jack up as his scores aren't responding.
- (Counter-normative) I would plan to continue to see Jack, because he is finding it helpful.

END OF DECISION TREE TASK

In this case anchoring and halo effects are not being tested and Jack presents with difficulties that are 'typically common' for step 2 of primary care. It is proposed that regardless as to whether participants have made mainly 'normative' or 'counter-normative' decisions in the experimental condition they will all predominantly make 'normative' decisions in the control version of Jack.

Control CV1 - Anchoring and Halo Effect

Session 1: Patient suitability for treatment.

You meet Jack for the first time for a screening appointment in order to assess his needs and suitability for a step 2 intervention. At the start of the consultation Jack states that he is attending due to problems with low mood and low energy. Jack explains that he has been isolating himself from family and friends. He describes feeling hopeless and that his future feels bleak. He enjoys the 5-areas assessment and gains insight from this. The problem statement is: "When I am in my crappy new job or when I am alone, I start to feel really down. I think that nothing will change, and I lack the motivation to do anything about it. The impact of this is that I am becoming increasingly isolated and lonely." You collaboratively agree that his goal for the PWP work would be to start to see his friends again. Jack tells you that he has been reading self-help books about CBT prior to the appointment and he has a good understanding of the theory. You complete the PHQ-9; he scores 14. You discuss specific treatment options and give him some psychoeducation on depression to read between sessions.

What would you do next?

- (Normative) I would plan to see Jack for 4-6 treatment sessions.
- (Counter-normative) I would class Jack as requiring a higher intensity treatment. I would refer him to the step 3 pathway. ***END OF DECISION TREE TASK.***

Session 2: Treatment fidelity.

After case management supervision, you have decided to commence behavioral activation (BA). At the second session you learn that Jack has read the psychoeducation you gave him about depression last week. Jack's PHQ-9 score has increased to 17. You introduce the BA self-help workbook.

Choose your next step:

- (Counter-normative Response) I would explore the barriers to the work and introduce cognitive restructuring.
- (Normative Response) I would explore the barriers to the work and continue with behavioural activation

Session 4: Hold or step up.

Jack attends at session 4. Overtime you have come to like and empathise with Jack and he tells you that talking to you is helpful but that he needs more time to change. His PHQ-9 score remains at 17 (compared to initial score of 14).

Choose your next step:

- (Normative) I would step Jack up as his scores aren't responding.
- (Counter-normative) I would plan to continue to see Jack, because he is finding it helpful.

END OF DECISION TREE TASK

Experimental and Control Case Vignettes – Chloe

Chloe is 26 years old and lives with her husband and their 6-year-old daughter. Chloe works as an immigration appeals officer and reports persistent anxiety that she has not done her job correctly. She is frequently concerned that someone might get deported as a result of her negligence. She often imagines the worst happening and states that when she has anxiety attacks, she feels sick, has headaches, feels butterflies in her stomach and is aware of her heart pounding. Chloe frequently gets hot and sweaty and says her anxiety makes it difficult to concentrate and do her job properly. Subsequently she often makes mistakes at work. Chloe struggles to properly engage with her daughter due to feeling so anxious and worries about the effect her anxiety is having on her family. This leads her to feel low in mood. Chloe also currently believes that she is not good enough for her husband and that he deserves someone better. Subsequently she has experienced fleeting thoughts of suicide, but has not made any plans, denies any true intent and could never make her family suffer such a loss. Chloe began experiencing panic attacks 3 month ago, often on Sunday nights before going into work the next day. She went to see her GP who prescribed her sertraline for moderately severe depression and associated panic attacks. The sertraline has been effective in helping to reduce Chloe's panic attacks, but her anxiety and low mood remain. She is otherwise physically fit and well and is not prescribed any other form of medication.

As with the experimental version of Jack, in the experimental version of Chloe anchoring and halo effects relate to the initial discovery that Chloe has complicating features due to complex elements in her history. This information aims to elicit a first impression that participants will base too much significance upon (Asch, 1964). According to Kahneman (2011) subsequent information will be mostly wasted. It is proposed that participants who make the 'counter-normative' decision to 'hold' Chloe at step 2 will base this decision upon their first impression of her that she is 'complex'. This is despite also learning that often works extra hours. This will be overlooked however due to anchoring and halo effects.

Experimental CV2 - Anchoring and Halo Effect

Session 1: Patient suitability for treatment.

You meet Chloe for the first time for a screening appointment in order to assess her needs and suitability for a step 2 intervention. At the start of the consultation Chloe states that she is attending due to problems with worry. After questioning about how things have been for her recently, Chloe discloses she is feeling under considerable stress. Chloe discloses that she has anxiety upon waking which stays with her throughout the day. She feels like her head is going to explode and her heart will jump out of her chest. She feels overwhelmed with fear, cannot work properly and cannot play with her daughter. She has experienced anxiety and depression on and off since her mother died when Chloe was 15. Around this time Chloe went through a phase of restricting what she ate, and she was bullied at school for being scrawny and aloof. She recalls that she self-harmed at this time. She explains that recently she has been using marijuana occasionally in order to manage her anxiety which has made her forgetful and contributed to her making mistakes at work. You introduce the 5-areas assessment and she engages well with this process. The problem statement is: "When I am at work or when I am at home and think about work, I start to feel really worried. In my job despite how hard I try I often think that I have failed my clients. I fear that I am good to nobody and this makes me feel hopeless and low in mood. The impact of this is that I am becoming increasingly low in mood and I believe that my family is suffering as a result." Chloe tells you she often works extra hours and, "Anyone else would do the same for another's wellbeing". You complete the GAD-7 and she scores 13. You discuss specific treatment options with Chloe and give her some psychoeducation on anxiety to read between sessions.

Choose your next step:

- (Normative) I would plan to see Chloe for 4-6 guided self-help treatment sessions.
- (Counter-normative) I would class Chloe as requiring a higher intensity treatment. I would refer her to the step 3 pathway for counseling.

END OF DECISION TREE TASK.

It is proposed that participants who make the 'counter-normative' decision to move from guided self-help to encouraging Chloe to engage purely in self-care are being influenced by the initial discovery (anchoring) that Chloe has complicating features (halo effect). This means that the suggestion that she is cold and standoffish is being measured (adjustment) according to the anchor. It is proposed that 'counter-normative' decision makers will subsequently decide that her treatment requires some adaptation. This is despite Chloe also demonstrating more favourable characteristics such as becoming distressed that a client has been deported and not recognise her own caring nature (e.g. she is compassionate and modest). This will be overlooked however due to anchoring and halo effects.

Session 2: Treatment fidelity.

After case management supervision, you have decided to see Chloe fortnightly for 35-minute sessions to commence individual guided self-help with her. At the second session Chloe is very distressed and tearful. Her GAD-7 score is now 15. She tells you that she recently learnt that one of her clients at work had been deported last week and this has caused Chloe to worry. Chloe has been off work as a result and tells you in some detail about her own financial difficulties.

Choose your next step:

- (Counter-normative Response) I would prioritise listening and empathising in this session in order to cement the alliance and prescribe some self-care time as homework.
- (Normative Response) I would complete a 5-areas of the work situation and provide a worry awareness diary to complete as homework.

For those participants who make 'counter-normative' decisions in the 'hold or step up' scenario evidence will have accumulated gradually (Kahneman, 2011) to continue seeing Chloe despite no reliable improvement. These clinicians will feel a pull to continue treatment because Chloe is complex (anchor) and they fear that her mental health will significantly decline if she faces long waiting lists for suitable treatments (Delgadillo, Gellatly & Stephenson-Bellwood, 2015). Signs to suggest that there is no reliable improvement include the fact that Chloe has doubts her anxiety will improve and that her GAD-7 score remains at 15. It is proposed that the dominating belief that Chloe is complex will override any subsequent information. For example, that due to your work she is becoming more psychologically aware but may benefit from a higher intensity treatment.

Session 4: Hold or step up.

Chloe attends at session 4; you are trying to deliver the next stage of the intervention. She has been able to engage in the homework that you set her during the last session, but her GAD-7 score remains at 15 and she has doubts about her anxiety improving. Chloe states that the sessions with you are a 'lifeline' and she is able to really confide in you. She continues to use marijuana occasionally in order to manage her anxiety.

Which statement is closest to what you would do next?

- (Counter-normative) I would plan to continue to see Chloe and use the COM-B model to discuss Chloe's doubts and support her in having a sense of herself as someone who can make the change.
- (Normative) I would organise another appointment to see Chloe but would use case management supervision to discuss the possibility of stepping her up, as her scores still aren't responding.

In this case anchoring and halo effects are not being tested and Chloe presents with difficulties that are 'typically common' for step 2 of primary care. It is proposed that regardless as to whether participants have made mainly 'normative' or 'counter-normative' decisions in the experimental condition they will all predominantly make 'normative' decisions in the control version of Chloe.

Control CV2 - Anchoring and Halo Effect

Session 1: Patient suitability for treatment.

You meet Chloe for the first time for a screening appointment in order to assess her needs and suitability for a step 2 intervention. At the start of the consultation Chloe states that she is attending due to problems with worry. After questioning about how things have been for her recently, Chloe discloses she is feeling under considerable stress. Chloe discloses that she has anxiety upon waking which stays with her throughout the day. She feels like her head is going to explode and her heart will jump out of her chest. She feels overwhelmed with fear, cannot work properly and cannot play with her daughter. You introduce the 5-areas assessment and she engages well with this process. The problem statement is: "When I am at work or when I am at home and think about work, I start to feel really worried. In my job despite how hard I try I often think that I have failed my clients. I fear that I am good to nobody and this makes me feel hopeless and low in mood. The impact of this is that I am becoming increasingly low in mood and I believe that my family is suffering as a result." Chloe tells you she often works extra hours and, "Anyone else would do the same for another's wellbeing". You complete the GAD-7 and she scores 13. You discuss specific treatment options with Chloe and give her some psychoeducation on anxiety to read between sessions.

Choose your next step:

- (Normative) I would plan to see Chloe for 4-6 guided self-help treatment sessions
- (Counter-normative) I would class Chloe as requiring a higher intensity treatment. I would refer her to the step 3 pathway for counseling.

END OF DECISION TREE TASK.

Session 2: Treatment fidelity.

After case management supervision, you have decided to see Chloe fortnightly for 35-minute sessions to commence individual guided self-help with her. At the second session her GAD-7 score is now 15. She tells you that she recently learnt that one of her clients at work had been deported last week and this has caused Chloe to worry. Chloe has been off work as a result and tells you in some detail about her own financial difficulties.

Choose your next step:

- (Counter-normative Response) I would prioritise listening and empathising in this session in order to cement the alliance and prescribe some self-care time as homework.
- (Normative Response) I would complete a 5-areas of the work situation and provide a worry awareness diary to complete as homework.

Session 4: Hold or step up.

Chloe attends at session 4; you are trying to deliver the next stage of the intervention. She has been able to engage in the homework that you set her during the last session, but her GAD-7 score remains at 15 and she has doubts about her anxiety improving. Chloe states that the sessions with you are a 'lifeline' and she is able to really confide in you.

Which statement is closest to what you would do next?

- (Counter-normative) I would plan to continue to see Chloe and use the COM-B model to discuss Chloe's doubts and support her in having a sense of herself as someone who can make the change.
- (Normative) I would organise another appointment to see Chloe but would use case management supervision to discuss the possibility of stepping her up, as her scores still aren't responding.

Appendix G. Inter-rater reliability scoring process.

A two point scale was employed to generate an agreement score (e.g. 0 = absent, 1 = present, 99 = N/A). Inter-rater agreement regarding the codes, themes and sub-themes was calculated. Percentage agreement and Krippendorff's alpha were used to calculate agreement. An online utility called ReCal2 ("Reliability Calculator for 2 coders") was used to compute intercoder/interrater reliability coefficients for nominal data coded by two coders. Prior to a meeting held between coders inter-rater percentage agreement was 93.5% but Krippendorff's alpha was only -0.02. Proposed acceptable levels of inter-rater agreement range from 0.70 - 0.80 (Davis, 1992; Selby-Harrington, Mehta, Jutsum, Riportella-Muller, & Quade, 1994). Following the meeting two contentious items were recoded resulting in an improved percentage agreement score of 97.8% and improved Krippendorff's alpha of 0.79.

Those numbers in **red** indicate disagreement amongst raters that was subsequently resolved. Those numbers in **blue** indicate disagreement amongst raters that was not resolved. Two contentious items were recoded. Subcategory 5: 'client preconceptions' moved from **category 1: 'client suitability', Subcategory: 4** to category 2: 'Accurately portraying a collaborative approach', Subcategory 5. This meant that code **pre meeting: 8'** also moved to category 2, final code: 10. **Final code: 29** remained unchanged after the meeting between the two coders despite disagreement regarding its coding. This was because inter-rater agreement according to Krippendorff's alpha was now acceptable despite this discrepancy.

See Appendix H for the coding key which lists which phase, stage, code, theme, category and subcategory corresponds with which number.

Phase	Stage	Preliminary code	Initial code	Potential theme	Category	Subcategory pre meeting	Code pre meeting	Subcategory post meeting	Final code	Rater 1 score Pre meeting	Rater 2 score Pre meeting	Rater 1 score Post meeting	Rater 2 score Post meeting
1													
		1								1	1	1	1
		2								1	1	1	1
		3								1	1	1	1
		4								1	1	1	1
2													
	1												
			1							1	1	1	1
			2							1	1	1	1
			3							1	1	1	1
			4							1	1	1	1
			5							1	1	1	1

Phase	Stage	Preliminary code	Initial code	Potential theme	Category	Subcategory pre meeting	Code pre meeting	Subcategory post meeting	Final code	Rater 1 score Pre meeting	Rater 2 score Pre meeting	Rater 1 score Post meeting	Rater 2 score Post meeting
			6							1	1	1	1
			7							1	1	1	1
			8							1	1	1	1
			9							1	1	1	1
			10							1	1	1	1
			11							1	1	1	1
			12							1	1	1	1
			13							1	1	1	1
			14							1	1	1	1
			15							1	1	1	1
			16							1	1	1	1
			17							1	1	1	1

Phase	Stage	Preliminary code	Initial code	Potential theme	Category	Subcategory pre meeting	Code pre meeting	Subcategory post meeting	Final code	Rater 1 score Pre meeting	Rater 2 score Pre meeting	Rater 1 score Post meeting	Rater 2 score Post meeting
			18							1	1	1	1
			19							1	1	1	1
			20							1	1	1	1
			21							1	1	1	1
			22							1	1	1	1
			23							1	1	1	1
			24							1	1	1	1
			25							1	1	1	1
			26							1	1	1	1
			27							1	1	1	1
			28							1	1	1	1
			29							1	1	1	1

Phase	Stage	Preliminary code	Initial code	Potential theme	Category	Subcategory pre meeting	Code pre meeting	Subcategory post meeting	Final code	Rater 1 score Pre meeting	Rater 2 score Pre meeting	Rater 1 score Post meeting	Rater 2 score Post meeting
			30							1	1	1	1
			31							1	1	1	1
	2												
				1						1	1	1	1
				2						1	1	1	1
				3						1	1	1	1
				4						1	1	1	1
				5						1	1	1	1
				6						1	1	1	1
				7						1	1	1	1
				8						1	1	1	1

Phase	Stage	Preliminary code	Initial code	Potential theme	Category	Subcategory pre meeting	Code pre meeting	Subcategory post meeting	Final code	Rater 1 score Pre meeting	Rater 2 score Pre meeting	Rater 1 score Post meeting	Rater 2 score Post meeting
				9						1	1	1	1
				10						1	1	1	1
				11						1	1	1	1
				12						1	1	1	1
	3			13						1	1	1	1
3													
					1					1	1	1	1
						1		1		1	1	1	1
							1		1	1	1	1	1
							2		2	1	1	1	1
						2		2		1	1	1	1

Phase	Stage	Preliminary code	Initial code	Potential theme	Category	Subcategory pre meeting	Code pre meeting	Subcategory post meeting	Final code	Rater 1 score Pre meeting	Rater 2 score Pre meeting	Rater 1 score Post meeting	Rater 2 score Post meeting
							3		3	1	1	1	1
							4		4	1	1	1	1
							5		5	1	1	1	1
						3		3		1	1	1	1
							6		6	1	1	1	1
							7		7	1	1	1	1
						4				**1**	**0**	**0**	**0**
							8			**1**	**0**	**0**	**0**
					2					1	1	1	1
						5		4		1	1	1	1
							9		8	1	1	1	1

Phase	Stage	Preliminary code	Initial code	Potential theme	Category	Subcategory pre meeting	Code pre meeting	Subcategory post meeting	Final code	Rater 1 score Pre meeting	Rater 2 score Pre meeting	Rater 1 score Post meeting	Rater 2 score Post meeting
							10		9	1	1	1	1
						6		5		1	1	1	1
							11		10	1	1	1	1
					3					1	1	1	1
						7		6		1	1	1	1
							12		11	1	1	1	1
							13		12	1	1	1	1
							14		13	1	1	1	1
						8		7		1	1	1	1
							15		14	1	1	1	1
							16		15	1	1	1	1

Phase	Stage	Preliminary code	Initial code	Potential theme	Category	Subcategory pre meeting	Code pre meeting	Subcategory post meeting	Final code	Rater 1 score Pre meeting	Rater 2 score Pre meeting	Rater 1 score Post meeting	Rater 2 score Post meeting
							17		16	1	1	1	1
						9		8		1	1	1	1
							18		17	1	1	1	1
							19		18	1	1	1	1
							20		19	1	1	1	1
					4					1	1	1	1
						10		9		1	1	1	1
							21		20	1	1	1	1
							22		21	1	1	1	1
							23		22	1	1	1	1
							24		23	1	1	1	1

Phase	Stage	Preliminary code	Initial code	Potential theme	Category	Subcategory pre meeting	Code pre meeting	Subcategory post meeting	Final code	Rater 1 score Pre meeting	Rater 2 score Pre meeting	Rater 1 score Post meeting	Rater 2 score Post meeting
							25		24	1	1	1	1
							26		25	1	1	1	1
							27		26	1	1	1	1
							28		27	1	1	1	1
						11		10		1	1	1	1
							29		28	1	1	1	1
							30		**29**	**1**	**0**	**1**	**0**
						12		11		1	1	1	1
							31		30	1	1	1	1
							32		31	1	1	1	1

Appendix H. Coding Process Thematic Analysis.

No.	Phase	Preliminary Code	Initial Code	Potential theme	Category	Subcategory	Final Code
1	Familiarising self with data	Suitability	Risk status of Jack realistic for IAPT.	Ambivalence	Client Suitability	Risk status	Risk status of Jack realistic for IAPT.
2	Generation of initial codes	Motivation	Risk status sounds suitable for standard patient.	Barriers	Accurately portraying a collaborative approach	Referral and screening process	Risk status sounds suitable for standard patient.
3	Searching for themes	Treatment selection	Screening assessment process for treatment suitability	Client pre-conceptions	Accurately depicting process of selecting treatment	Motivation	Referral
4		Ecological validity	'Jack's perspective re. wait time and expectancy to see a counselor 'bang on'.	Collaborating with the client	Ecological Validity	Setting goals with client	Screening assessment process for treatment suitability.
5			Third option – private counseling might be offered.	Decision tree scoring		Client Preconceptions[6]	'Jack's perspective re. wait time and expectancy to see a counselor 'bang on'.
6			Setting collaborative goal would only be done if Jack onboard with treatment (motivated) and presenting with mild to moderate depression.	Evidence-base		Therapist decision-making	Third option – private counseling might be offered.

[6] Was Category 1: 'client suitability'; Subcategory: 4 prior to rater meeting.

No.	Phase	Preliminary Code	Initial Code	Potential theme	Category	Subcategory	Final Code
7			More dialogue with client in an actual assessment as to their motivation right from beginning.	Motivation		Barriers	Buying into the model.
8			Aim is for Jack to sound dubious but will have a go.	PWP Characteristics		Decision tree scoring	Setting collaborative goal would only be done if Jack onboard with treatment (motivated) and presenting with mild to moderate depression.
9			Some expectation that client might not have been given all the correct information. This might be why dubious and therefore would be seen as understandable by PWP.	Realism of vignette		Realism of vignette	First intervention should be in line with goals and so would start with BA.
10			First intervention should be in line with goals and so would start with BA.	Referral and screening process		Evidence base	Some expectation that client might not have been given all the correct information. This might be why dubious and therefore would be seen as understandable by PWP[7].
11			You do swap around; pressure gets to you.	Risk Status		PWP Characteristics	Easier to see change if you start with BA.
12			Therapy does not always go to plan – clear cut.	Setting goals with the client			You do swap around; pressure gets to you. Therapy does not always go to plan – clear cut.
13			Client may not initially understand CBT concepts.	Therapist decision-making			Influence of client sociodemographic (e.g. client's lower social status might deter PWP from pursuing treatment and keep trying with client).

[7] Was 'Code pre meeting: 8' prior to meeting amongst reviewers.

No.	Phase	Preliminary Code	Initial Code	Potential theme	Category	Subcategory	Final Code
14			Weave in more about PWP thinking about barriers as per COM-B approach with client.				Client may not initially understand CBT concepts.
15			More background about what might be making things worse for Jack. E.g. fact that PWP has given him 40 pages to read of info when motivation low!				Weave in more about PWP thinking about barriers as per COM-B approach with client.
16			Continuity issues raised that do not allude to either intervention so that right or wrong option selection not highlighted.				More background about what might be making things worse for Jack. E.g. fact that PWP has given him 40 pages to read of info when motivation low!
17			Discussions about making content more realistic and richer such as doing homework in session and add more info in about what got in the way of not doing homework.				Clarification as to what treatment content would look like to make incongruent with either behavioural activation approach or cognitive restructuring. Keep it vague enough but still following a thread of something.
18			Also look at applying COM-B approach to look at barriers.				Discussion about wording in the text so that it does not align with any one approach specifically.
19			Try not to be too specific re. certain approaches (e.g. problem solving) as this begins to sound like going down a particular intervention route.				Also realistic about what might lead PWPs to be suspicious of selecting right and wrong answers (e.g. get rid of relapse prevention option).
20			More detail about exactly what would be offered if stepped up (e.g. counseling or more CBT).				Frequency of therapy.

No.	Phase	Preliminary Code	Initial Code	Potential theme	Category	Subcategory	Final Code
21			Factor in process you would go through of reviewing their progress and motivation in sessions with client before discussing stepping up in case management supervision.				Continuity issues raised that do not allude to either intervention so that right or wrong option selection not highlighted.
22			Collaborative decision with client is the aim.				Discussions about making content more realistic and richer such as doing homework in session and add more info in about what got in the way of not doing homework.
23			Pull to offer more discussed if client is saying how much they appreciate 'chats' with PWP despite scores not responding to treatment.				Try not to be too specific re. certain approaches (e.g. problem solving) as this begins to sound like going down a particular intervention route.
24			Also questions as to what approach might be best re. allocation at step 3? Might not be relevant, however.				More detail about exactly what would be offered if stepped up (e.g. counseling or more CBT).
25			Clarification as to what treatment content would look like to make it congruent with either behavioural activation approach or cognitive restructuring. Keep it vague enough but still following a thread of something.				Factor in process you would go through of reviewing their progress and motivation in sessions with client before discussing stepping up in case management supervision. Collaborative decision with client is the aim.
26			Discussion about wording in the text so that it does not align with any one approach specifically.				Pull to offer more discussed if client is saying how much they appreciate 'chats' with PWP despite scores not responding to treatment.
27			Also realistic about what might lead PWPs to be suspicious of selecting right and wrong answers (e.g. get rid of relapse prevention option).				Also questions as to what approach might be best re. allocation at step 3? Might not be relevant, however.

No.	Phase	Preliminary Code	Initial Code	Potential theme	Category	Subcategory	Final Code
28			Feedback that vignette flows well. Compromise between what is realistic versus what is in line with literature (e.g. suggesting client seeks alternative support elsewhere).				Feedback that vignette flows well. Compromise between what is realistic versus what is in line with literature (e.g. suggesting client seeks alternative support elsewhere).
29			Discussion about case management supervision and process of stepping up or even stepping out. Linking process to holding research but also think about realistic decisions re. holding client and how case management used.				Discussion about case management supervision and process of stepping up or even stepping out. Linking process to holding research but also think about realistic decisions re. holding client and how case management used.
30			Discussion about difference between newly qualified PWP and more experienced PWP in how they engage clients. Newly qualified keener to influence change in client, more experienced PWP putting more on client (e.g. "what do you want from the process?")				Discussion about difference between newly qualified PWP and more experienced PWP in how they engage clients. Newly qualified keener to influence change in client, more experienced PWP putting more on client (e.g. "what do you want from the process?")
31			How does stage of career affect decision-making of PWP? Influence of client sociodemographic (e.g. client's lower social status might deter PWP from pursuing treatment and keep trying with client).				How does stage of career affect decision-making of PWP?

Appendix I: Study A. In-depth thematic analysis

Client suitability.

These responses fell into three sub-themes.

Risk status. PWP teaching staff (Participant 1 and 2) commented that the risk status of Jack was realistic for IAPT:

"What was the risk again? So fleeting thoughts?" (P1)

"Yeah, but not made any plans and denies any true intent." (Facilitator) "Yeah, we get that quite a lot." (P1)

They added that Jack's risk status sounded suitable for a standard patient who would be seen under IAPT:

> "We will get patients that are more risky than we should be seeing in primary care and obviously then get – try to get them the appropriate help but that's (referring to Jack's profile) probably a sort of standard profile, you know, so it's someone who's depressed and normally has passive thoughts of escape, something like that but not with any sort of active, you know." (P2)

Referral and screening process. They explained the usual process when a client is referred to IAPT by their GP and the screening assessment process for treatment suitability in their region:

> "Yeah, so we wouldn't triage from a GP, we'd just assess them straight off [---] It depends what the definition is, so in XXXX, for example, we have – they come into IAPT, they have an assessment, which would be like a suitability assessment, other services might do a triage, where a screening might mean a few different things. I think probably in this situation, it's at the point where they're seeing us, and we're sort of getting information and deciding to – whether they're suitable for treatment or not. Because like a triage would be – would almost be a step before that". (P2)

The teaching staff agreed that the vignette accurately depicted a common situation in IAPT at the point of client screening:

"Very true to life." (P1)

"Ok and seems to be expecting to see a counsellor." (Facilitator)

"He's hit the nail on the head." (P1)

Motivation: One staff member talked about how important it was that the client 'bought into the model' and stated that the 'Jack' vignette gave a sense of this:

"Yeah, because I think there's something around – you wouldn't initiate treatment for someone who's not motivated or who's not willing to buy into the model, but the fact that he sort of then – because he's sort of done the 5 areas and the problem statement and the goal, he's done - the narrative that he's responded well to the 5 areas, I guess the – the options are good…" (P2)

'The other staff member suggested that if participants felt 'Jack' was not buying into the model then a third option could be support from the voluntary sector:

"I suppose the – the option might be if he doesn't want what your offering and thought he was going to see a counsellor and wanted to see a counsellor you could maybe discuss voluntary sector counselling if that was something…" (P1)

Accurately portraying a collaborative approach. These responses fell into two sub-themes.

Setting goals with the client. Both staff members commented that the chosen intervention (cognitive restructuring) did not appear in line with Jack's goals and suggested an alternative option:

"See, cognitive restructuring doesn't seem to be congruent with that goal." (P2)

"You'd probably go on to the behavioural activation stuff." (P1)

"Ok, so maybe do it the other way round then?" (Facilitator)

Staff indicated the importance of collaborative goal setting at the point of assessing patient suitability:

> "I suppose it's – going back to his goals, so he sets the goal, even though his expectation of the appointment is different to the reality, so he's come in thinking here's an opportunity to offload, but he's sort of – you set a collaborative goal together. So, is that then assuming that he's then on board with the process or...?" (P2)
>
> "Has he gone along with it (the process of goal setting)?" (P1)

Client preconceptions. One staff member commented that in IAPT there was often some expectation that the client might not have been given correct information. This might be why the client is dubious and therefore this would be seen as understandable by the PWP:

> "Well, I think it depends on the PWP […] I think people – half the time they are told they're coming to see a counsellor […] So it is quite natural that you will have someone who is a little bit dubious but will give it a go." (P1)

Accurately depicting process of selecting treatment. Responses fell into three sub-themes.

Therapist decision-making. One staff member suggested why therapists often choose behavioural activation (BA) to begin with:

> " And quite often it's easier to see a change with BA, so you might start there with depression." (P1)

The current researcher discussed with staff whether it was realistic that in the vignette participants have the option to change their intervention. One staff member advocated for this and spoke about how common it was for therapists to change their minds:

"Yeah, because I think you should carry on with the BA, but also people do – you do swap around when the pressure gets to you. Someone may be – maybe you start talking about some negative thoughts or not sleeping and you jump on that, so. It doesn't feel kind of clear cut." (P1)

The other staff member added that sociodemographic factors related to a client can also have an impact on the decisions PWPs make:

> "And demographics has an effect, because if someone's in a surgery where someone – say you were in 2 surgeries in XXXX for example, one where people are very – maybe have a history of high education in the area and people are more willing to engage more, maybe you hold out more hope for those patients than in an area that is more socially deprived… the demographics definitely affect how much you sort of continue with an intervention or whether you give up." (P2)

Barriers: 'Staff spoke of some of the barriers that can come up regardless of what treatment a PWP selects for a client:

> "Or sometimes people will come back and say they didn't like it, but what they actually mean is they didn't understand it, or they need an explanation, or we've used too much jargon, or they might not be able to read…" (P2)

Teaching staff suggested that the vignette could include reference to the Com-B approach as a way of overcoming potential barriers to the work:

> […] we talk about using something called Com-B which is looking at someone's ability to – that's Com-B, and that's looking at someone's ability to understand, engage, I mean, it's quite in depth, but as a - obviously […] because that's looking at barriers and understanding and motivation and stuff like that."

Decision tree scoring. 'Staff offered advice on how to structure treatment related content so as to make it congruent with both behavioural activation and cognitive restructuring. That way it would still be possible to score normative/counter-normative choices but participants would not suspect whether they had made the 'right' choice or not participants would not suspect whether they had made the 'right' choice or not:

> "So maybe you decide to focus on the next step of the intervention." (P1)
>
> "Or like he's – he brings his diary back, because that diary could be behavioural or cognitive." (P2) (coded twice)

> "Maybe 'you go through an example in session'; that's enough. […] I suppose, because you wouldn't want to say – I suppose if it was BA, you might look at changing the hierarchy round or picking something out, or if it was cognitive restructuring, you'd challenge a thought in session. You want to mention doing something in session but be vague enough you don't, as 'I' said, allude to the intervention." (P1)

Ecological Validity. These responses fell into three sub-themes.

Realism of vignette. 'One staff member raised continuity issues in the vignette that meant the text would no longer apply to participants if they did not choose behavioural activation as a treatment option:

> "Yeah. You know if people select the previous one and go on to this, do they then realise they might have chosen the wrong option or?" (P1)
>
> [...]
>
> "So maybe you decide to focus on the next step of the intervention." (P2)
>
> "Or like he's – he brings his diary back, because that diary could be behavioural or cognitive […] Almost like – because I guess what you're looking at is, or what you're trying to sort of highlight is his lack of work between sessions, his lack or – well, he's basically not done it (his homework), has he, so that's the issue, isn't it, rather than the specific thing he's not done." (P1)

Staff members suggested ways of making the content richer and more realistic such as doing homework in the session and adding more information about what got in the way of the client not doing homework:

> "So, if he's not brought his BA diary, you might talk about how his week's been; if he's not brought his cognitive restructuring, you might do a bit." (P1)
>
> "So, the point is that you're almost trying to make up for the fact that he's not done it, by doing it with him, or doing it for him, basically". (P2)

[…]

"Um, maybe, hmm, you could look at what got in the way. Yeah, that's a bit of Com-B, yeah, look at what got in the way of completing it." (P1)

Staff members also commented upon how common the pull to offer more therapy is for therapists, even when client outcome measure scores show the client is not responding to therapy:

"Do you think there's something in there as well about the narrative about him telling you that he likes – not likes you, but is almost very complimentary of – because there's a pull when – there's a pull from patients when you're going to finish treatment, or when treatment's not working, to stick with them because they like – they're almost – they almost like flatter you a little bit or say they like you […] Because I've had people who've I've said, I've come to the end of treatment and they're saying 'oh, I'm really sad about that, um, I'm – and you know the scores are staying the same, you know you're not helping them, but they'll say that you are, they'll say that they really appreciate our chats, or…" (P2)

"Or, oh, one more." (P1)

"Yeah, and that's sometimes – that's a real pull." (P2)

Evidence base. Teaching staff discussed ideas for the 'hold' option of the vignette that were in line with the research literature but also realistic regarding what a PWP might do when it came to holding a client:

" […] I need to stay in line with what the research literature is saying and think about maybe could you factor in another possibility, meaning step out, as you say. " (Facilitator)

[…]

"I don't know if someone would do a relapse prevention for someone that's not basically shown reliable or sort of – and is showing that improvement […]" (P2)

PWP Characteristics. ‘One staff member suggested that it might be interesting to examine whether the stage of a person’s career might impact their decision-making process:

> “It would be interesting looking at the um, the decision making and how long people have been qualified for. So, whether people who have been qualified longer, right at the beginning, like ‘if you don’t want it, you can go for counselling with mind’, and newer people ‘let’s just try’.” (P1)

The other staff member gave an example of this from their own career practicing as a PWP:

> “I think when I started off, I know for a fact, like when I qualified 2012, I know there was people I saw in my training year that I saw for longer than I should have because I was trying to get some movement, and in hindsight, actually, I didn’t get any, and I wanted to sort of to keep trying, really.” (P2)

Appendix J. Thematic Analysis Process

Data emerging from the focus group were analysed using thematic analysis. The six phases described by Braun and Clarke (2006) were followed and results are listed below. These comprised:

1. Familiarization with the data set;
2. Initial codes generated;
3. Themes searched for;
4. Themes reviewed;
5. Themes defined and named;
6. Report produced.

This qualitative method intends to identify, analyse, organize, interpret and report patterns (i.e., themes) in the data (Clarke & Braun, 2017). Thematic analysis was chosen as it requires systematic, in-depth and intricate interpretations of the data (Clarke & Braun, 2017). An essentialist/naïve realist approach to inquiry was employed. This assumes there is a reality in the data and the researcher takes an active role in identifying and reporting these experiences and their meanings (Braun & Clarke, 2006). This was chosen given that PWPs are expected to follow specific assessment and treatment procedures. Furthermore, an 'objective', inductive and data-driven perspective of the participants' experience of the dynamic measure was required. Whilst the limitations of this position were acknowledged (Madill, Jordan, & Shirley, 2000) it was hoped it would improve the ecological validity of the dynamic measure.

Phase 1: familiarizing self with the data

Preliminary codes:

1. Suitability
2. Motivation
3. Treatment selection
4. Ecological validity

Phase 2: Generation of initial codes

Stage 1: Coded for

1. Risk status of Jack realistic for IAPT.
2. Risk status sounds suitable for standard patient.
3. Screening assessment process for treatment suitability.
4. 'Jack's perspective re. wait time and expectancy to see a counselor 'bang on'.
5. Third option – private counseling might be offered.
6. Setting collaborative goal would only be done if Jack onboard with treatment (motivated) and presenting with mid to moderate depression.
7. More dialogue with client in an actual assessment as to their motivation right from beginning.
8. Aim is for Jack to sound dubious but will have a go.
9. Some expectation that client might not have been given all the correct information. This might be why dubious and therefore would be seen as understandable by PWP.
10. First intervention should be in line with goals and so would start with BA.
11. You do swap around; pressure gets to you.
12. Therapy does not always go to plan – clear cut.
13. Client may not initially understand CBT concepts.
14. Weave in more about PWP thinking about barriers as per COM-B approach with client.

15. More background about what might be making things worse for Jack. E.g. fact that PWP has given him 40 pages to read of info when motivation low!
16. Continuity issues raised that do not allude to either intervention so that right or wrong option selection not highlighted.
17. Discussions about making content more realistic and richer such as doing homework in session and add more info in about what got in the way of not doing homework.
18. Also look at applying COM-B approach to look at barriers.
19. Try not to be too specific re. certain approaches (e.g. problem solving) as this begins to sound like going down a particular intervention route.
20. More detail about exactly what would be offered if stepped up (e.g. counseling or more CBT).
21. Factor in process you would go through of reviewing their progress and motivation in sessions with client before discussing stepping up in case management supervision.
22. Collaborative decision with client is the aim.
23. Pull to offer more discussed if client is saying how much they appreciate 'chats' with PWP despite scores not responding to treatment.
24. Also questions as to what approach might be best re. allocation at step 3? Might not be relevant, however.
25. Clarification as to what treatment content would look like to make it congruent with either behavioural activation approach or cognitive restructuring. Keep it vague enough but still following a thread of something.
26. Discussion about wording in the text so that it does not align with any one approach specifically.
27. Also realistic about what might lead PWPs to be suspicious of selecting right and wrong answers (e.g. get rid of relapse prevention option).

28. Feedback that vignette flows well. Compromise between what is realistic versus what is in line with literature (e.g. suggesting client seeks alternative support elsewhere).
29. Discussion about case management supervision and process of stepping up or even stepping out. Linking process to holding research but also think about realistic decisions re. holding client and how case management used.
30. Discussion about difference between newly qualified PWP and more experienced PWP in how they engage clients. Newly qualified keener to influence change in client, more experienced PWP putting more on client (e.g. "what do you want from the process?")
31. How does stage of career affect decision-making of PWP? Influence of client sociodemographic (e.g. client's lower social status might deter PWP from pursuing treatment and keep trying with client).

Stage 2: Potential emerging themes/repeated patterns

1. Ambivalence
2. Barriers
3. Client pre- conceptions

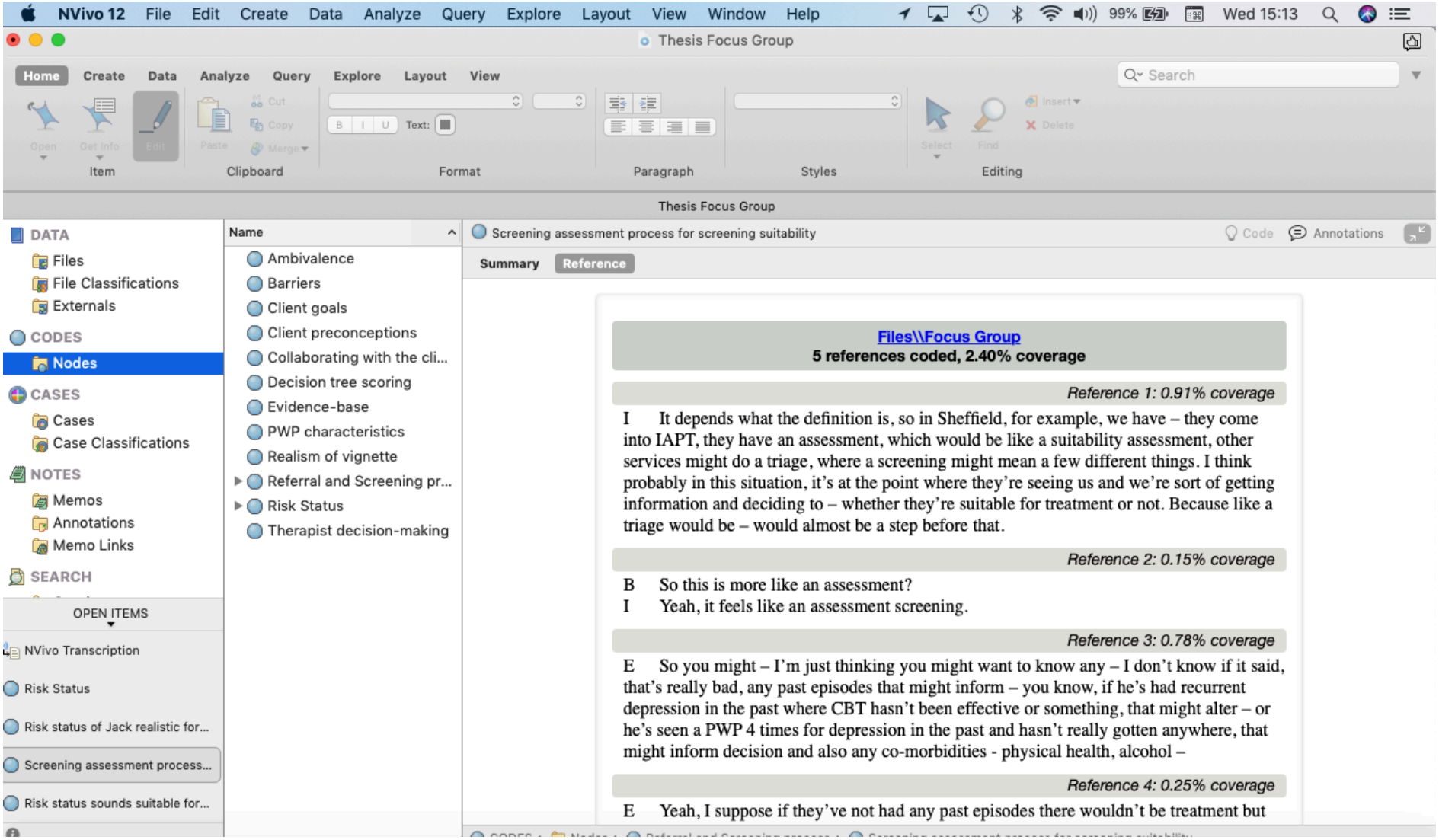


Figure 1. NVivo 12 was used to manage data. Screenshot shows it at Stage 2 of thematic analysis process.

4. Collaborating with the client
5. Decision tree scoring
6. Evidence-base
7. Motivation
8. PWP Characteristics
9. Realism of vignette
10. Referral and screening process
11. Risk Status
12. Setting goals with the client
13. Therapist decision-making

Phase 3: Searching for themes

Category 1: *Client Suitability*

Subcategory 1: *Risk status*

1. Risk status of Jack realistic for IAPT.
2. Risk status sounds suitable for standard patient.

Subcategory 2: *Referral and screening process*

3. Referral
4. Screening assessment process for treatment suitability.
5. 'Jack's perspective re. wait time and expectancy to see a counselor 'bang on'.

Subcategory 3: *Motivation*

6. Third option – private counseling might be offered.
7. Buying into the model.

Category 2: ***Accurately portraying a collaborative approach***

Subcategory 4: *Setting goals with the client*

8. Setting collaborative goal would only be done if Jack onboard with treatment (motivated) and presenting with mild to moderate depression.
9. First intervention should be in line with goals and so would start with BA.

Subcategory 5: *Client Preconceptions*

10. Some expectation that client might not have been given all the correct information. This might be why dubious and therefore would be seen as understandable by PWP.

Category 3: ***Accurately depicting process of selecting treatment***

Subcategory 6: *Therapist decision-making*

11. Easier to see change if you start with BA.
12. You do swap around; pressure gets to you. Therapy does not always go to plan – clear cut.
13. Influence of client sociodemographic (e.g. client's lower social status might deter PWP from pursuing treatment and keep trying with client).

Subcategory 7: *Barriers*

14. Client may not initially understand CBT concepts.
15. Weave in more about PWP thinking about barriers as per COM-B approach with client.
16. More background about what might be making things worse for Jack. E.g. fact that PWP has given him 40 pages to read of info when motivation low!

Subcategory 8: *Decision tree scoring*

17. Clarification as to what treatment content would look like to make it congruent with either behavioural activation approach or cognitive restructuring. Keep it vague enough but still following a thread of something.

18. Discussion about wording in the text so that it does not align with any one approach specifically.
19. Also realistic about what might lead PWPs to be suspicious of selecting right and wrong answers (e.g. get rid of relapse prevention option).

Category 4: ***Ecological Validity***

Subcategory 9: *Realism of vignette*

20. Frequency of therapy.
21. Continuity issues raised that do not allude to either intervention so that right or wrong option selection not highlighted.
22. Discussions about making content more realistic and richer such as doing homework in session and add more info in about what got in the way of not doing homework.
23. Try not to be too specific re. certain approaches (e.g. problem solving) as this begins to sound like going down a particular intervention route.
24. More detail about exactly what would be offered if stepped up (e.g. counseling or more CBT).
25. Factor in process you would go through of reviewing their progress and motivation in sessions with client before discussing stepping up in case management supervision. Collaborative decision with client is the aim.
26. Pull to offer more discussed if client is saying how much they appreciate 'chats' with PWP despite scores not responding to treatment.
27. Also questions as to what approach might be best re. allocation at step 3? Might not be relevant, however.

Subcategory 10: *Evidence base*

28. Feedback that vignette flows well. Compromise between what is realistic versus what is in line with literature (e.g. suggesting client seeks alternative support elsewhere).

29. Discussion about case management supervision and process of stepping up or even stepping out. Linking process to holding research but also think about realistic decisions re. holding client and how case management used.

Subcategory 11: *PWP Characteristics*

30. Discussion about difference between newly qualified PWP and more experienced PWP in how they engage clients. Newly qualified keener to influence change in client, more experienced PWP putting more on client (e.g. “what do you want from the process?”)
31. How does stage of career affect decision-making of PWP?

Appendix K: The Cognitive Reflection Test (CRT; Frederick, 2005)

Appendix L: The Rational and Intuitive Decision Styles Scale (DSS; Hamilton, Shih and Mohammed, 2016)

Appendix M: The Mini-IPIP 20-item Short Form Scale (Donnellan, Oswald, Baird and Lucas, 2006)

Appendix N: Participant information and consent

What is the study about?
You are invited to participate in a study exploring the clinical judgement and decision-making of Psychological Wellbeing Practitioners (PWPs). This research aims to gain further understanding regarding the factors that might affect clinical decision-making.

Who is conducting the study?
Benjamin Michael (trainee clinical psychologist) is conducting this study with the support of two research supervisors based within the psychology department of the University of Sheffield. The study will form part of the requirements for the Doctor of Clinical Psychology degree of Benjamin Michael.

Who can I contact if I have questions about the study?
Benjamin Michael can assist you with any enquiries you may have regarding the use of the data or the survey itself. He also welcomes any of your comments about the completion of the survey. Please feel free to contact him (bmichael1@sheffield.ac.uk).

If you have a complaint about the study please contact Benjamin Michael initially. His research supervisors Steve Kellett (s.kellett@sheffield.ac.uk) and Jaime Delgadillo (j.delgadillo@sheffield.ac.uk) may also be contacted however. If you feel that your complaint has not been handled to your satisfaction then please contact the Head of the Psychology Department, Glenn Waller (g.waller@sheffield.ac.uk).

What does the study involve?
If you wish to participate, you will be asked to complete a **(time to be decided after pilot study)** minute online survey**.**

How will my privacy be protected?
The information gathered from this survey is confidential and anonymous. When you submit your completed survey your name and email address will not be stored.
Your results will be published in Benjamin Michael's doctoral theses as part of a larger data set and may also appear in peer-reviewed journals. However, no individual participant details will be identified in any publication of results. The data obtained will only be stored and accessed by Benjamin Michael, Steve Kellett and Jaime Delgadillo. The anonymous data you provide may be used in future research.

GDPR
As new data protection legislation came into effect across the EU, including the UK on 25 May 2018; this means that we need to provide you with some further information relating to how your personal information will be used and managed within this research project. This is in addition to the details provided above.

The University of Sheffield will act as the Data Controller for this study. This means that the University is responsible for looking after your information and using it properly. In order to collect and use your personal information as part of this research project, we must have a basis in law to do so. The basis that we are using is that the research is 'a task in the public interest'.

Is my participation voluntary?

Participation in this study is entirely voluntary. If you do decide to participate, you are free to withdraw at any time and you will not be asked to provide a reason for your withdrawal.

The University of Sheffield Research Ethics Committee has approved the ethical aspects of this study.

Will financial/in kind payments be offered to participants?
As an incentive to take part in the study there will be a £1 donation made to the mental health charity, Rethink Mental Illness per participant for the first 50 participants who complete the study.

I agree to participate in this research, knowing that:

- I understand that my responses will be confidential.
- My participation is voluntary, and I am free to withdraw at any time.
- I give permission for members of the research team to have access to my anonymised responses.
- I understand that my name will not be linked with the research materials and that I will not be identifiable in the report or any further reports that result from the research.
- I have read the information sheet and I am aware that I may contact Benjamin Michael if I have any questions.
- I agree for the data I submit to be used in future research.

☐ Yes

☐ No

NEXT

Appendix O: Normality plots

Histogram and Q-Q plots CV1/CV2 testing assumption of normality

overall dynamic measure score

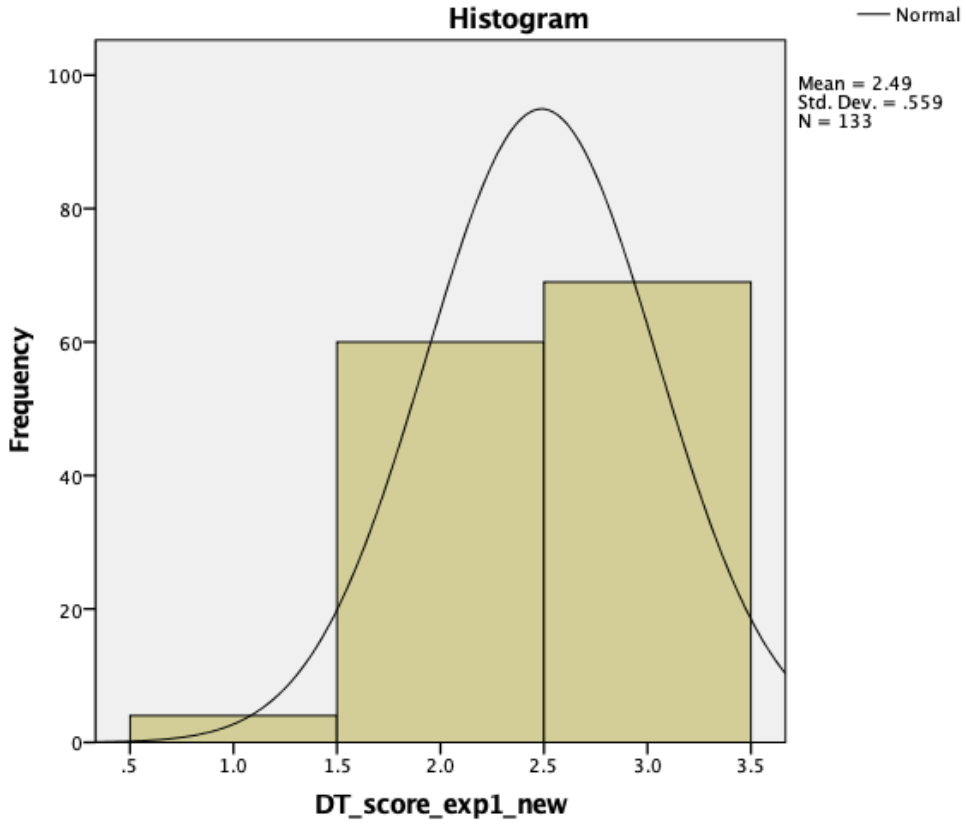


Figure 1. Histogram CV1 overall dynamic measure score

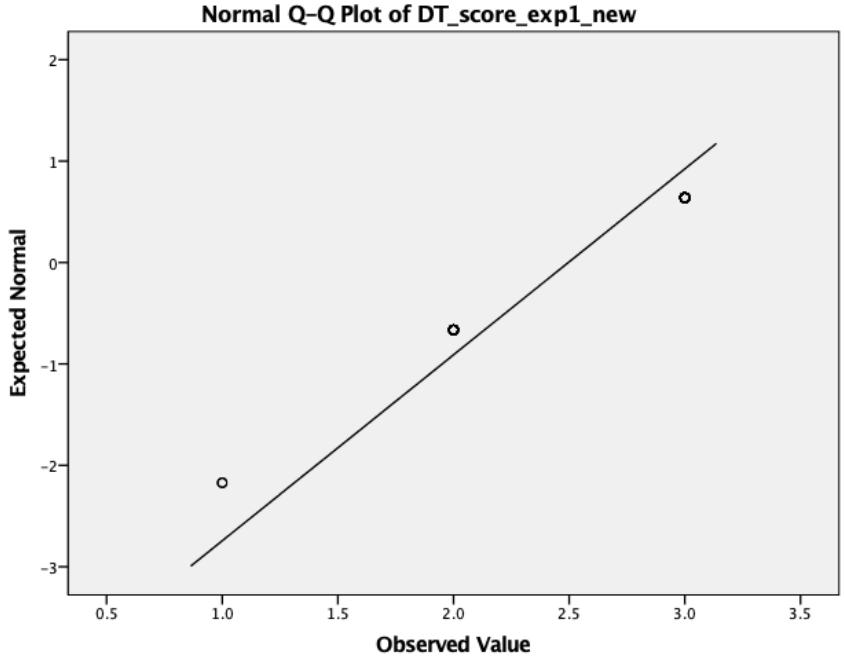


Figure 2. Normal Q-Q Plot of CV1 overall dynamic measure score

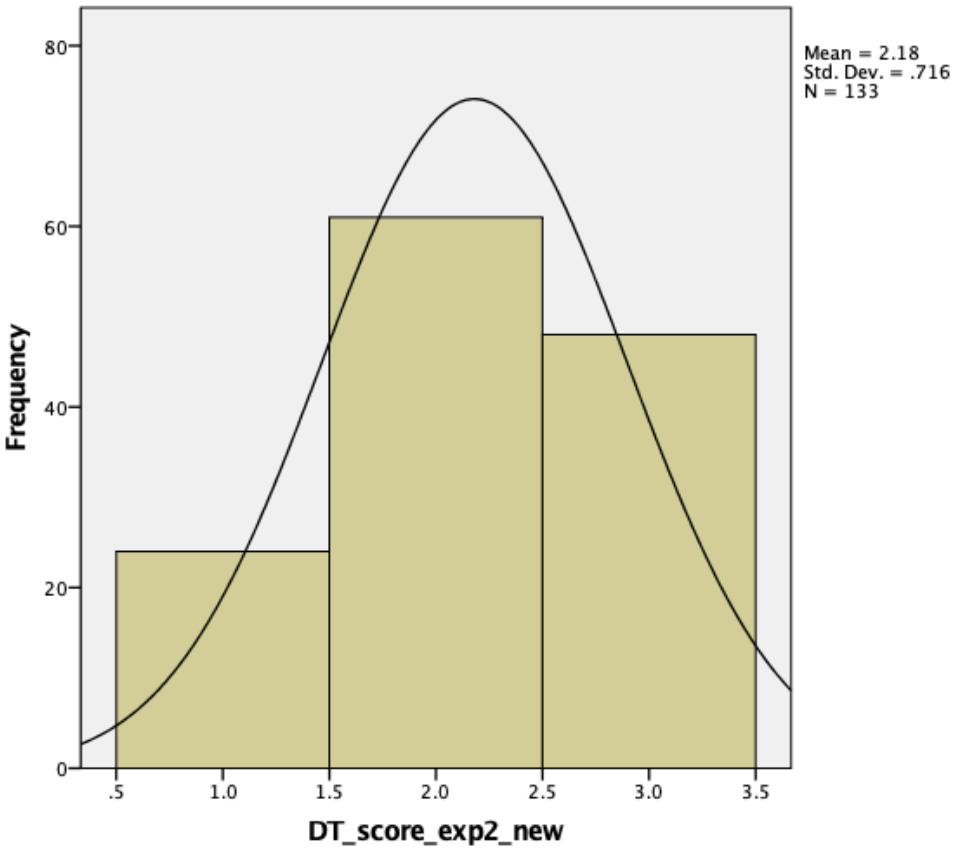


Figure 3. Histogram CV2 overall dynamic measure score

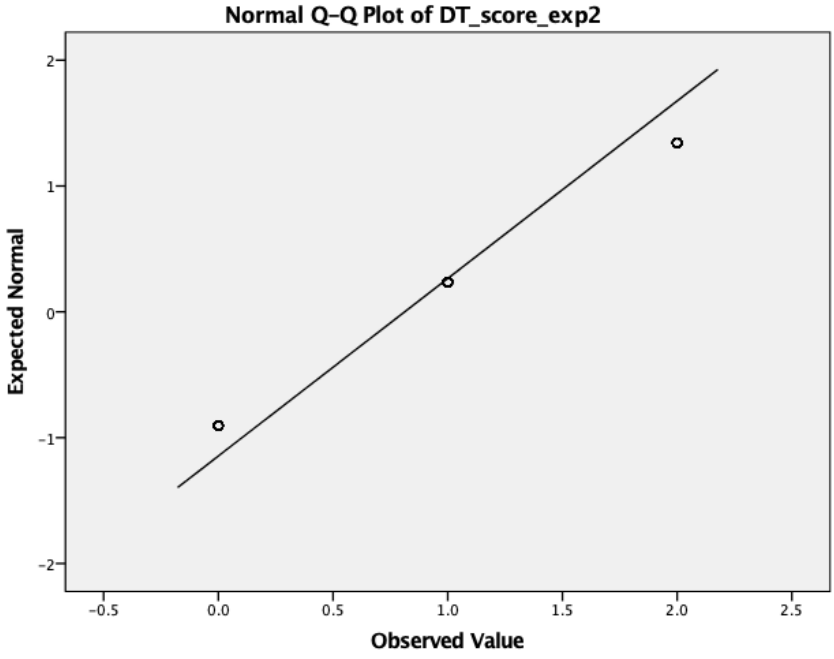


Figure 4. Q-Q Plot of CV2 overall dynamic measure score

Histogram and Q-Q plots testing assumption of normality static measures

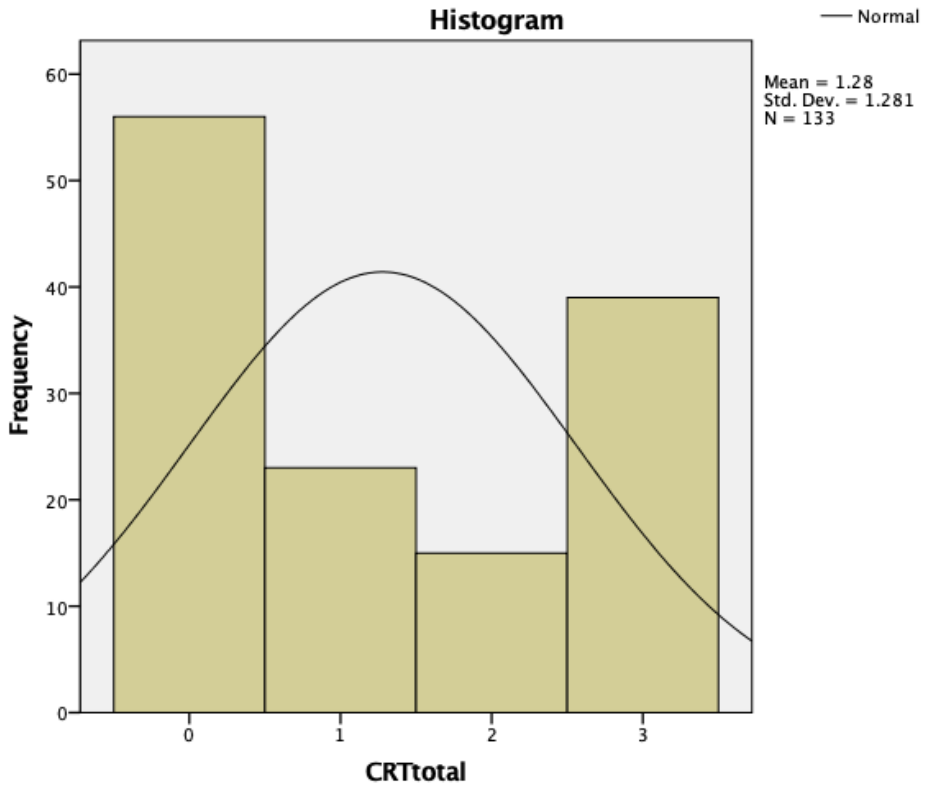


Figure 5. Histogram to check for normality: CRT score

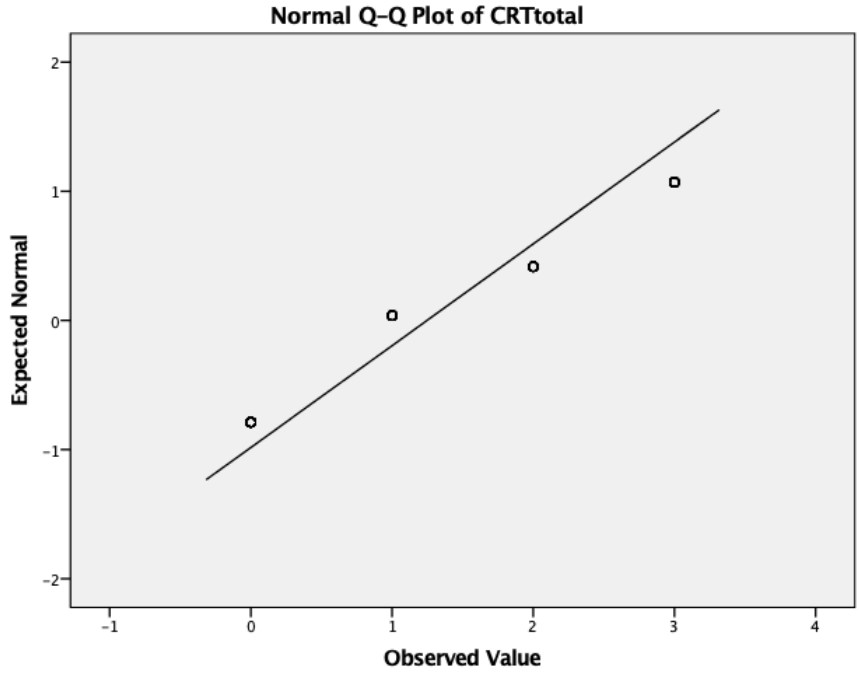


Figure 6. Normal Q-Q Plot to check for normality: CRT score

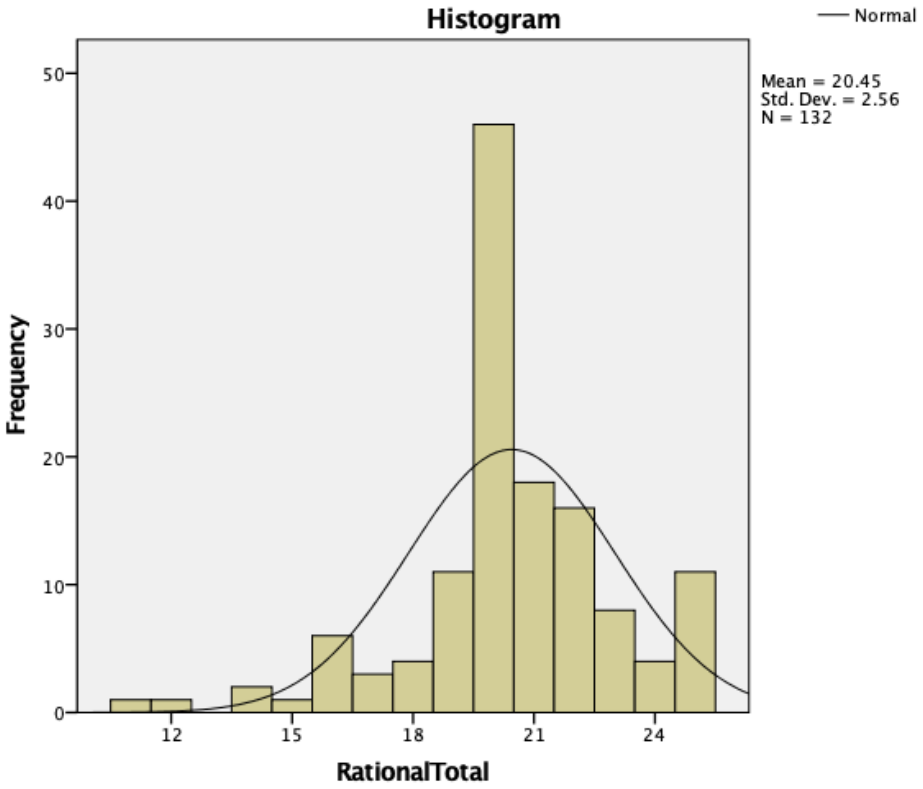


Figure 7. Histogram to check for normality: DSS Rational score

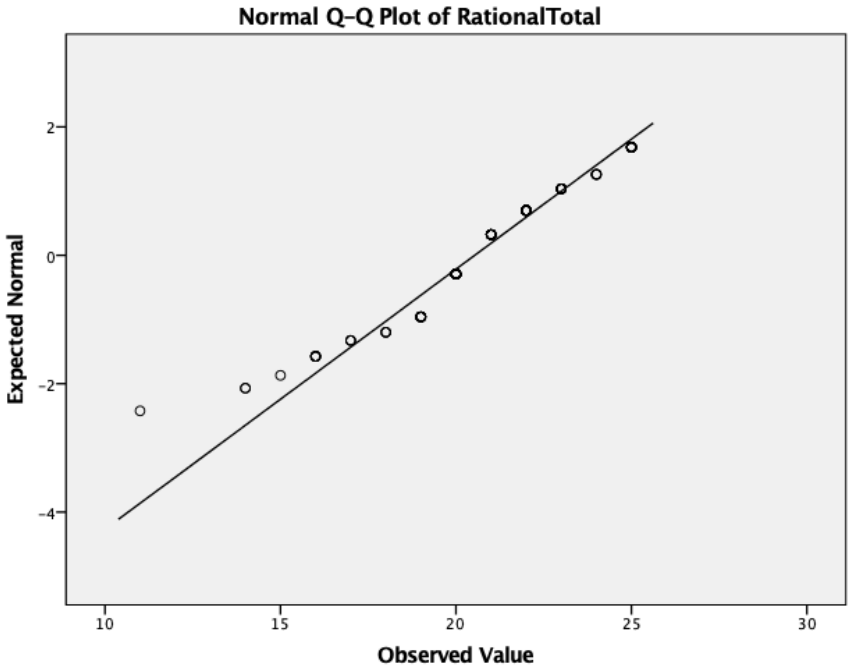


Figure 8. Normal Q-Q Plot to check for normality: DSS Rational score

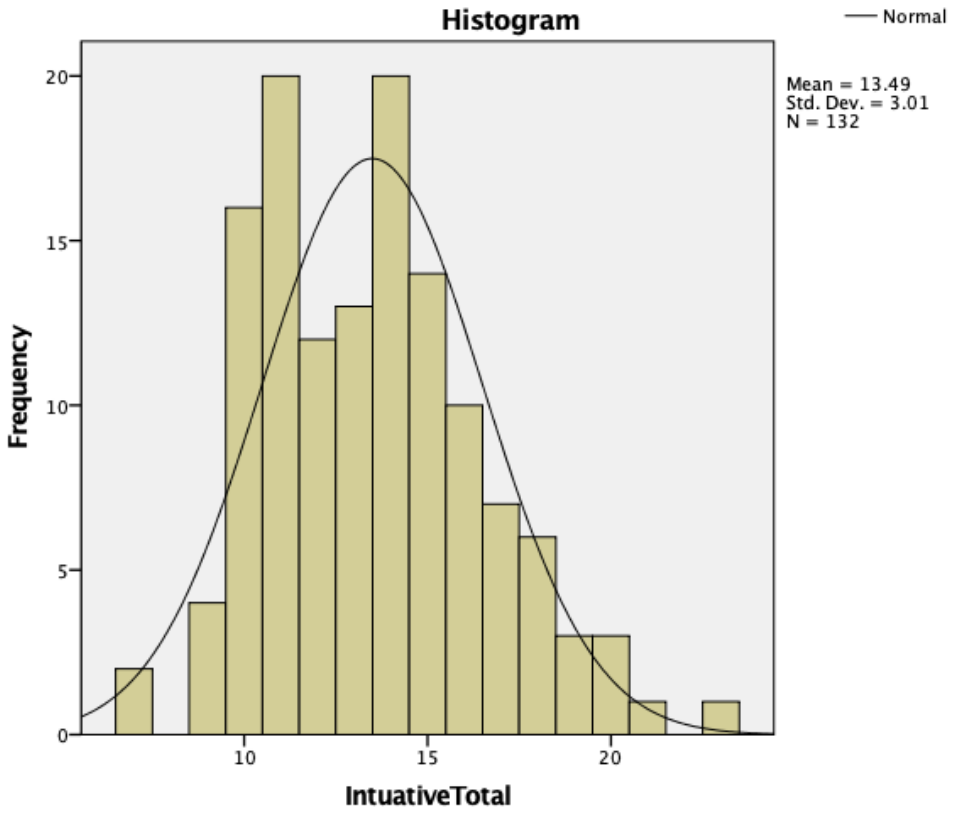


Figure 9. Histogram to check for normality: DSS Intuitive score

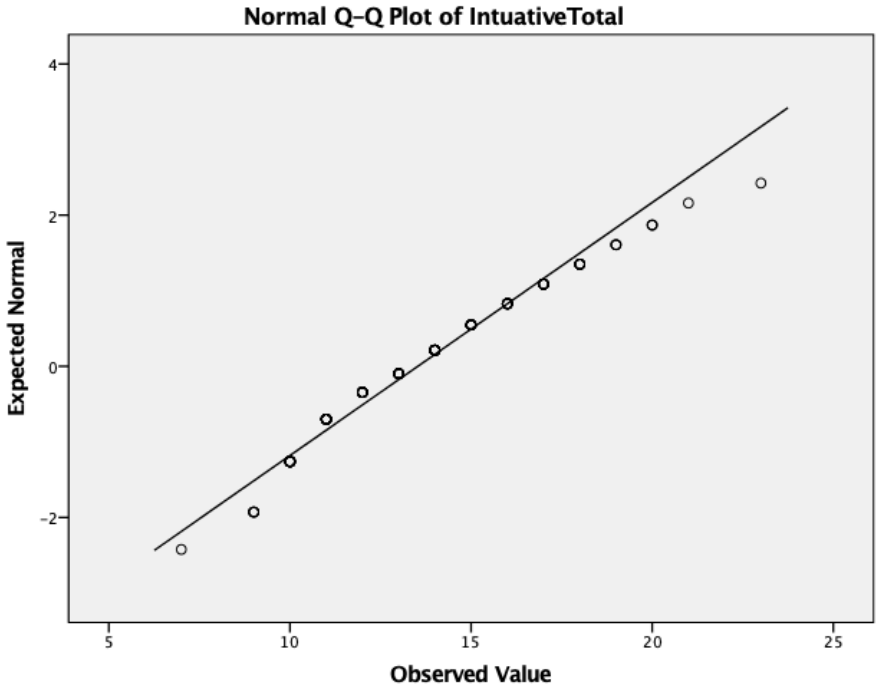


Figure 10. Normal Q-Q Plot to check for normality: DSS Intuitive score

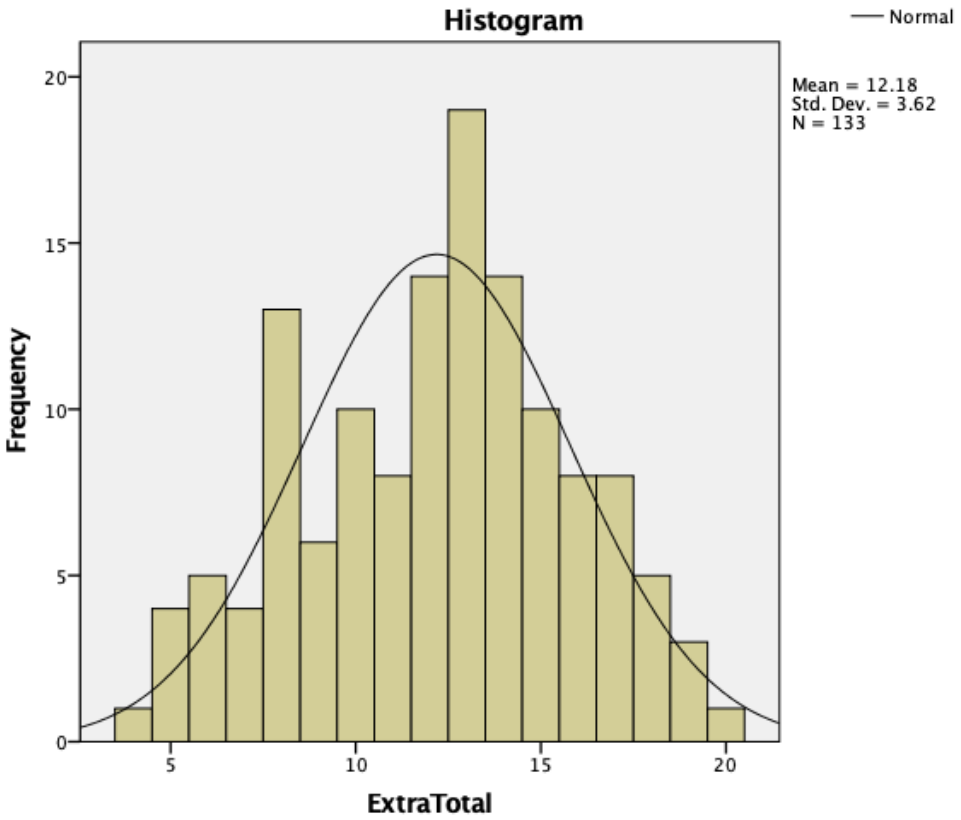


Figure 10. Histogram to check for normality: Extraversion score

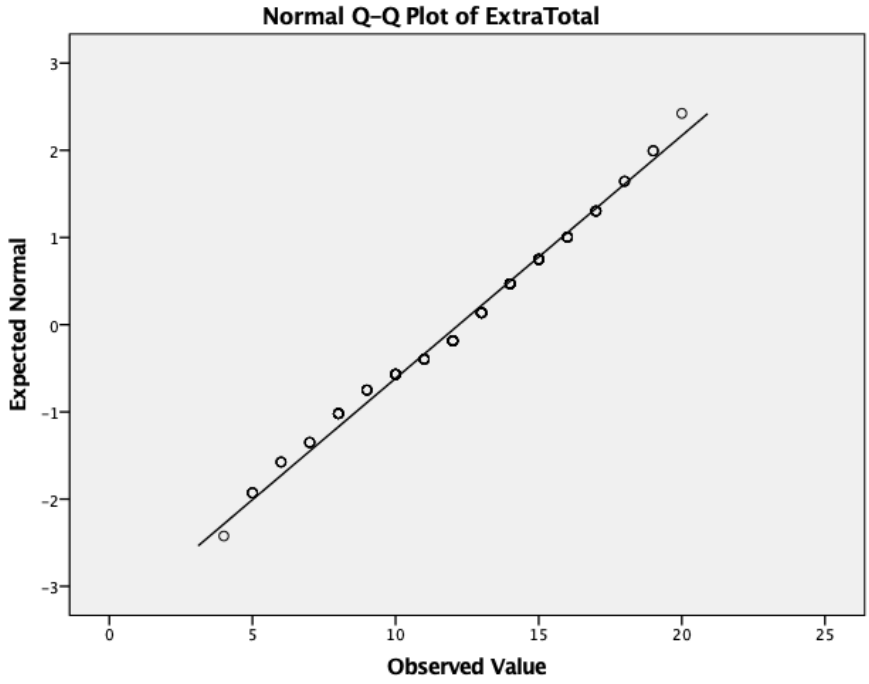


Figure 11. Normal Q-Q Plot to check for normality: Extraversion score

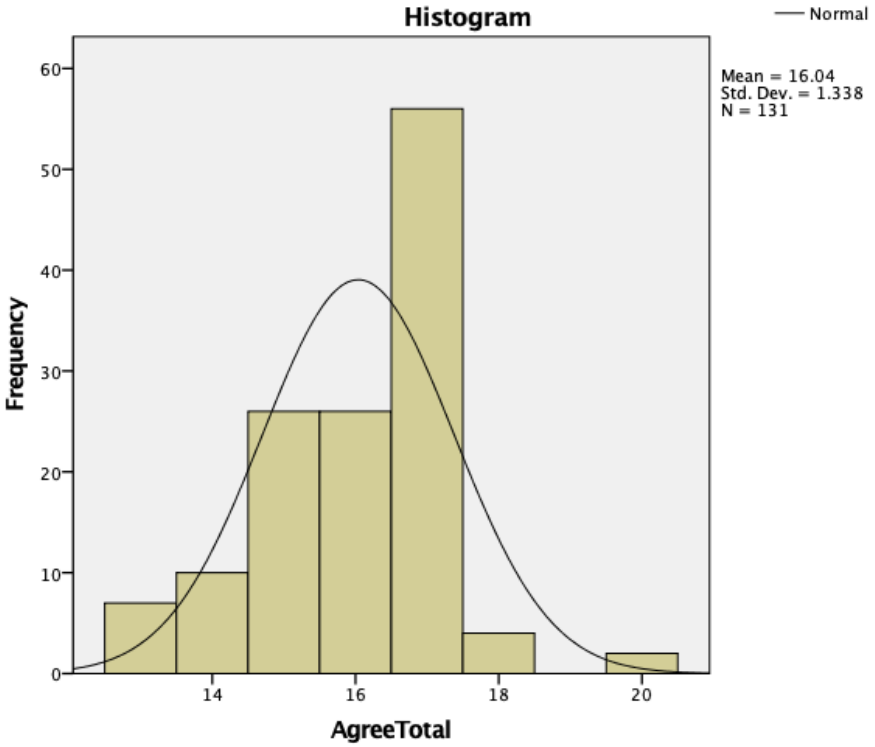


Figure 12. Histogram to check for normality: Agreeableness score

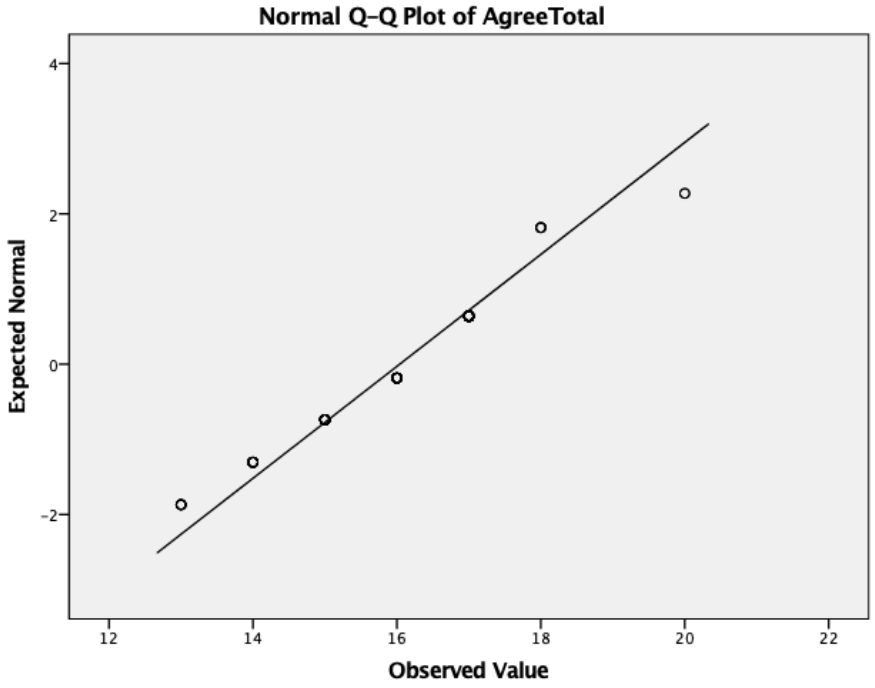


Figure 13. Normal Q-Q Plot to check for normality: Agreeableness score

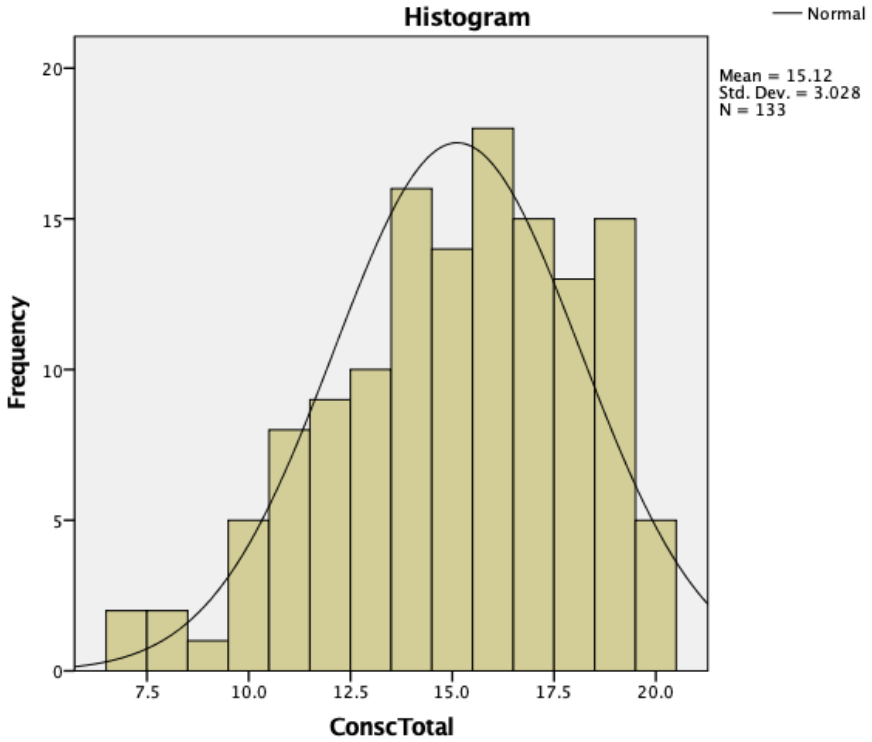


Figure 14. Histogram to check for normality: Conscientiousness score

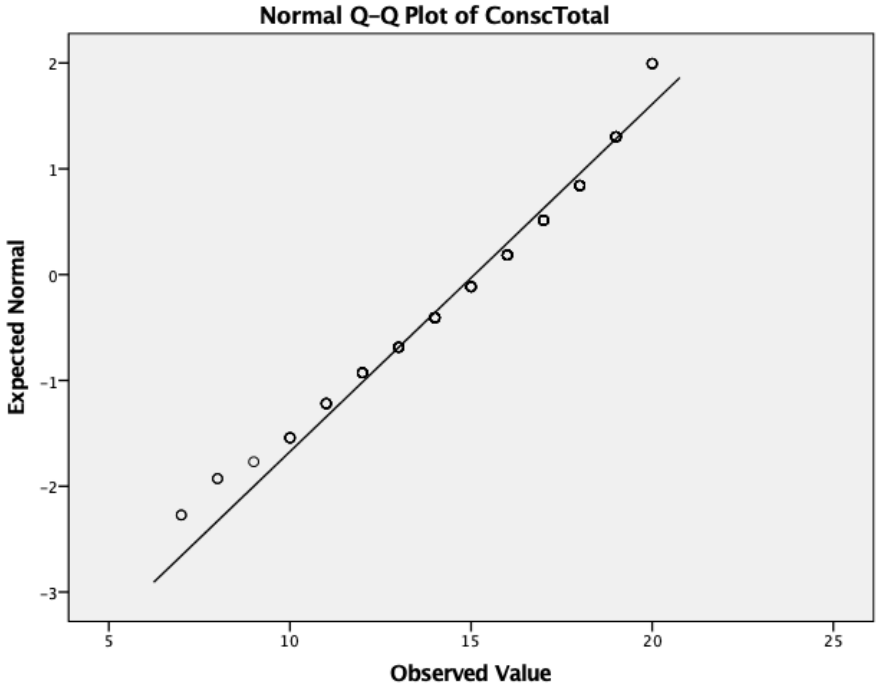


Figure 15. Normal Q-Q Plot to check for normality: Conscientiousness score

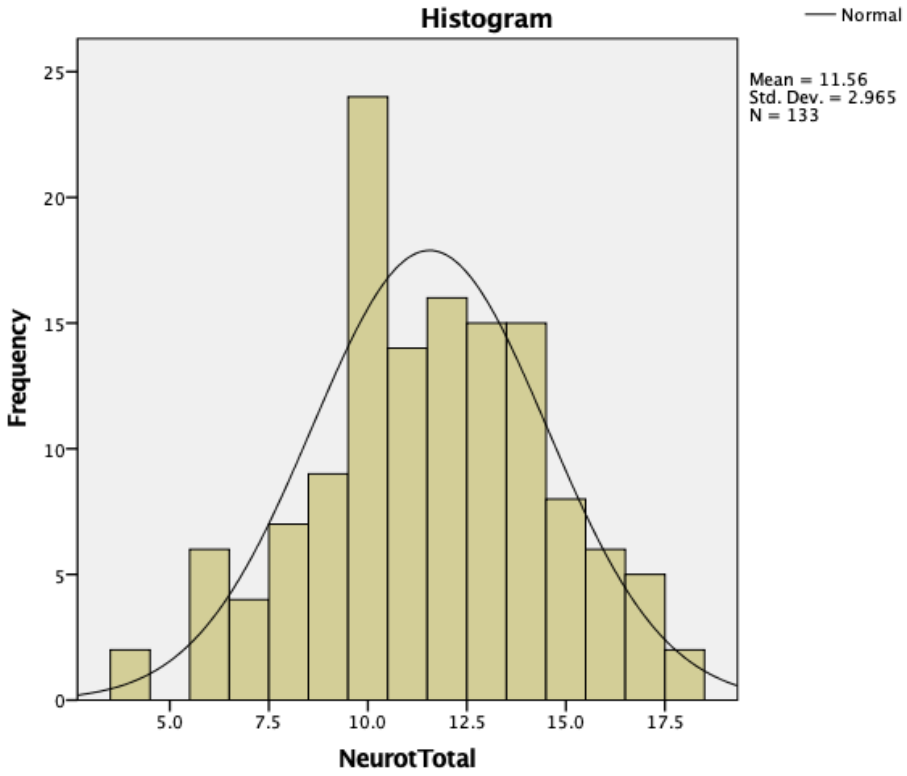


Figure 16. Histogram to check for normality: Neuroticism score

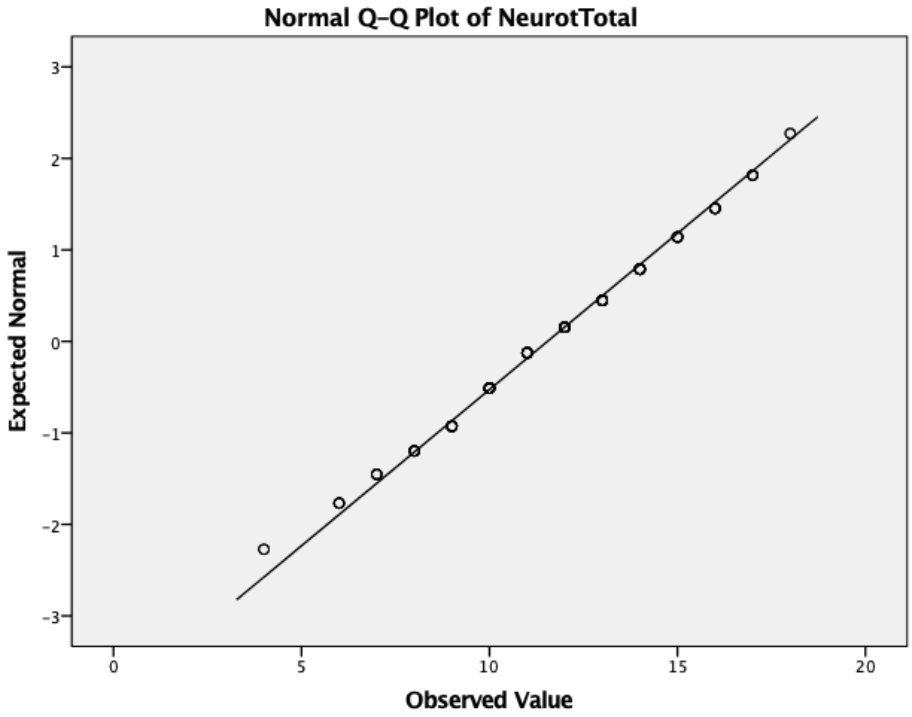


Figure 17. Normal Q-Q Plot to check for normality: Neuroticism score

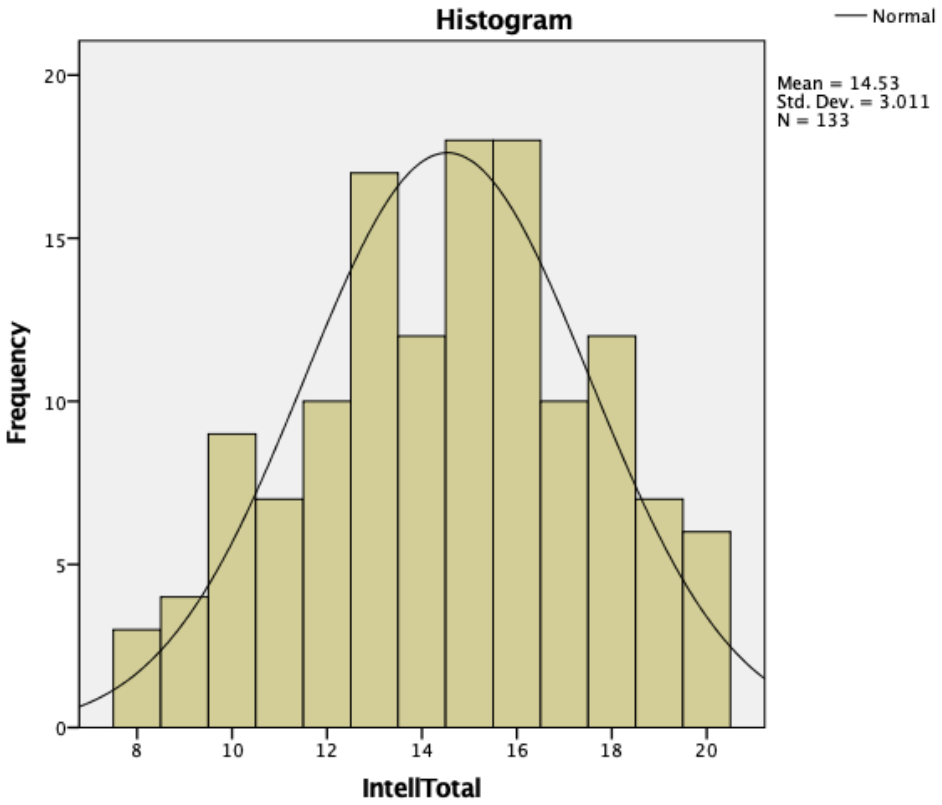


Figure 18. Histogram to check for normality: Intellect score

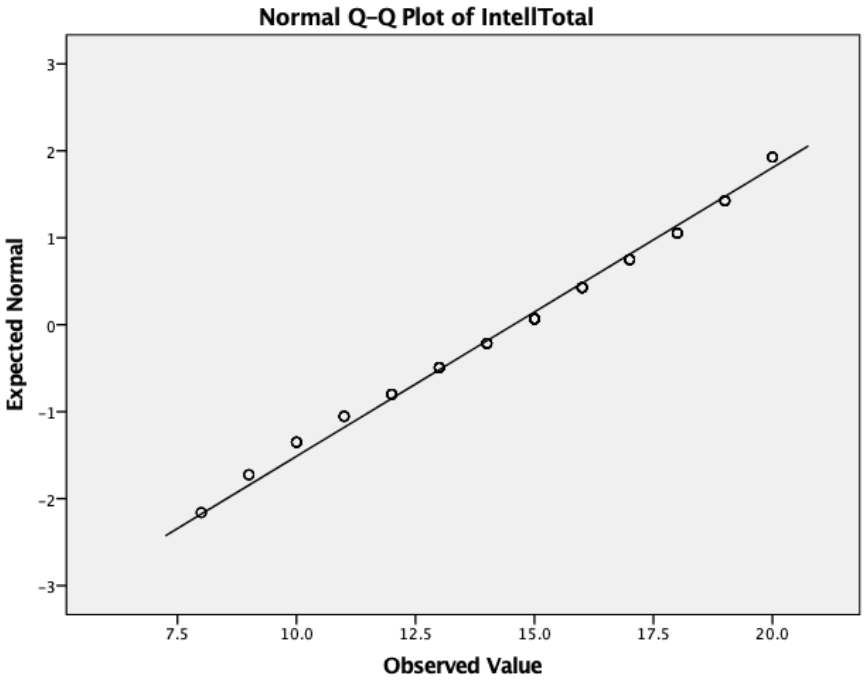


Figure 19. Normal Q-Q Plot to check for normality: Intellect score

Histograms with superimposed normal curve and P-P Plots testing the assumption of normality of the residuals for multiple regression CV1/CV2

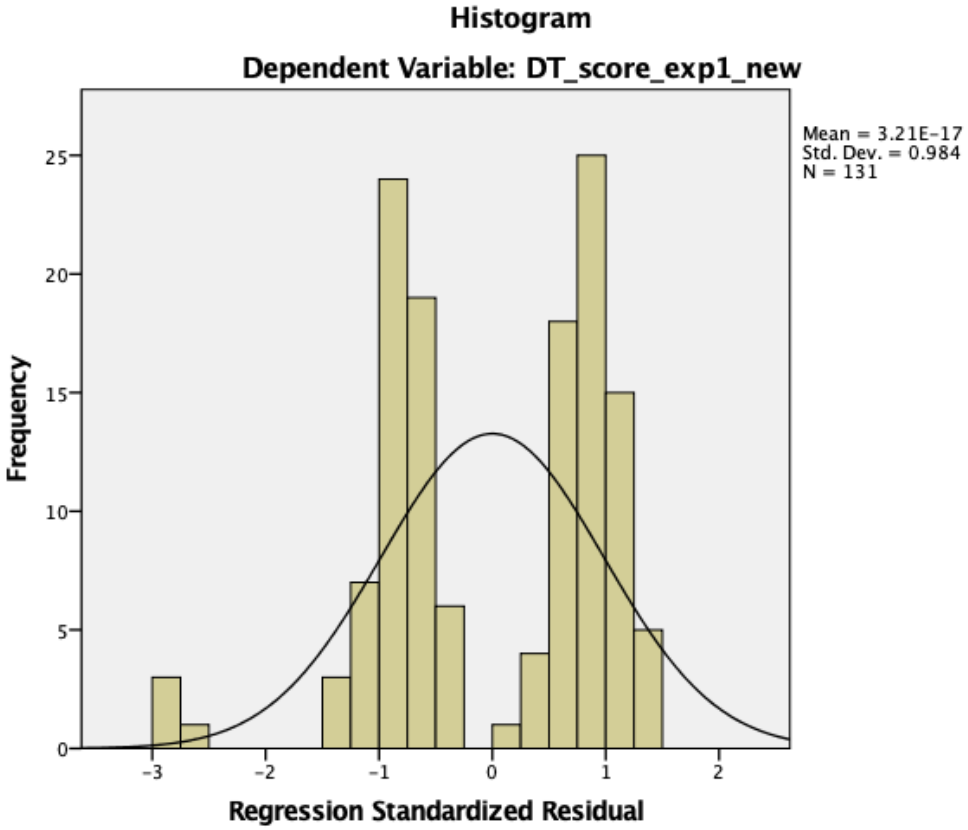


Figure 20. Histogram testing normality of the residuals for CV1

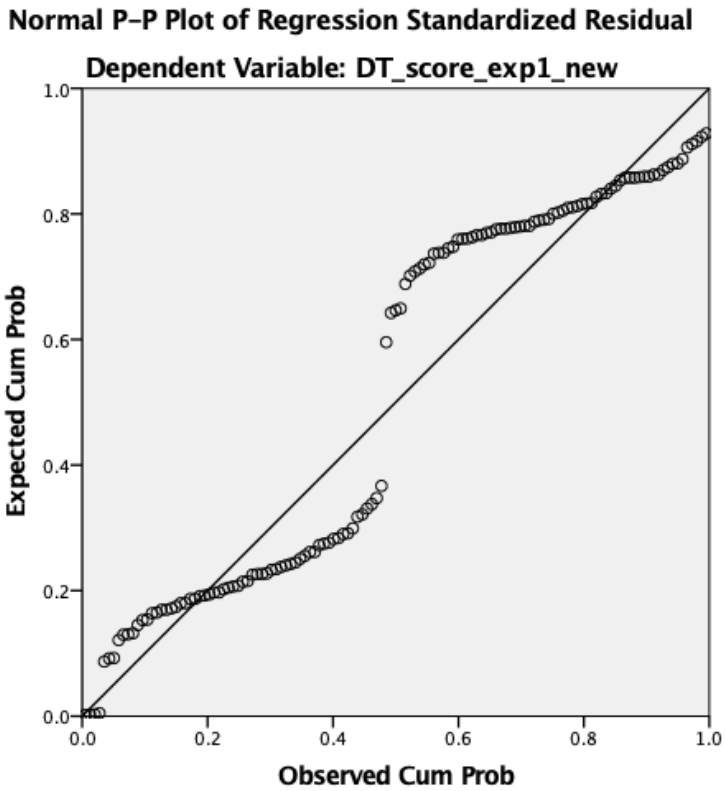


Figure 21. P-P Plot testing normality of the residuals for CV1

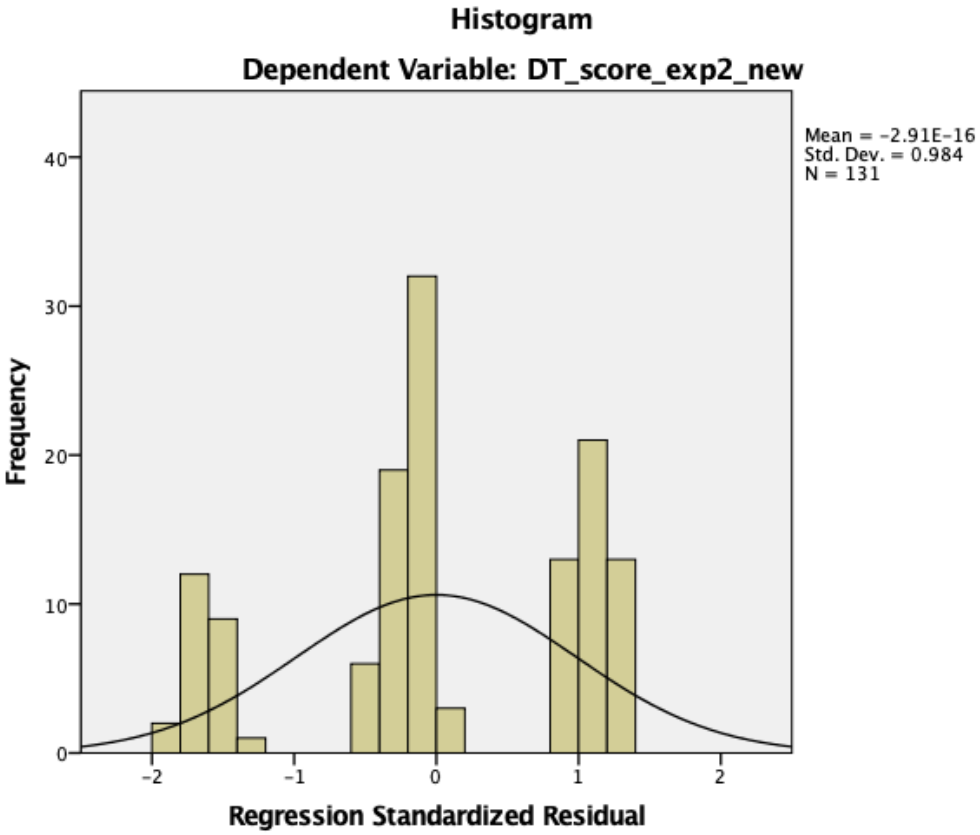


Figure 22. Histogram testing normality of the residuals for CV2

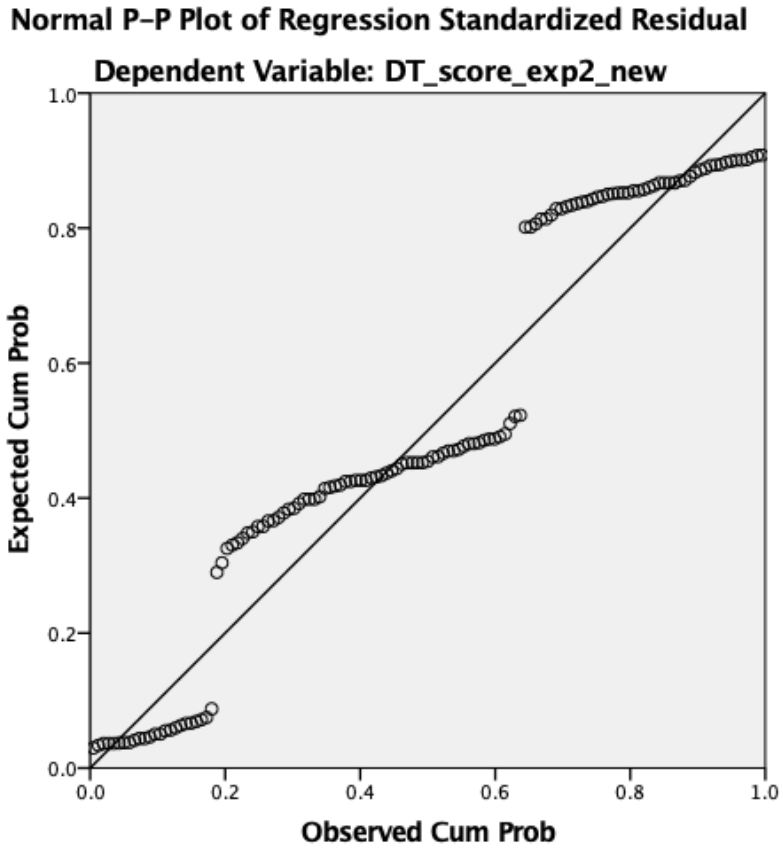


Figure 23. P-P Plot testing normality of the residuals for CV2

Appendix P: Linearity Plots

Examples of scatterplots testing assumption of a linear relationship between dynamic and static measure scores

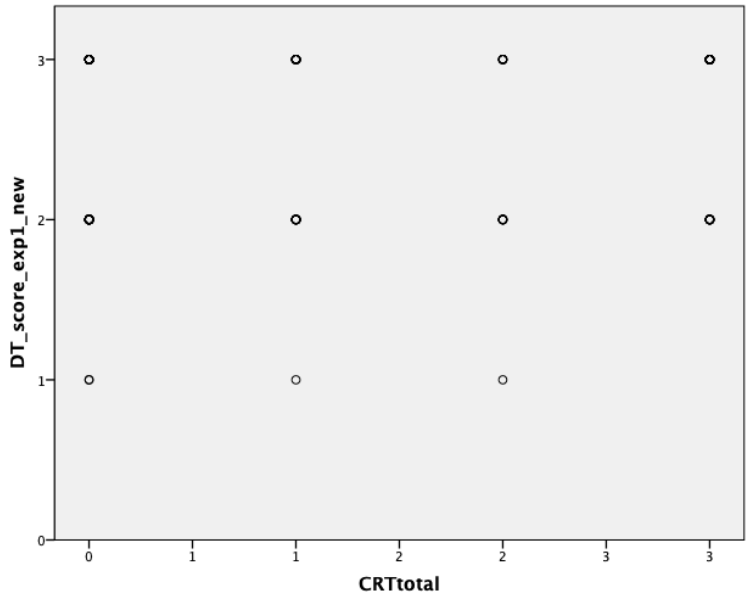


Figure 1. Scatterplot testing linear relationship between CV1 and CRT

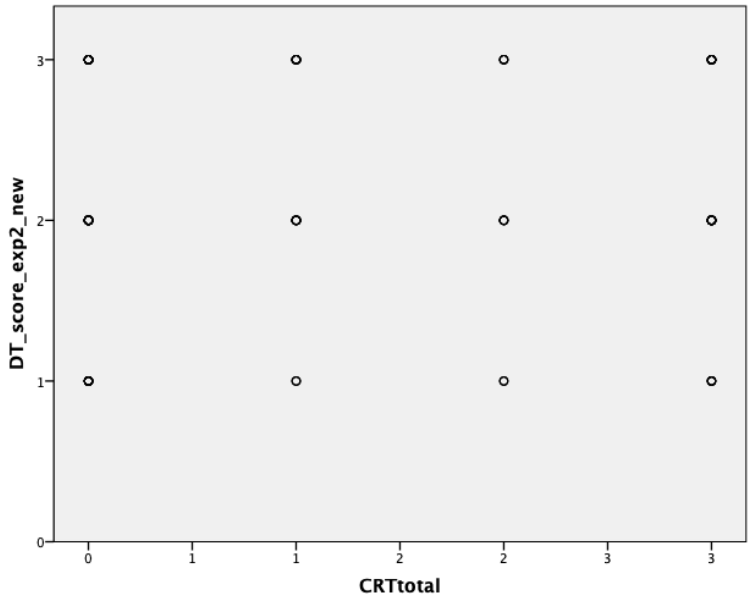


Figure 2. Scatterplot testing linear relationship between CV2 and CRT

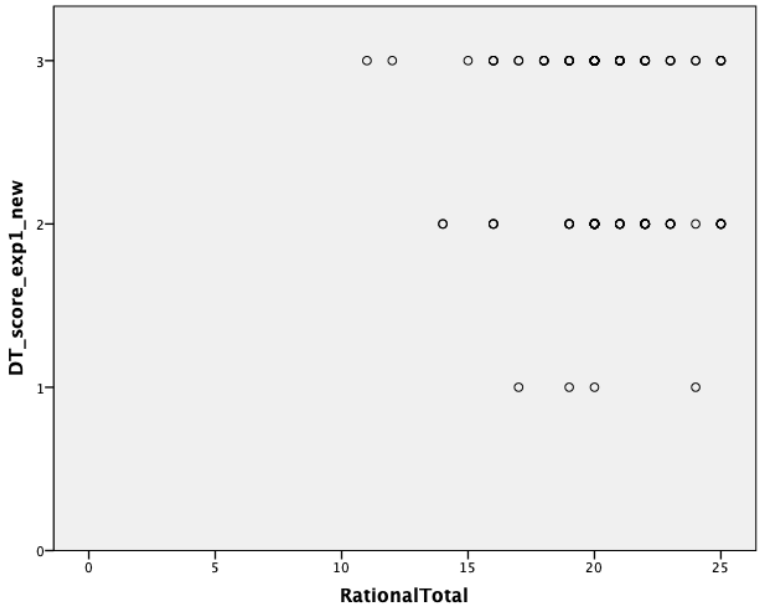


Figure 3. Scatterplot testing linear relationship between CV1 and DSS Rational

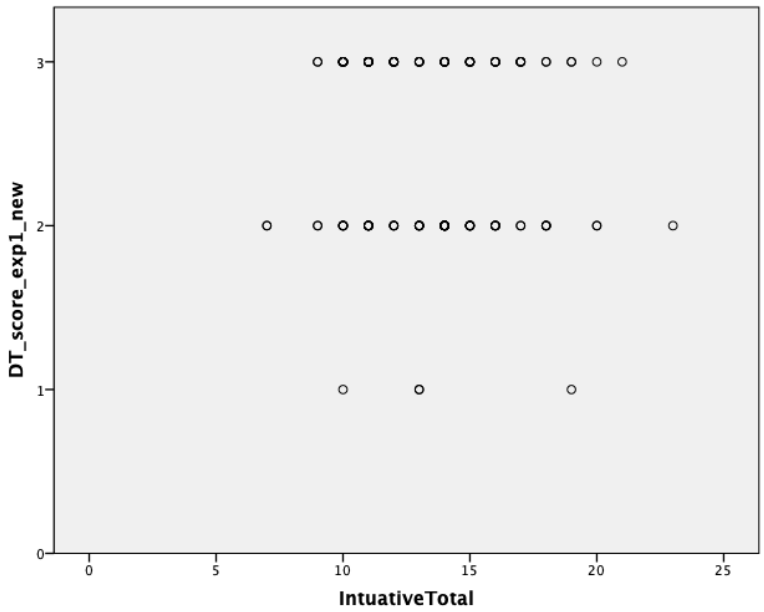


Figure 4. Scatterplot testing linear relationship between CV1 and DSS Intuitive

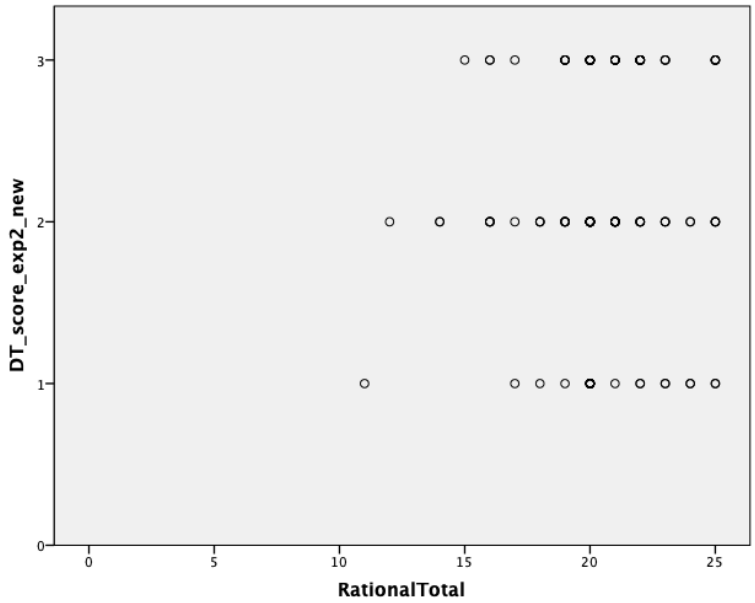


Figure 5. Scatterplot testing linear relationship between CV2 and DSS Rational

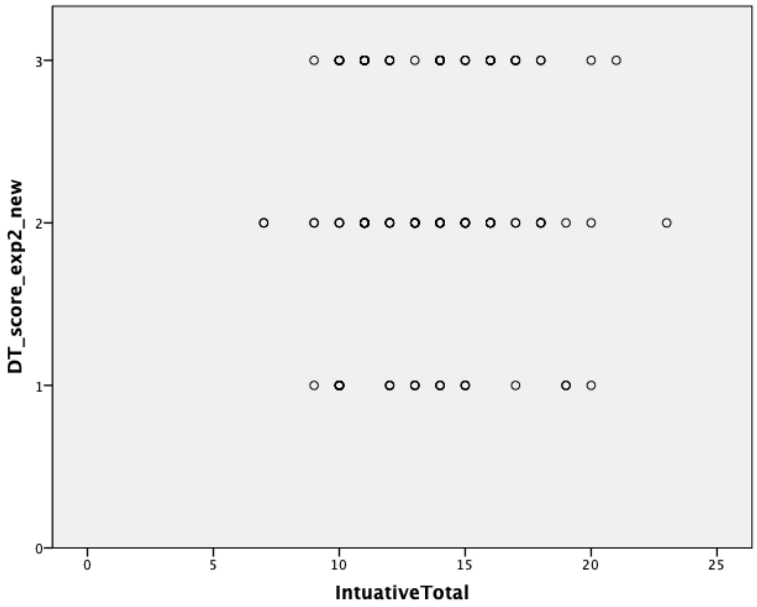


Figure 6. Scatterplot testing linear relationship between CV2 and DSS Intuitive

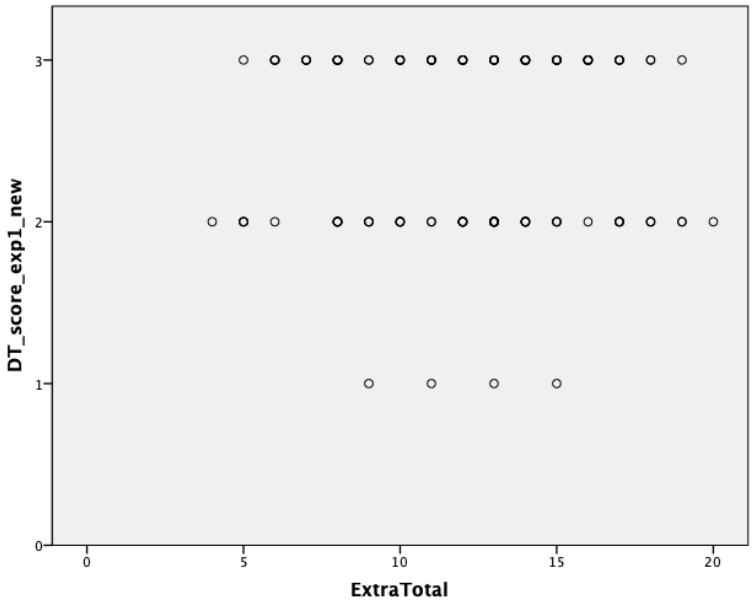


Figure 7. Scatterplot testing linear relationship between CV1 and Extraversion

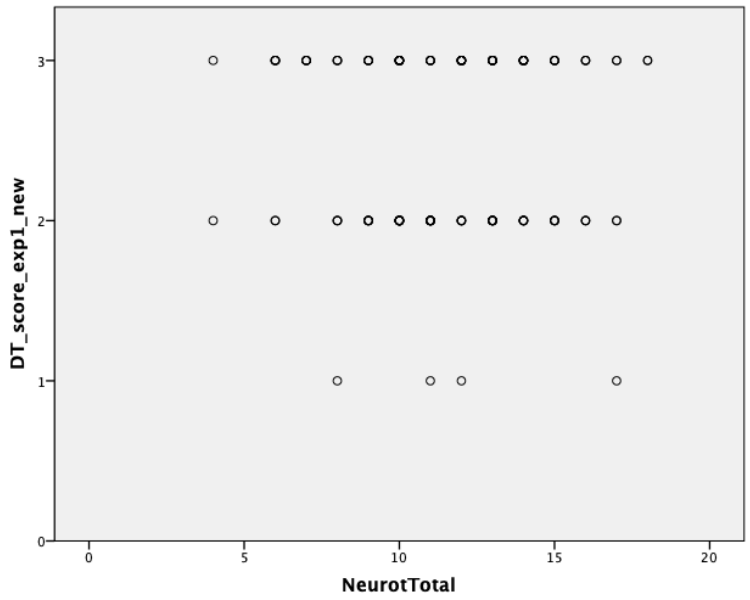


Figure 8. Scatterplot testing linear relationship between CV1 and Neuroticism

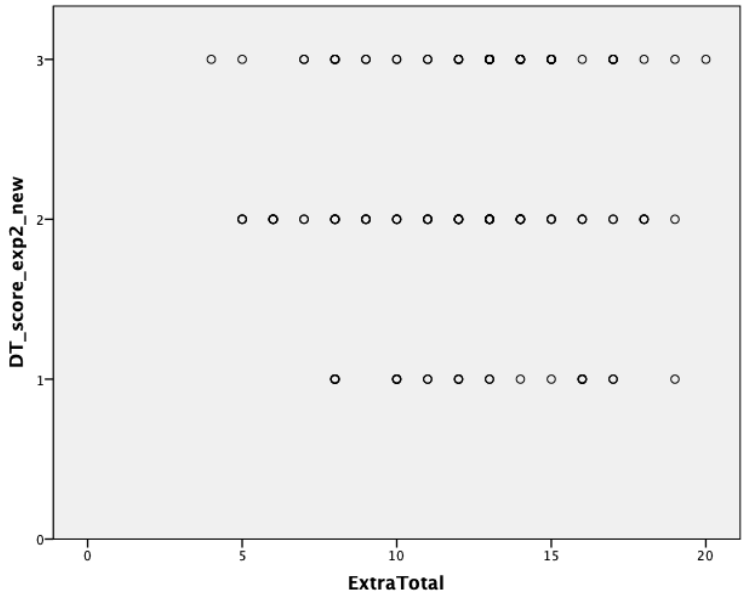


Figure 9. Scatterplot testing linear relationship between CV2 and Extraversion

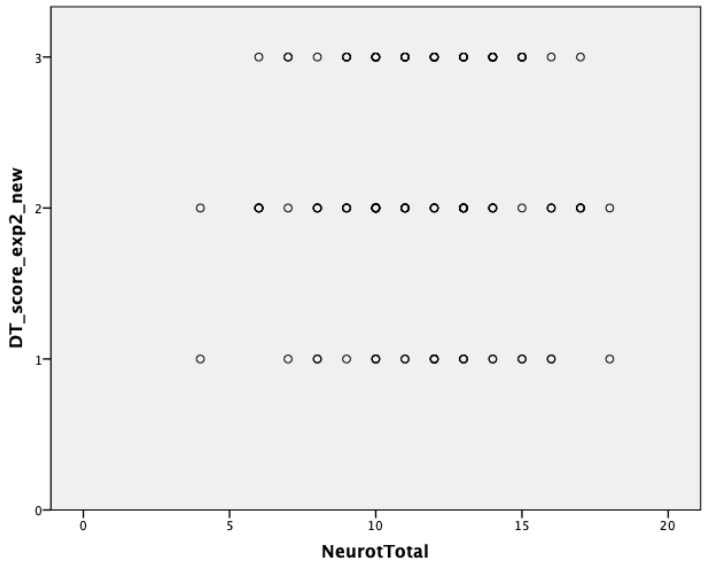


Figure 10. Scatterplot testing linear relationship between CV2 and Neuroticism

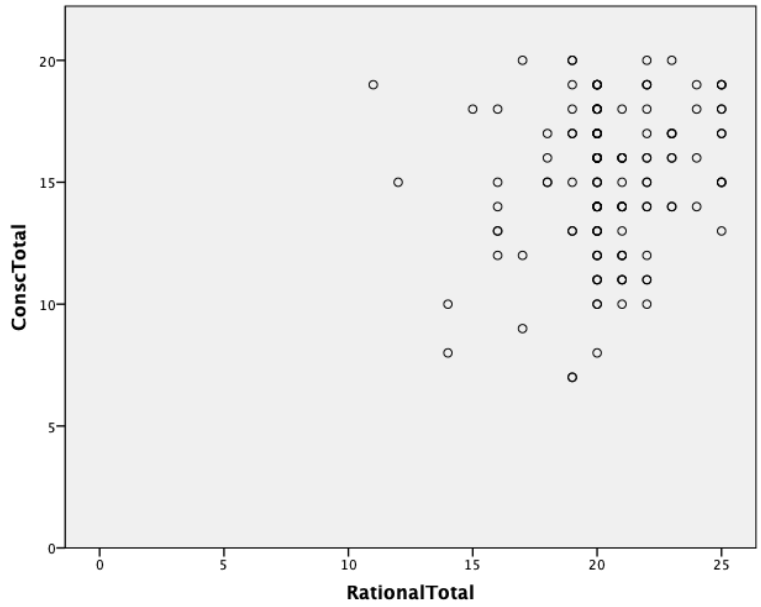


Figure 11. Scatterplot testing linear relationship between Conscientiousness and DSS Rational

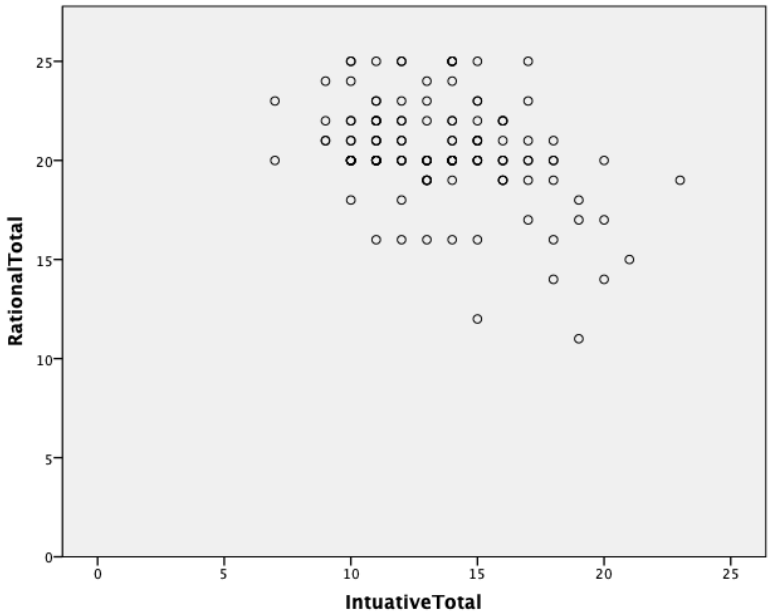


Figure 12. Scatterplot testing linear relationship between DSS Rational and DSS Intuitive

Scatterplot testing assumption of linearity and homoscedasticity between dependent and independent variables "collectively" in multiple regression for CV1/CV2

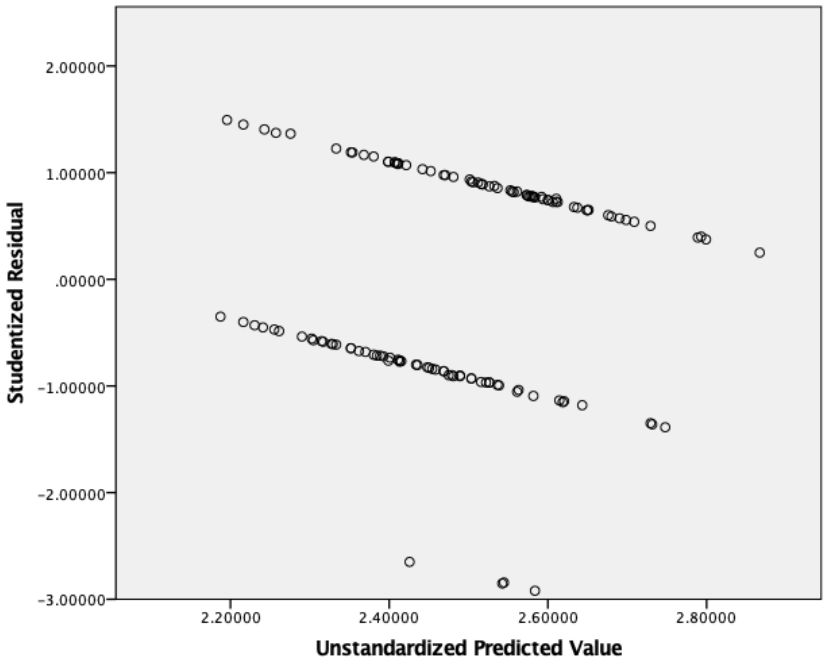


Figure 1. Relationship between dependent variable and independent variables for CV1

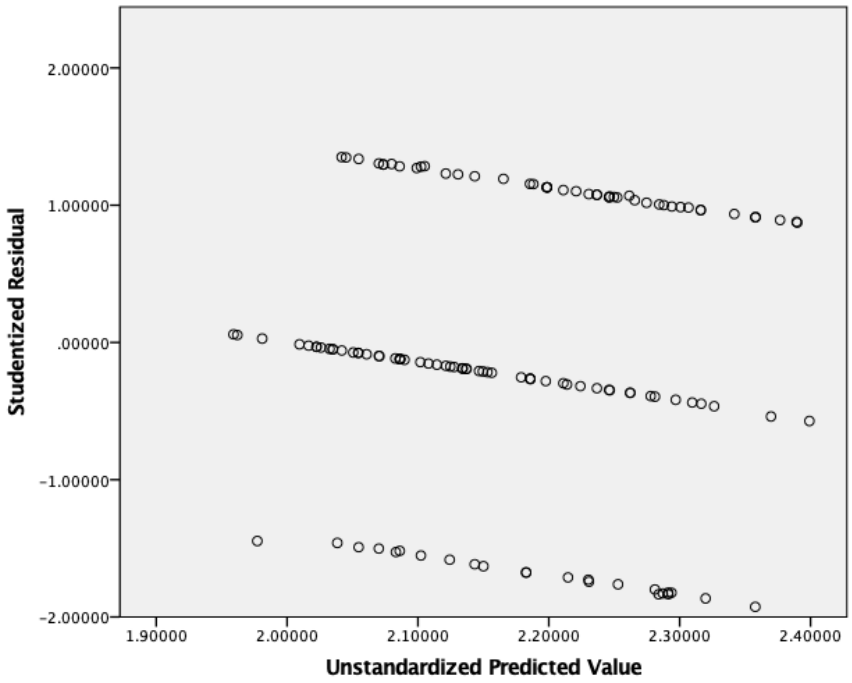


Figure 2. Relationship between dependent variable and independent variables for CV2

Scatterplot testing assumption of linearity between dependent variables and "each" of the independent variables for CV1/CV2

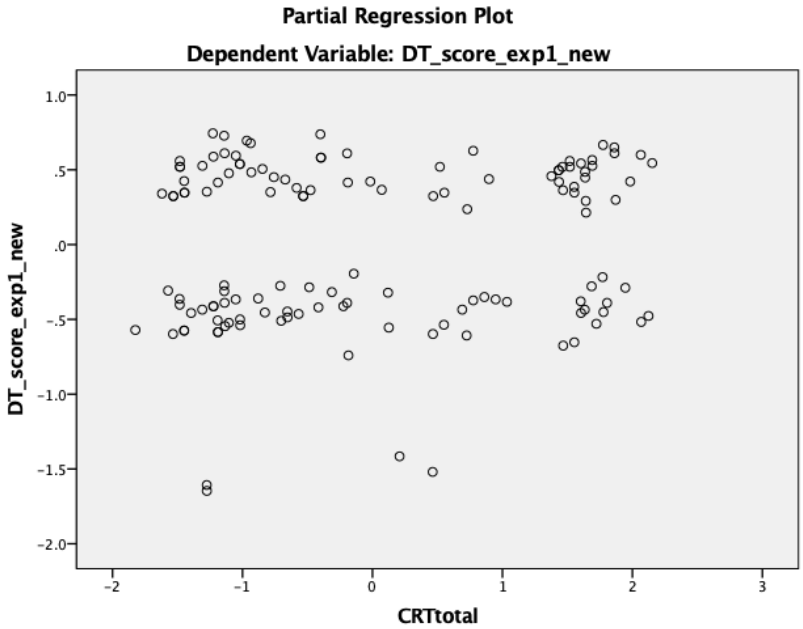


Figure 3. Relationship between dependent variable and CRT for CV1

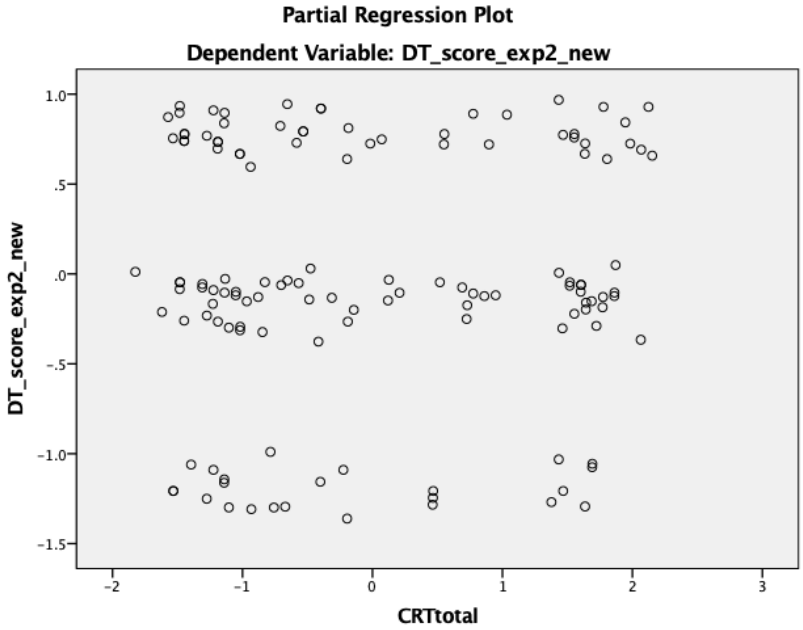


Figure 4. Relationship between dependent variable and CRT for CV2

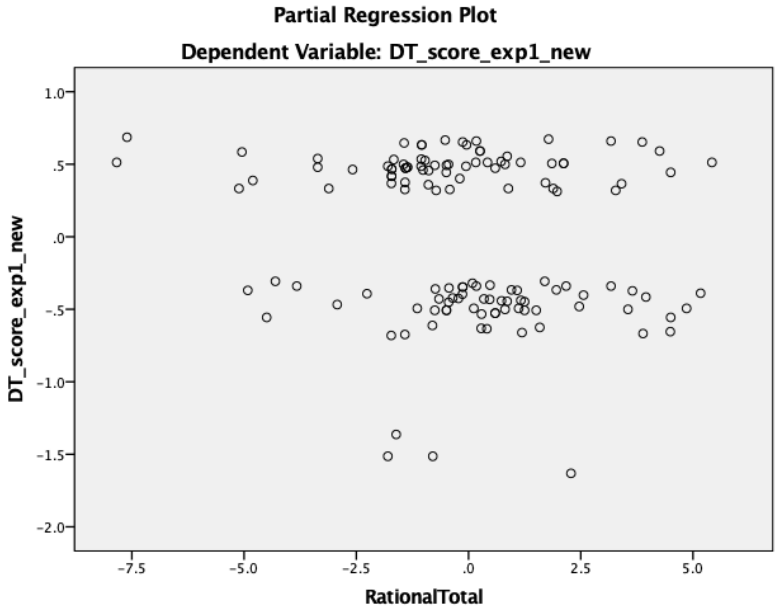


Figure 5. Relationship between dependent variable and DSS Rational for CV1

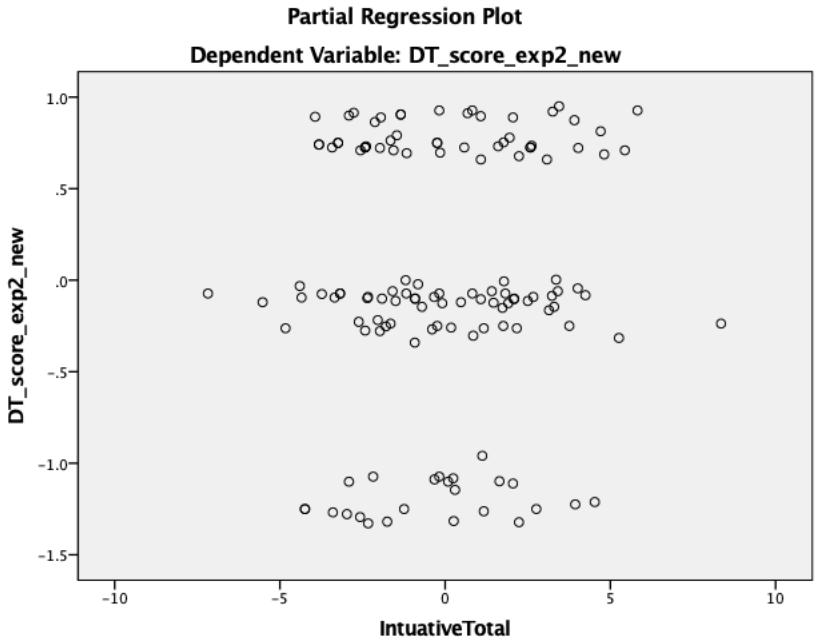


Figure 6. Relationship between dependent variable and DSS Rational for CV2

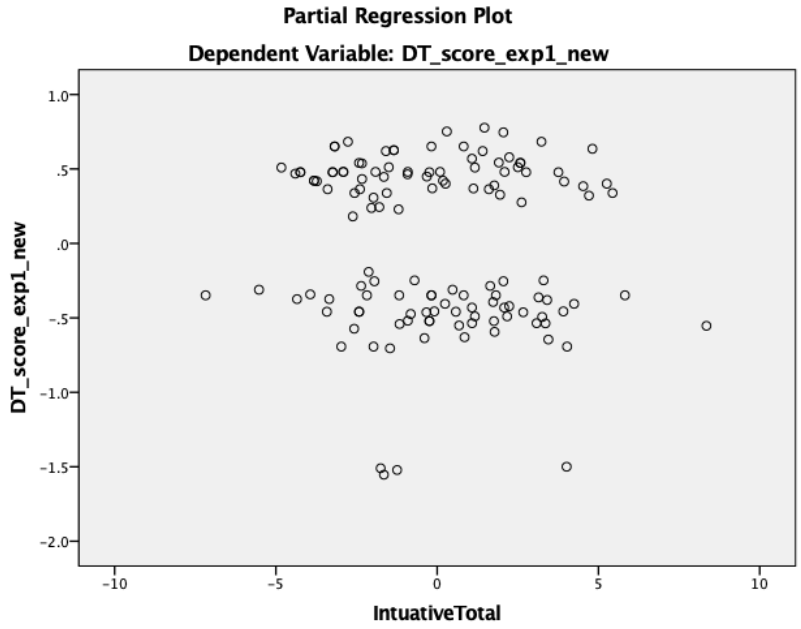


Figure 7. Relationship between dependent variable and DSS Intuitive for CV1

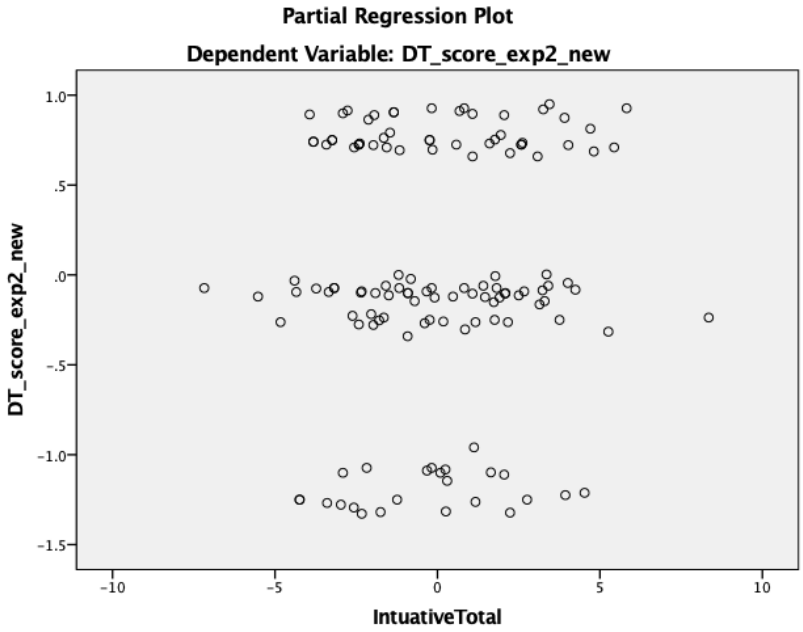


Figure 8. Relationship between dependent variable and DSS Intuitive for CV2

Appendix Q.

Summary of Pearson Correlations between static and dynamic measures

Variables	1	2	3	4	5	6	7	8	9	10
1. Dynamic Measure CV1	-									
2. Dynamic Measure CV2	.04	-								
3. CRT Measure	.12	.03	-							
4. DSS Rational Subscale	-.15	.06	.06	-						
5. DSS Intuitive Subscale	-.04	.04	-.20*	-.36**	-					
6. Extraversion Total	-.03	.04	-.03	.05	.11	-				
7. Agreeableness Total	-.03	-.14	-.09	.01	.14	.29**	-			
8. Neuroticism Total	.02	.03	.12	-.15	.01	-.29**	-.05	-		
9. Conscientiousness Total	-.06	.06	-.05	.19*	.02	-.06	.04	-.07	-	
10. Intellect Total	.04	-.20*	.10	-.01	.19*	.09	.10	-.01	-.08	-

*p < .05. **p < .01. ***p < .001