Development and Evaluation of a Valid and Reliable Footprint Measurement Approach in Forensic Identification

Sarah Mai-Lin Reel

059000068

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

York St John University

School of Health and Life Sciences

December 2012

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given where reference has been made to the work of others.

Publications arising from this work:

Reel, S., Rouse, S., Vernon, W. & Doherty, P. (2010) Reliability of a two-dimensional footprint measurement approach. *Sci Justice,* 50, 113-8.

Reel, S., Rouse, S., Vernon, W. & Doherty, P. (2012) Estimation of stature from static and dynamic footprints. *Forensic Sci Int,* 219, 283.e1-283.e5.

The publications arose from the work that is presented in Chapters 6 and 7. All authors contributed to the writing for these publications.

Acknowledgements

I am most grateful for the invaluable support and thought-provoking guidance generously given to me by my supervisors, Professor Patrick Doherty, Dr Simon Rouse and Professor Wesley Vernon OBE. Without their unfailing enthusiasm and encouragement from the beginning to the end, this doctoral thesis would not have been possible.

I am indebted to the volunteers who took part in the study, and to my research assistants, Nor Razaob, Snehal Pakhare, Hannah Smith and Samantha Lambert.

I would also like to thank friends, colleagues and family members for their belief in me, especially my parents.

Finally, I would like to express my love and gratitude to my sons Joe and Adam who lit the spark that allowed me to pursue my interest in forensic podiatry. Most importantly I would like to thank my husband Jim, who was forced to share my journey without complaint and with unremitting love and support.

Abstract

Introduction: Bare footprints found at a crime scene can be used as forensic evidence to link a person to the incident using comparison methods. Identification relies upon methods of evaluation including measurement; however the science underpinning measurement in this field has not been fully explored.

Method: A critical review of the literature revealed various measurement approaches and also demonstrated little or no measurement rigour in terms of reliability and validity. Therefore a novel pragmatic method for collecting and measuring two-dimensional bare foot impressions was developed by the researcher to provide the necessary tool for use in this field. Evaluation involved three static and three dynamic footprints collected from thirty female and thirty one male volunteers using an inkless paper system. The footprints were digitised and widths, lengths and angles constructed and automatically measured using freely available measurement software. Measurement rigour was pursued using modern validity and intra-/inter-rater reliability approaches followed by an evaluation of the tool by experts in the field. These explorations are presented within the thesis as separate investigations.

Results: Statistically significant differences occurred between paired static and dynamic linear measurements (df 60) with t values ranging from 3.08 to 23.17, $p < 0.01$. The highest correlations with stature were shown to be the linear measurement from the heel to fifth toe print in the dynamic footprints ($r = 0.858$, $p < 0.01$). The reliability analysis found high intra-rater agreement using intraclass correlation coefficient (ICC) 0.99 with a 95% standard error of measurement 0.84 mm, 95% limits of agreement (LOA) −0.91 to 0.65.

Conclusion: The research establishes a valid and reliable two-dimensional measurement approach, useful for footprint identification purposes and also as a baseline method for further research in this field.

Table of Contents

List of Tables

List of Figures

Preface

'Courted Expert Steps on Toes With Footprints. Louise Robbins is known in the legal profession as a 'hired gun,' an expert witness whose testimony has helped prosecutors put more than a dozen men behind bars and on Death Row. Though prosecutors and judges have hailed this world-renowned forensic anthropologist for the development of a new science, defense attorneys and experts believe innocent men are facing death sentences and imprisonment because of her courtroom statements that she can identify people from their footprints.' Chicago Tribune, April 6th, 1986 (Gibson, 1986, page 1).

'Courts Trample Life's Work. An RCMP expert says feet leave an impression in shoes as distinct as a fingerprint. Appeal courts in Ontario and the US say his evidence isn't reliable enough. An Ottawa case is being appealed to the Supreme Court.

By the late 1990s, after developing a database of thousands of footprints, Sgt. Kennedy was qualified as an expert in court and testified at several trials in Canada and in other countries. However, after two appeal courts set aside murder convictions based on Sgt. Kennedy's testimony – one in Ottawa – the validity of the discipline is again in question.' The Ottawa Citizen, March 22, 2004, (Rupert, 2004, page 1).

The above excerpts taken from two North American newspapers summarise the precarious backdrop that lies at the heart of this thesis. The courtroom convictions described in these articles were deemed unsafe because the science underpinning the method of footprint identification had not been established. Given this situation it is apparent that this gap in understanding needs to be filled.

Chapter 1
Introduction

In a forensic context, comparison of the shapes of bare foot impressions has been utilised in criminal investigations for identification purposes in order to associate, or disassociate a person with a scene of crime (Bodziak, 2000; Kennedy et al. 2003; DiMaggio & Vernon, 2011). An overview of the processes involved in bare footprint identification is detailed in Appendix A.1. The analysis and evaluation of footprints in this context involves both a subjective and objective interpretation by way of measurement. The latter aspect of this interpretation forms the basis of the research contained in this thesis.

Before embarking on any formal research, the researcher had attended a forensic podiatry workshop in which the delegates were invited to pair up and take inked footprints from one another. The prints were later measured and compared using a simple pen-and-ruler method. The researcher and her 'foot printing' partner failed to label the footprints produced as advised by the instructor. As a result the sheets of footprint impressions were difficult to differentiate as they shared similar measurements and overall shape, including toe patterns. The researcher and fellow delegate were of a similar height, weight and ethnic background. This experience fascinated the researcher, especially in light of the work by Kennedy et al. which suggested that there is a 1 in 1.27 billion chance that one person shares the same footprint shape with another, inferring that footprints are unique (Kennedy et al., 2005).

The work of Kennedy et al. came under the media spotlight after Kennedy gave testimony in several court cases in Canada and the United States (Hansen, 1993; McRoberts et al., 2004). Between 2001 and 2009, four guilty verdicts centred on Kennedy's footprint evidence were appealed and overturned. Kennedy had erroneously applied his team's research conclusions considering the uniqueness of inked bare footprints collected under clinical conditions, to insole prints inside footwear. It was on this basis that the validity of his barefoot impression analysis was scrutinised, and the evidence ruled unreliable, described by the presiding judges as 'junk science' (State v Berry, 2001; State v Jones, 2001; R v Dimitrov, 2003; State v Jones, 2009). These decisions reflect

ongoing attempts in the United States (US) to develop a rule that will exclude unreliable evidence from a trial. Previously this has been driven by toxic tort litigation (Daubert v Merrill Dow Inc, 1992) but more recently government driven initiatives have lead to policies regarding admissibility in both the US and the United Kingdom (National Research Council, 2009; The Law Commission Report, 2011), the rationale and context of which will be discussed in the thesis literature review.

Footprint evaluation for identification purposes involves the concepts of measurement. During the course of the research, it became clear that there were differing interpretations of the concepts of measurement in the three areas of forensic practice, science and law-driven policy and initiatives. Although in the context of measurement the terms 'reliability' and 'validity' sat happily in each of the three camps, they were defined differently. Perhaps this discord is unsurprising; as far back as the fourteenth century others noted that science and law made unhappy bedfellows (Redmayne, 2001; Berger & Solan, 2008) and this apparent disparity between measurement concepts in research and measurement applied in the field prompted further interest. Primarily a health professional, the researcher was familiar with the principles and practices of evidence-based practice in medicine and had expected parallels with the identification forensic sciences. However, it was clear that these principles were absent in the practice of the majority of the forensic identification disciplines. At the time of writing, this situation is currently being addressed by the UK Forensic Science Regulator (Rennison, 2011).

The forensic podiatry workshop had offered a different method of measuring footprints to that published by Kennedy et al. prompting the researcher to wonder if other approaches were being employed. If so, which footprint measurement approach would enable a measurement tool to be acceptable in forensic practice, science and law? If no such measurement method exists, the foundations on which the science of footprint impression evidence is based would appear fragile.

To answer this inquiry, the following research objectives were proposed;

1) To critically review the literature for footprint impression measurement approaches

2) To evaluate the extent to which reliability and validity have been utilised in measurement

3) To develop a new pragmatic approach to footprint impression measurement underpinned by high levels of validity and reliability.

Chapter 2
Critical Appraisal of the Literature

In the forensic arena, two-dimensional bare foot impressions are analysed and compared with others for identification purposes, as indicated in Chapter 1 and summarised in Appendix A.1. The objective interpretation of the footprint evaluation relies on measurement of the footprints of interest. The following literature review will explore the differing measurement concept philosophies between three areas; law, science and forensic practice. It will then review literature pertaining to measurement approaches currently used by forensic practitioners and also explore footprint measuring methods beyond the realms of forensic science. Finally this chapter will present a framework for the thesis, built upon the emergent questions from the literature review.

2.1 Measurement concepts within forensic identification science

This section will explain the circumstances in which bare footprints are found at a crime scene and reflect on current practices in the interpretation and evaluation of these prints.

2.1.1 Identification through marks left by a bare foot

Criminal investigations rely on the sound gathering and analysis of evidence relevant to the crime in order for a subsequent prosecution to be made (DiMaggio & Vernon, 2011). In some crime scene situations, a perpetrator may leave physical traces of their presence. If the perpetrator was unshod at the scene, it is possible that they may have transferred residue between the foot and the substrate (such as blood), to the floor (Bodziak, 2000). This then leaves a mark of the plantar surface of that person's foot which is, in effect, a 'two-dimensional representation of a three-dimensional structure' (Cole, 2007, page 272). If such evidence is made on a soft surface such as mud or sand, a three-dimensional footprint will result; however the work involved in this thesis will focus on two-dimensional footprints as these may be more commonly found at crime scenes (DiMaggio, 2004). Two-dimensional prints are made by the transfer of residue to a hard surface such as a floor or door. Bare footprints may

be found at scenes of sexual offences or murder, and are more prevalent at crime scenes in countries of a warmer climate (Sharma, 1970; Qamra et al., 1980). They can be left on a hard surface by a variety of substances such as dust, oil, blood, paint and mud (Qamra et al., 1980).

A foot impression may be detected by crime scene personnel, for example the Scene of Crime Officer in the UK, or the Crime Scene Investigator in the US. Once detected, attempts are then made to recover the footprint for further analysis; as such evidence may associate or exclude a suspect from that crime scene. This process usually involves enhancement of the footprint followed by photography by specialists in the field (DiMaggio & Vernon, 2011). Enhancement may involve the use of chemicals such as Luminol which reacts with blood, or the use of different lighting variables (Bodziak, 2000). The photographed crime scene foot impression (known as the question or unknown print) is subsequently analysed. It is then compared with actual and photographed foot impressions collected from a person linked with the incident. These donor footprints can be made in the same, or a similar substance to the traces left at the scene of crime, but most often in ink. The donor footprints are described as the exemplar or known prints (DiMaggio & Vernon, 2011).

Recovered bare footprints only occasionally display ridge patterns like those found in fingerprints (Sharma, 1970; Kerr, 2000; Johnson, 2008) therefore evaluation and comparison of footprints usually involves measurement of the outline or shape. Such analysis is undertaken by a variety of forensic disciplines including forensic podiatrists, marks examiners, anthropologists and specially trained members of the police force (Robbins, 1978; Laskowski and Kyle, 1988; Kennedy, 1996; Borkowski, 2002). The role of the forensic podiatrist has become more prevalent in recent times, catalysed by various peer-reviewed research-based publications and regulatory and professional body recognition (Vernon, 2009). Forensic podiatry is defined as 'the application of sound and researched podiatry knowledge and experience in forensic investigations, to show the association of an individual with a scene of crime, or to answer any other legal question concerned with the foot or footwear that requires knowledge of the functioning foot' (Vernon & McCourt, 1999, page 47). This definition refers to the discipline as a whole and includes four main areas;

identification using podiatry treatment records, footwear identification, forensic gait analysis and bare footprint identification.

Despite publications in this area dating back more than ninety years, forensic podiatry is regarded as a specialism 'still in its relative infancy' (DiMaggio & Vernon, 2011, page v). Formal recognition of the discipline was acknowledged in 2007 after approval from the Board of the International Association for Identification, the largest and oldest professional organisation for members of the forensic science community (Polski, 2007).

Previous errors highlighted in the media by other novice disciplines such as those interpreting bite mark (Pretty, 2006) and ear print impression evidence (Moenssens, 1995; Stripp, 2008), have catalysed the need to establish solid foundations both in practice (based on empirical research) and in its regulatory and educational initiatives.

At the time of writing, there is no formal mandatory regulation of forensic practitioners in this field; however the Health and Care Professions Council's 'standards of conduct, performance and ethics' document (HCPC, 2012) relevant for the regulation of podiatrists, apply also to those engaged in the practice of forensic podiatry (Urwin, 2012). The UK Forensic Science Society supported by the forensic regulator, has recently championed moves for competency examinations with the aim of creating a public-facing register for successful applicants (Ostell, 2011). M level studies in forensic podiatry are now offered at University of Huddersfield (2012) and combined with various post-graduate research projects the speciality is actively building the foundations necessary for acceptance in the wider forensic community and beyond.

2.1.2 Challenges to forensic identification science

Criticism of the use of Kennedy's 'junk science' in court as outlined in Chapter 1 of this thesis is not a new phenomenon in the field of identification. In the 1980s, anthropologist Dr Louise Robbins appeared as a footprint and footwear evidence expert in more than twenty cases in the US and Canada, mainly for the prosecution. Her evidence contributed to the sending of twelve people to prison and one to death row (Allen, 2004). In some cases, her testimony constituted the only physical evidence linking the defendant to the crime (Hansen, 1993). Unfortunately, her beliefs regarding footprints and shoe outsole

wear patterns in an identification context had no scientific basis and were subsequently found to be 'unreliable' by a panel of more than one hundred forensic experts, who concluded that her footprint identification techniques lacked validity (McRoberts et al., 2004, page 8). The word 'unreliable' used by the panel describes their opinion that the evidence was untrustworthy. This definition differs from the scientific use of the word which reflects the extent of repeatability between one or more tests. Confusion regarding the concepts of reliability and validity across the disciplines will be discussed further in this thesis. There is debate as to whether the unquestioned acceptance of this type of evidence amongst jurors outweighed the probative value in the Robbins and Kennedy cases, a phenomenon now popularly referred to as the CSI effect (Cole & Dioso-Villa, 2009). However, it is apparent that the lack of a scientific basis underpinning bare footprint impression evidence may have been the primary reason why the prosecutions in these cases were unsafe.

Bare footprint evidence was not the only forensic identification discipline under scrutiny. All forensic identification sciences barring nuclear DNA analysis, have been criticised as having a lack of scientific basis, in that empirical testing including investigations of reliability and validity have not been explored (Pretty, 2006; Cole, 2007; Saks & Faigman, 2008; Saks and Koehler, 2008). Included in this group are fingerprint, handwriting, bite mark, voiceprint, tool mark, firearm mark, tyreprint, footprint and shoeprint evidence, and have been referred to as the 'non-science forensic sciences' (Saks & Faigman, 2008, page 150). The criticisms have been more pronounced in the US in which the forensic identification sciences have traditionally evolved in the back rooms of police departments with little or no academic involvement from higher education institutes (Saks & Faigman, 2008; National Research Council, 2009). Certain applications of forensic practice that have developed from this non-academic background have recently been highlighted by the media. For example, the laboratory of the Federal Bureau of Investigation in Washington, D.C., considered a gold standard within US forensic analysis community, was criticised after results from hair and fibre analyses were ruled unsafe (Hsu, 2012). According to Hsu, 'hundreds of defendants' have been misidentified as a result of the laboratory failing to adopt a scientific approach to hair and fibre analyses (Hsu, 2012, page 1). Forensic identification sciences in the UK are not

without criticism either. The R v T appeal case, for example, highlighted the consequences of an approach being presented as scientific, when in fact it was not (R v T, 2010). The appellant claimed he was wrongfully linked with shoe impressions found at the crime scene. In the initial trial in which the defendant was found guilty, an expert experienced in foot wear marks had concluded that there was 'a moderate degree of scientific evidence to support the view that the (Nike trainers recovered from the appellant) had made the footwear marks' (R v T, 2010, page 9). An approach based on likelihood ratios was used to form this opinion. Although not a novel concept, the expert had formulated his opinions by using inferential statistics from a relatively small database (eight thousand, one hundred and twenty two shoe prints) and applied these to a real-world setting in which approximately forty two million pairs of shoes are sold each year. The witness did not inform the court of the size of the database from which his conclusions were drawn. A statement prepared by Professor Jamieson of the Forensic Institute (UK) read, 'I am not disputing (the expert's) opinion, but the scientific basis of it. It is my opinion that the state of development of this expertise is insufficient to ascribe any more than a rough approximation to the probative value of the evidence, and such opinions cannot be considered scientific' (page 12). The expertise in this case was opined to be that of a subjective, rather than an objective nature and given that this had not been made transparent, the conviction was quashed. This emphasised the need for more scientific research to be undertaken, not only in shoe impression evidence but also of other identification disciplines. Saks & Faigman (2008) claim the basic hypotheses for all such sciences, including latent fingerprint identification, have never been tested in any rigorous or systematic way. Not only was there a lack of empirical scientific testing within these forensic sciences, but also a requirement for protocols to guide the practitioner.

A case outlined in another recent appeal, demonstrated a counter-argument for the necessity of a scientific approach in identification (Otway v R, 2011). In this particular case, the appellant argued that the closed circuit television footage placing him at the scene of crime was dependent upon a subjective rather than an objective scientific analysis, again supported by Professor Jamieson. However, in this case, the appeal was overturned as the admittance of the subjective comparison evidence was deemed acceptable. The decision was

based on the fact that the expert witness had presented his opinions with transparency. Despite a lack of a scientific basis, the evidence presented was allowed by virtue that it had been clear that this was an opinion based on expertise alone.

Emerging from a health-related background, the researcher was familiar with working with an evidence-based medicine (EBM) model as informed by organisations such as the National Institute for Health and Clinical Excellence (NICE). Guidelines produced by NICE involve the systematic reviewing of the most current and valid research findings as the foundation for clinical decisions (National Institute for Health and Clinical Evidence, 2009). In areas of medicine not covered by NICE guidelines, a clinical practitioner would aspire to working within the doctrine underpinning the EBM model (Greenhalgh, 2004). It did not seem implausible then, that a similar principle is applied to the forensic identification arena, with practitioners basing their approaches on peer-reviewed methods that have demonstrated scientific rigour. This next section considers whether an evidence-based practice (EBP) model exists for use in forensic practice or its related policies.

The definition of EBM by Sackett et al. (1996) states that it is 'the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients...(by) integrating individual clinical expertise with the best available external clinical evidence from systematic research' (page 71). EBM begins and ends with the patient. It is clear that when dealing with forensic impression evidence, the terms 'patients' and 'clinical evidence' within the definition are not applicable. However, Sackett et al. (1996) advise that their approach delivers a framework only, and in the forensic scenario these words can be substituted with 'the interpretation of forensic evidence' instead of 'care of individual patients', and 'forensic science evidence' instead of 'clinical evidence'. The required research evidence in all arenas should be of the highest scientific quality (Sackett et al., 1996). That being said, Guyatt et al. (2000) note that sometimes research evidence is not enough, and a practitioner's experience and judgment should also contribute to their decision-making in practice. The outcome of the Otway appeal case outlined above supports this notion (Otway v R, 2011). At the time of Kennedy's and Robbins' testimonies, research regarding footprint impression evidence was limited and it

can be argued that these two experts acted in good faith in applying theories of evidence-based practice, as well as their own expertise. However, there was no evidence-base to link knowledge pertaining to bare footprints with insole prints; this being a leap of faith without scientific justification.

2.2 Measurement concepts within the area of law

This section will explain how recent criticisms of measurement evidence are being addressed through law-driven initiatives.

2.2.1 Law-driven recommendations

In simple terms, the aim of criminal law enforcement is to identify people who have committed offences and to prevent erroneous convictions of the innocent (Ashworth, 2006). The legal system relies on evidence to support the case for either the prosecution or defence teams. Individualisation forensic evidence aspires to be scientific, both in the method of recovery and in the analysis. The National Research Council of the US identified an important qualification for the admissibility of and reliance upon forensic evidence in criminal trials. The organisation stated that it is 'the extent to which a particular forensic discipline is founded on a reliable scientific methodology that gives it the capacity to accurately analyse evidence and report findings' (National Research Council, 2009, page 87). Two questions emerge from this statement; could footprint evidence successfully meet the demands of this qualification, and how are the terms reliability and accuracy defined in this context? The results from data analyses produced from the research and presented later in this thesis will attempt to answer the former question. The answer to the latter will be discussed within the remainder of this chapter.

The advancement of science has produced various standards for scientific evidence to be admitted in a court of law. The first of these in the US in 1923 saw a landmark case declare the results of a lie detector test *unreliable*. The presiding judge decided the technology had 'not gained general acceptance in the scientific community' (Frye v. United States, 1923, page 6). Further developments culminated in the case proceedings of Daubert v. Merrell Dow Inc. (1992) giving rise to the Daubert test which examines the *reliability* of an

item of evidence or a scientific technology intended to be used in a factual dispute. The test requires evidence to meet these four requirements:

1. Verification of the theory or technique through tests,
2. peer review and publications,
3. known error levels,
4. general acceptance within relevant scientific community.

The Daubert test goes some way in gate keeping the acceptance of new forensic technologies such as foot impression evidence. However, application of the Daubert ruling has been sporadic and its utilisation has been inconsistent and lacking in clarity (National Research Council, 2009). To add to the ambiguity the Supreme Court responsible for the ruling described the Daubert standard as flexible, therefore offering no clear substantive standards by which trial judges admit or exclude evidence or expert testimony. *Reliability* as defined by the four points of the Daubert test can be considered a reflection of trustworthiness incorporating concepts of validity rather than the scientific understanding of reliability which will be explored further in this section. Parallels with the Daubert test and the philosophies underpinning the theories of EBP are apparent. Law-driven initiatives as exemplified by the four points of the Daubert test are clearly in favour of EBP- type criteria from which to admit evidence into a law-court. Whether these criteria are observed in practice is debatable.

The state of forensic sciences in the US was scrutinised in a highly critical report published by the National Academy of Sciences (NAS) (National Research Council, 2009). In particular, the report condemned forensic techniques which determine the source item that leaves a trace at a crime scene, such as a footprint. The committee criticised the fragmentation and inconsistent practices of these individualisation forensic sciences and also the 'noticeable dearth of peer-reviewed published studies establishing the scientific bases and validity of many forensic methods' (page 8). It identified many weaknesses in areas of accuracy, reliability as well as validity in these disciplines and highlighted the general lack of knowledge concerning the accuracy of various techniques even under ideal conditions.

The NAS committee agreed that the lack of protocols regarding the validation of scientific techniques prior to admissibility in court was completely unsatisfactory. Recommendation 3 of the report calls for 'the development and establishment of quantifiable measures of the reliability and accuracy of forensic analysis' but does not offer further explanations of these terms (National Research Council, 2009, page 23).

Although the NAS publication is a US commissioned report, it does have implications for UK-based forensic science disciplines (Roberts, 2009). The Forensic Science Advisory Council which included representatives of the Home Office, the Bar, the Crown Prosecution Service, the police, the Forensic Science Service (now disbanded), Forensic Alliance and LGC Forensics, commented that in light of the NAS report, 'there are concerns over the validation of technology used in cases', and recommended that 'the level of confidence in a new technology has to be clarified' (Forensic Science Regulation Unit, 2009, page 4).

A previous report by the House of Commons Science and Technology Committee also discussed concerns regarding the apparent lack of scientific validation within forensic communities in the UK. Paragraph 173 relays the worries of the Association of Chief Police Officers; 'To a large extent we are at the mercy of the criminal justice system as we have no agreed method of getting new techniques validated' (Gibson et al., 2005).

More recently in March, 2011, the UK Law Commission published a report regarding concerns over evidentiary reliability in response to various wrongful convictions in the UK (R v Dallagher, 2002; R v Clark, 2003; R v Cannings, 2004; R v Kempster, 2008), prompted by the House of Commons Science and Technology Committee report. Again, the term reliability here is used to describe trustworthiness. The Law Commission (2011) offered principal recommendations for restructuring the law relating to expert evidence in criminal proceedings. The recommendations regarding expert opinion evidence admissibility include *reliability testing* to decide if the expert opinion is scientifically sound. To clarify the test further, the publication recommends the trial judge should examine 'the extent and quality of the data on which the expert's opinion is based, and the validity of the methods by which they were

obtained' and also 'if the expert's opinion relies on an inference from any findings, whether the opinion properly explains how safe or unsafe the inference is (whether by reference to statistical significance or in other appropriate terms)' (The Law Commission, 2011, page 140). The parallels with this publication and the NAS report are palpable. Regardless of ambiguity over definitions of reliability and validity, both advocate that scientific evidence should underpin expert opinion evidence proffered in a court of law.

Whilst the work of Kennedy et al. had been peer-reviewed and published, the application of the knowledge resulting from their bare footprint studies to that of foot insole impressions could be described as a leap of faith. It was eventually recognised as such in the US/Canadian court appeals and would not have stood up to the above recommendations set out by the UK Law Commission and the NAS report regarding the admissibility of expert evidence. The researcher was interested to discover if the methods used for routine evaluation and comparison of footprints for forensic identification purposes were truly lacking in validity, as the reports had suggested. This led her to explore the area of footprint identification in both in practice and in research where she hoped to find valid support for the applied technology.

Anecdotal evidence from UK, US and Australian forensic examiners[1] revealed a variety of measurement approaches currently employed to evaluate and compare footprint shapes for identification purposes, as noted also by Vernon (2007). The lack of consensus amongst practitioners regarding a universally accepted measurement approach was of concern to the researcher, especially in light of the governmental reports. Recommendation 2c of the Law Commission report calls for evidence to be rejected from a trial if 'the expert's opinion relies on the results of the use of any method (for instance, a test, measurement or survey), which has not taken proper account of matters, such as the degree of precision or margin of uncertainty, affecting the accuracy or reliability of those results' (The Law Commission, 2011, page 139). This concurs with the forensic science regulator's codes of practice and conduct document which requires that all measurement methods used for forensic identification purposes should provide evidence of validity (Rennison, 2011). In

[1] Footprint evidence meeting, 94th International Association for Identification Educational Conference, Tampa, FL, 20th August, 2009.

order to compare measurement, a tool that has been scientifically validated must be employed. If many approaches were currently being utilised in the field, could this suggest a lack of an appropriate method or conversely a range of equally rigorous methods? It was hoped a search of relevant literature would reveal an answer.

2.3 Measurement concepts within the area of science

It is apparent that the words *reliability* and *validity* are important in both forensic practice and the policies outlined above. But what do these mean in terms of scientific measurement research? Related to reliability and validity are other terms such as *precision*, *accuracy* and *consistency*. This next section will discuss these concepts in the context of the thesis.

2.3.1 Reliability, validity, precision, accuracy and consistency

In the forensic field and in a court of law, the terms *reliable* and *valid* are often used, but confusingly are not defined in the same way as the scientific determination of the meanings. As discussed in section 2.2.1, reliability in legal terms often refers to evidentiary reliability. The Law Commission further defines this by suggesting that the evidence 'must be sufficiently reliable, that is, sufficiently trustworthy, to justify being admitted before a jury.' (The Law Commission Report, 2009, page 34). In his widely cited publication 'Unified theory of scientific evidence', Black defines validity in law thus; 'The evidence is scientifically valid if it results from sound and cogent reasoning' (Black, 1988). This definition is not far removed from the scientific meaning; however, Black proposes substituting the word *valid* for *reliable* in the context of scientific evidence - a suggestion supported by the Law Commission consultation paper (The Law Commission Report, 2009)

Similarly, in the scientific context, the word *validity* is inextricably linked with *reliability*, but the two terms would not be interchangeable, as in the above example. This illustrates the various differences between the disciplines of law and science, despite important scientific inferences that exist within the legal setting. For a test or technology to have rigour, it must be able to demonstrate both reliability and validity. In science, if an experimental procedure, measurement or test is not reliable, then it cannot be valid. However, although

reliability is necessary, it cannot ensure validity alone (Robson, 2002). *Reliability* refers to the consistency of scores over repeated testing (Baumgartner, 1989) and relates to the freedom of the scores from measurement error (Wood, 1989). *Validity* can be defined as the extent to which a test measures what it is supposed to measure (Hicks, 2005). According to the literature, there are various forms of validity that can be divided into two groups; experimental and test measurement validation (Safrit & Wood, 1989; Stevens, 1993). Examples of experimental validity include internal, ecological, statistical conclusion and external (Stevens, 1993). Test measurement validation relates to the appropriateness of the interpretation of the scores of a test, for example; construct, face, criterion, and content validity (Safrit & Wood, 1989; Stevens, 1993; Robson, 2002). Concerns that have been raised regarding the validity and reliability of scientific evidence in court usually pertain to a technology or technique used for evaluation, e.g. fingerprint or footprint comparison involving physical measurement; therefore validity and reliability are discussed here mainly in terms of test measurement.

Measurement is the 'process of assigning numbers to properties of objects, organisms or events according to some rule', facilitating objectivity and thus enabling verification by others (Safrit, 1989, page 371). This process allows researchers to relay results with greater precision. The term *precision* can be described by the appropriateness of the scale to the task and the extent to which a measurement obtained by one measurement is matched by a second measure (Streiner & Norman, 2006). Leedy notes that reliability deals with accuracy in the sense of degree of precision, (Leedy, 1993) but Streiner and Norman argue that a measurement can be precise but is not necessarily reliable. They conclude precision does not reflect the ability of the test used to differentiate among individuals (Streiner & Norman, 2006) although the use of the standard error of measurement (SEM), to be further discussed in this thesis, can be construed to facilitate this (Denegar & Ball, 1993; Myers et al., 2007). Related to the term precision, is *consistency*, another important characteristic when evaluating measurement tests. Consistency is a term that is used to reflect a lack of, or in reality a small variation between the paired scores of repeated tests and can be demonstrated by the use of 95% limits of agreement graphs (Bland & Altman, 2003). A small standard deviation of paired differences

between a rater's test scores would imply his results were consistent. The term *accuracy* is often used synonymously with validity, for example if the scores from a new measurement test are found to be comparable with the scores from an established gold standard test, and is an indication of the closeness of measurement results to the real value. For example, Myers et al. (2007) compared the test scores collected manually by clinicians with simultaneous collections by a machine (the gold standard). They estimated the accuracy of the test scores by calculating the mean difference and absolute mean difference between the results from the clinician and the results from the instrument-obtained data. However, Messick (1995) argues for the investigation of various forms of validity to assess a new technique or measurement, and the accuracy estimated for each test of validity, although a universally accepted definition of accuracy has yet to be identified. In the absence of a gold standard, or indeed any comparison test, determination of accuracy of a new technology may not be possible. In the presence of a gold standard test, a measurement approach can be defined as being precise but not accurate, accurate but not precise, both or neither. For example, if a trial displays systematic error, then a sample size increase generally improves precision but does not increase accuracy. Accuracy and precision are often used synonymously in the areas of law, forensic practice and the policy drivers of forensic practice, but are deliberately contrasted in the scientific disciplines.

Concepts of validity and reliability are discussed in further detail in the next section (2.3.2).

2.3.2 Validity concepts in measurement

Comparison of test scores from a new technique with a gold standard is essentially an investigation of criterion validity. Other types of criterion-related evidence are concurrent validity, in which test scores are collected simultaneously with results from another test, and predictive validity in which data can be predicted from previously acquired results. The extent of criterion and its related validities can be demonstrated through correlational testing.

Criterion validity is one of the three Cs of the trinitarian view of test validation; the others being content and construct validity (Cronbach & Meehl, 1955). Content validity usually precedes the collection of data and cannot be

quantified. It explores the degree to which a technique should measure or address the concept it is hoping to measure and then converts the themes brought about by this exploration into concrete research questions (Stevens, 1993). Construct validity arises from this exploration in that the operational concepts being measured are examined to see if they accurately reflect the theoretical concepts (constructs) suggested by the content validity analysis (Stevens, 1993). Often the lack of content validity prohibits this initial exploration and the constructs are thus derived from the analyses of wider hypotheses, for example those derived from a criterion validity determination. Construct validity can be split further into separate components: discriminate validity and convergent validity. Discriminate validity demonstrates an absence of a relationship among measures which in theory, should not be related. Convergent validity is the opposite of this – it is the proven agreement among scores, gathered independently of each another, where measures would be expected in theory to be linked (Wood, 1989).

Face (logical) validity, another member of the test validations, can be split into four interpretations; validity by assumption, validity by definition, appearance of validity and validity by hypothesis (Mosier, 1947). The looseness of these definitions implies a lack of consensus as to the exact meaning of the term *face* validity and is now regarded as an obsolete concept of validity (Wood, 1989).

A test may rely on an intuitive judgement as to whether it measures what it claims to measure. When this judgement is made by informed individuals, it is known as 'expert' validity (Stevens, 1993). External validity may also involve the judgements and opinions of experts. This type of validity explores how far the results of the research can be generalised beyond the findings for the sample of the study in terms of populations, concepts and situations (ecological validity). Internal validity examines the strength of the design of the research. It explores the manner in which the research question was structured and the choices made regarding methods of data collection (Stevens, 1993).

It is apparent that measurement validity can be broken down into many parts. Thorough examination of appropriate types of validity within a measurement approach will enhance its relevance. For the purposes of this thesis, various

forms of measurement validity will be explored, guided by the available literature.

2.3.3 Measurement reliability

In a court of law, if evidence is deemed reliable it is thought of as trustworthy, dependable and consistent. In general terms, this is true of *reliability* when referring to measurement concepts in research. When a line is measured on a page, one may be satisfied that the corresponding notch on the ruler truly denotes the length of that line. One may however, be uncertain that the line was measured correctly and the process may be repeated again. If the same measurement is recorded, it is noted that this measure was consistent with the last and one assumes that this result can be trusted. It is reliable.

However, another may measure the same line with a different ruler and declare the line to be of a different length to one's own result. The measurement process is repeated and the same result obtained but again, different from one's own.

Each person's set of length measurements are reliable, but there appears to be no reliability from person to person. This could be because of a difference occurring in each individual's judgement as to where the line ends – maybe one person may have a slight visual impairment for example. Or it may be that the rulers were produced in separate factories and were calibrated differently. These differences or variances can be described as measurement error.

If there is no error we can say that a measurement is perfectly reliable. Unfortunately when a measurement involves the judgement of human beings, it is rarely error free. This is more apparent when the measurement is repeated many times over – plotted results often falling into a bell-shaped curve. An individual comparing a series of their own repeated results can be described as an assessment of intra-rater (intra-operator, intra-tester) reliability (Hicks, 2005).

Measurement error can occur more readily where two or more people measure the same thing. Comparing results among a number of people to measure consistency can be defined as inter-rater (inter-operator, inter-tester) reliability, examining the extent of agreement between raters (Hicks, 2005).

However measurement error not only occurs when human beings are the operators; the example above suggests the rulers used for measuring may have been calibrated differently. Measurement instruments often possess finite calibration for increased precision and accuracy. Two instruments manufactured in the same place may yield dissimilar results from one another due to temperature, vibration, and other environmental differences in the factory setting (Bruton et al., 2000). It is therefore essential that reliability studies of any new measurement apparatus are produced to ensure that such errors can be accounted for, and are small enough to detect real changes in the test results (Bland & Altman, 1986).

Clearly, a measurement tool that displays totally unrepeatable results also has no validity. But it has been acknowledged that there will always be some degree of error when dealing with continuous measurements (Baumgartner, 1989). Therefore reliability could be described 'as the amount of measurement error that has been deemed acceptable for the effective practical use of a measurement tool' (Atkinson & Nevill, 1998, page 219).

Unfortunately there is no firm definition of 'an acceptable level of measurement error', and published levels of statistical significance concerning reliability may not actually be acceptable in a laboratory or a clinical setting. So, as with much hypothetical research, statistical interpretations of reliability studies must be considered carefully before making conclusions. For example, reliability is often population specific, and this must be taken into account when looking at several comparison studies. Also, there is a variety of statistical approaches used for interpreting measurement. For continuous data (weight, mass, time, distance) the intraclass coefficient (ICC), Bland & Altman 95% limits of agreement (LOA), SEM, and the coefficient of variation (CV) are the most commonly used indices of reliability (Bruton et al., 2000). There seems to be a lack of consensus as to which of these tests is the most appropriate. Various authors have therefore argued for a variety of tests to be employed, in order for a more definitive picture of reliability to be produced, rather than just one single estimate (Safrit & Wood, 1989; Rankin & Stokes, 1998; Bruton et al., 2000).

In summary, there are many layers to the concepts of validity, reliability and accuracy. In the research arena, statistical interpretation must be taken in context with the design of the measurement study in question. The overall range and magnitude of the scores should be considered before deciding if a test indicating high reliability really is statistically significant. Reporting validity (including tests of accuracy and reliability) is the ideal requirement of hypothesis testing in experiments relating to measurement. In forensic practice, validity refers to 'the ability of the process to measure the data in question' (Speckels, 2011, page 204) which is enhanced by adequate training, certification and accreditation of practitioners (Redmayne, 2001; Speckels, 2011). If a measurement or forensic test is repeatable, it is accepted to be reliable as the two words in this arena are interchangeable. Policy makers refer to *reliability* in terms of trustworthiness; a far reaching concept with no clear boundaries, unlike the academic perception of reliability. However, if a measurement or test is repeatable, it is considered reliable, valid and consistent (National Research Council, 2009; The Law Commission Report, 2011).

There are clear differences between the three areas (research, practice and policy) regarding measurement concepts (summarised in Appendix A.2). The areas of research and forensic practice are driven by policy. Measurement concepts as described by the likes of Daubert, NAS report, the Law Commission report and the UK forensic science codes of practice and conduct document, appear to have a greater commonality with the same concepts in research than in forensic practice. Despite this, shared philosophies amongst the three areas regarding measurement are relatively few, especially between research and practice. Redmayne reflects on this disparity thus; 'It may be that the lack of academic base for forensic science ...has led us to see diversity where there should be unity' (Redmayne, 2004, page 35).

There is an apparent imbalance between the determinations of external and internal validity in that forensic science testing is weighted in favour of the former and research methods in this area favour the latter. It is clear that methods used in the individualisation sciences including footprint identification, require rigorous testing both experimentally and in the field to redress this balance. Thorough investigation into the validity and reliability of a method or measurement can offer error estimates or margins of uncertainty, for example in

the form of confidence intervals. Reporting the extent of variance and error is integral to this process and can allow researchers, judges and jurors to make more meaningful decisions and mollify the concerns highlighted by recent government initiatives.

2.4 Searching the Literature

The researcher was now intent in exploring the literature in terms of footprint measurement approaches, focusing on reported reliability and validity estimates within the articles. Of particular interest was the question of whether or not there indeed existed a 'noticeable dearth of peer-reviewed published studies establishing the scientific bases and validity of many forensic methods', as observed by the NAS report (National Research Council, 2009, page 8).

Aided and tutored by a York St John University research librarian, the researcher searched various databases using search strategy worksheet grids with Boolean connecters (MEDLINE Course Materials, 2005). For example, keywords such as footprint* AND measure*, footprint* AND identification, forensic AND impression*, forensic AND podiatry, footprint* AND evidence, footprint* AND criminol*, foot* AND anthropol*, etc. were entered into search engines. Databases included MEDLINE, CINAHL, AMED, Wiley Online Library (Law and Criminology), ScienceDirect, WestLaw, SIGLE and Google Scholar. In addition, ZETOC alerts were set up for currency. The researcher was also in receipt of publications from organisations including the International Association for Identification and the Forensic Science Society. Greenhalgh & Peacock (2005) observed the importance of a multi-method approach to literature searching and noted that systematic reviewing cannot rely solely on the articles issued from computerised databases. Therefore serendipitous searching, including hand searching and author searching guided by references from other articles, recommended by authors including Rumsey (2008) and Montori et al. (2005), was adopted by the researcher in addition to traditional searching strategies.

Critical appraisal tools such as those afforded by Critical Appraisal Skills Programme (Burls, 2006), Appraisal of Guidelines Research and Evaluation (The AGREE Collaboration) and CATmaker (Oxford Centre for Evidence-Based

Medicine) were considered useful in guiding the researcher through the appraisal process. The articles were eventually critiqued using an appraisal tool modelled on Hicks (2005) and Law et al. (1998) as this tool questioned all aspects of the publications relevant to the research (Appendix B.1).

A specific tool for scoring evaluated forensic science literature could not be found by the researcher. Instead, a recognised system used for grading medical literature, the Oxford Levels of Evidence system (OLE) was chosen, recommended by Phillips et al. (1998) (Appendix B.2). This method offers a hierarchy system for evaluating and assessing research evidence in the medical field. It is recognised that this method of grading scientific evidence is useful when attempting to systematically review the pertinent literature (Harbour & Miller, 2001). Evaluation of this nature has been recommended for forensic science literature also (Smith, 1996; Cole, 2007). The lowest level of evidence is given to expert opinion (level 5) and the highest level of evidence (level 1) awarded to systematic reviews of randomised control trials (RCTs). The sub-sections of the highest level mostly concern studies involving RCTs, and it can be argued that disciplines outside medicine (such as the forensic sciences) share no commonality with the levels of evidence prescribed by the Oxford centre, making this method of evaluation irrelevant (Kroke et al., 2004). However, RCTs are not exclusive to the medical sciences and have been carried out in forensic research (Roman et al., 2008). In response to a report by the UK government reflecting on management, quality and use of science in the Home Office and Ministry of Justice, the Royal Statistical Society advised that the Home Office 'strongly promote the use of well designed experiments, including randomisation, for evaluation purposes' (Review of Science in the Home Office, 2003, page 14). Conversely, RCTs can be considered to be unfeasible, unethical, and at times impossible to carry out (Cole, 2007). Therefore the OLE system for grading the quality of the relevant literature was not used in isolation, as it afforded a determination of one aspect of article appraisal only.

To ensure confidence in the results reported by the relevant articles, an additional evaluation tool was used to determine validity and reliability in specific areas not described by the OLE system. This tool was modelled on a worksheet for evaluating therapeutic articles developed by the Ohio College of

Podiatric Medicine (OCPM) (Turlik & Kushner, 2000). The scores using this tool range from 0 to a maximum of 24 (Appendix B.2).

An appraisal of articles concerning human footprint measurement with OLE and OCPM grading is presented in the next section.

A general literature search using the databases AMED, CINAHL and MEDLINE was carried out with the search terms 'footprint* AND human NOT carbon NOT gene* NOT immunol*'. This preliminary search uncovered eight hundred and sixty one results. The search was narrowed further by eliminating articles that did not fall into the subject heading of 'foot', resulting in sixty two articles. These were then reduced to thirty seven by excluding articles that fell under the major headings of, for example, 'leg', health knowledge', haplorhini', etc. In a further serendipitous search, a further twenty three relevant publications were found.

2.4.1 Methods of footprint collection and justification for further refining of the main literature search

The resulting sixty articles uncovered footprint studies that utilise various methods for capturing two-dimensional footprint impressions, including inked and electronic methods. These studies often explore the function of the medial longitudinal arch (Lin et al., 2004; Chen et al., 2006), or assess arch height (Shiang et al., 1998). Geometric measurement of electronic footprints are used for footwear design (Hawes et al., 1994a; Mochimaru & Kouchi, 1997), biometrics (Nakajima et al., 2000; Jung et al., 2004) and also for classification of foot types (Mathieson et al., 2004; Wearing et al., 2004). Inked footprint studies are found in the research areas of forensic identification (Kennedy, 1996; Qamra et al., 1980), anthropology (Krishan, 2007; Fawzy & Kamal, 2010) and medicine (Kippen, 1993).

Studies that analyse two-dimensional footprints describe various methods of collection. In Jasuja et al.'s study participants are asked to stand in their bare feet on jute bags, soaked in water soluble black ink before making prints on a smoothly plastered floor (Jasuja et al., 1997) or a white sheet (Jasuja & Manjula, 1993). Water soluble black ink has also been favoured in other studies (Kippen, 1993; Barker and Scheuer, 1998) while poster paint was used by Nikolaidou & Boudolos (2006) in their study investigating foot types of schoolchildren. The Harris and Beath Mat (Harris & Beath, 1947) has been

used in static (standing) and dynamic (walking) footprint studies (Cobey & Sella, 1981; Welton, 1992; Chockalingam & Ashford, 2002,) while cyclostyling ink was used in Qamra et al.'s study (1980) in the analysis of identification factors in bare footprint impressions.

The use of cyclostyling ink, poster paint and water soluble ink spread on the plantar surface of a foot will create a good two-dimensional footprint impression if the same person then steps onto a hard surface (linoleum, wood, laminate flooring) covered with a capturing medium such as light-coloured paper. It has been suggested however, that variability occurs between footprints if the foot is over- or under-inked. Anecdotal reports suggest that an over-inked foot will produce a footprint with larger dimensions than an under-inked print from the same foot (Bodziak, 2000), although formal research regarding this phenomenon is yet to be published.

The Harris and Beath mat was primarily designed in 1947 to record foot to ground pressure patterns aiding clinical diagnoses (Harris and Beath, 1947; Rose et al., 1985; Welton, 1992). It consists of a rubber mat made up of regularly placed horizontal and vertical ridges set at three different heights. The surface of the mat is inked and covered with a semi-absorbent sheet of paper. The volunteer stands or walks on the mat to form a print and the greater intensities of pressure on the sole of the foot incurred by ground reaction forces are recorded as darker areas. Problems arise when measuring footprint dimensions using the Harris and Beath mat because the actual outline of the foot is often incomplete, due to the nature of the grid-like ridges imposed by the mat on the impression.

Perhaps a cleaner method giving rise to more clearly defined footprints compared with traditional inking methods and the Harris and Beath mat is the inkless paper system (Kennedy et al., 2003; Kennedy et al., 2005). This consists of a mat containing an odourless, colourless ink which covers the plantar surface of the foot. The volunteer then places the foot on the inkless paper containing a chemical substrate, immediately developing a clearly defined black footprint upon contact (Figure 2.1). Targeting the small area of the inkless paper may be problematic for dynamic footprint collection and is further discussed in Chapter 3, section 3.1.

Figure 2.1 Impression made by Inkless Paper System

Inked footprints are often measured by the construction of lines drawn over the prints using a pen, ruler and protractor for example in the studies published by Jasuja et al. (1997), Barker & Scheuer (1998) and Krishan (2008a; 2008b). This conventional method, however, can incur operator error and thus affect reliability of the resulting measurements (Grčar et al., 2006). Digitisation of footprints can overcome these problems with various options available, ranging from highly sophisticated medical imaging processes to simpler, automated measurement software (Hawes et al., 1992; Mochimaru & Kouchi, 1997; Sforza et al., 1998).

The option of using an electronic footprint capture method with incorporated computer interfaces for automatic calculation of foot dimensions is seemingly a more advantageous approach than using a more conventional method, such as inked two-dimensional footprint capture. Repeatability is increased when automatic measurements are performed, compared with the manual measurement of inked footprints, as operator bias is minimised (Grčar et al., 2006). However, the sensors in the responding surface that replicate areas of foot contact produce blocks of information on the resultant image and do not represent the true outline of a footprint. This is supported by the literature. For example, Chu et al. (1995) discuss the issues concerning the poorly delineated and irregular boundaries of electronic footprints and advise that this may affect repeatability and introduce error when gathering data. Urry & Wearing (2005) report that the Musgrave Footprint method of footprint collection (Musgrave Systems Ltd, Wrexham, North Wales) produces irregular borders, stating on

page 204 that '...the linear, angular and area measurements obtained from them may differ from an ink print of the same foot.' This may account for the statistically significant differences between the mean long plantar angle (inner and outer tangent angle) values of the inked verses the electronic footprints reported in their study. Current systems used for the collection of plantar pressure distribution data, similarly project footprint images that are not analogous to their inked print counterparts. Visual examples of these systems can be found on the internet and include Rothballer Footscan® (www.mar-systems.co.uk), Footscan® (www.rsscan.co.uk) and Emed® (www.novel.de).

Studies involving electronic footprint capture facilitate greater statistical testing of the chosen method of measurement compared with existing forensic footprint measurement studies, due to the inclusion of software components. However, in the forensic research arena, inked footprints are more representative of the types of two-dimensional footprints recovered at a crime scene formed from mud, dust, blood, oil, paint, etc, than electronically captured prints. As discussed in section 2.1.1, it is preferable for the forensic comparison between a donor print and a crime scene print to be made in a similar substance, to recreate the mark left by the unknown foot as far as possible. Often this is unachievable because of ethical considerations, for example if the mark appears to have been made in blood, or because the original substance cannot be identified. In these cases, ink is the preferred substance to recreate a footprint (DiMaggio & Vernon, 2011). The validity of the transference of electronic measures to inked footprints is questionable and it is for this reason that the researcher has critically appraised publications pertaining to footprint measurement approaches obtained by inking methods only.

2.4.2 Footprint evaluation methods

Evaluation of two-dimensional footprint impressions for comparison purposes are carried out in the field of forensic identification by one or more of the following methods; overlay, Robbins, Gunn, Rossi and the Optical Centre Method (Vernon, 2007). The first part of this literature review will appraise the literature underpinning these five approaches and offer a hierarchy of levels of evidence, using the aforementioned OLE and OCPM systems for grading literature. Using the basic search terms forensic* AND footprint* resulted in thirty three articles. Exclusion criteria such as papers pertaining to genetics,

footwear, chromatography and ballistics narrowed the results of the search further to just two articles. Retrieval strategies outside the realms of computerised databases including reference list and author searching increased the final results to six. A critique of the main footprint evaluation methods in a forensic context derived from these six publications are discussed next.

2.4.2.1 The Overlay Method

In the overlay method, the forensic practitioner places a sheet of clear acetate over the question print. The question print may be recovered from the crime scene and presented either in its true form (for example a bloody footprint on a piece of flooring, removed in its entirety), or as a latent print (for example on lifting film) but more often than not, as a life-sized evidence photograph (Bodziak, 2000). A fine-tipped marker pen is used to trace around the outline of the exemplar print and the acetate then transferred and placed over the question print (Vernon, 2007). Authors Smerecki and Lovejoy (1985) published a police case report in which the methodology used to link a partial footprint with the suspect is reviewed. The footprint found at the crime scene was of the forefoot only and clothed in a sock. Inked impressions were taken of two suspects. Half of the impressions were bare footed and half were taken when wearing a sock. One suspect was eliminated immediately as her forefoot shape 'did not resemble' the questioned print. However, the other suspect's print appeared to have similar features to the partial print found at the crime scene. A further test was designed to ascertain the likelihood that a person from a more general population could possibly share the same contours of the questioned print. Inked impressions were therefore collected from ninety-five female and five male volunteers with US shoe sizes ranging from three to ten, height from under five foot to above six foot and weight between seven and fourteen stones. A sheet of acetate was placed over the print to be analysed and the outlines of the shape of the forefoot traced. Thirteen different characteristics were used as identifying landmarks, for example, the degree of hallux valgus/varus relative to the metatarsal impression. A second independent test to improve the rigour of the design used a further thirty five pairs of female footprints. From this larger combined sample of one hundred and thirty five paired footprints, fifty single-blind and random paired prints were selected. Of these, only twelve pairs of prints from the sample of fifty were examined of which some pairs contained the

suspect's footprint, others without. All of the suspect's prints were identified using this method. The main author, a US detective sergeant, concluded that 'the results demonstrate the reliability of the non-exclusion method applied to partial prints as no non-exclusions were incorrectly made' (Smerecki and Lovejoy, 1985, page 189). The definition of reliability is not confirmed in this article as it offers no supporting statistics. It highlights the gap between science and practice and identifies the strong relationship between validity and reliability. The high probability of identifying a person's footprints in amongst such a small yet widely heterogeneous sample is surely a certainty. As previously discussed in this chapter, the UK court case, R v T, demonstrated the error in forming these types of specific conclusions from a general population base. Reliability in the case reported by Smerecki and Lovejoy was dependent upon the validity of the design of the study which may have been flawed, as the experimental database consisted of only twelve pairs of footprints. The authors state that the probability of the suspect owning the footprint found at the crime scene was set at 2205:1 and the jury convicted the female suspect of murder. It is not clear how this likelihood ratio was calculated; the method using the transparencies is not fully described nor the method of footprint collection. Statistical analyses of the findings are not published. Using the OLE system for grading literature, this could be classed as of level 5 quality; 'Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles",' (Phillips et al., 1998, page 1). Using the OCPM method for determining the validity of an article, this paper scores 1 out of a maximum of 24 points.

A study by Maltais & Yamashita (2010) describes the involvement of fifteen forensic practitioners who compared latent (question) prints with inked impressions with the intention of identifying a correct match for each latent print. More than one inclusion in each set of ten inked prints to a latent print was not considered an error, as positive identifications were not asked for. All latent prints were correctly identified by the experts, with the exception of one. The article concludes that the barefoot comparison 'techniques' employed in the study are valid (page 408). This comment is possibly referring to convergent validity. Unfortunately the 'techniques' used are not fully described, although the paper suggests that comparisons are carried out by observing the 'class

characteristics of the bare foot's shape or morphology as manifested in the impression of the weight-bearing areas of the foot' (page 363). Examples of class characteristics in footprint impressions are the shape and pattern of the toes, the overall length of the foot and the shape and positioning of the balls and heels of the foot (Bodziak, 2000). The overlay method may have been an appropriate technique for this type of comparison exercise; instead observational skills only were employed in this study by Maltais & Yamashita (Yamashita, 2009). There are no statistical inferences made in the article and using the OLE system for grading literature, this may also be classed as level 5. A result of 1 is derived using the OCPM grading system for validity. With the advancement of digital photography, the overlay method can be adapted in the field of software enhancement and superimposition, for comparison of exemplar and question prints (Vernon, 2007). No validation tests have been published to date regarding this method.

2.4.2.2 Robbins Method

In 1978, anthropologist Louise Robbins collected inked footprints from five hundred male and female volunteers ranging from eight to seventy-nine years of age, for comparison with prehistoric footprints found in caves in the United States of America (Robbins, 1978). The aim of this study was to differentiate the number of prehistoric persons responsible for the large number of prints found in the caves and also to attempt to determine age, sex, stature and weight from the questioned footprints by comparing with the known contemporary group of prints. The author used her own method for taking measurements of both groups of footprints for comparison. Length, width and angle measurements from various points of reference were manually drawn and quantitatively analysed using acetate metric grids.

The footprints were also subjectively analysed by examining the shapes and contours of different parts of the footprint, for example in describing the shape of the big toe pad.

Footprint lengths and widths between subjects displayed comparable measurements (within 5mm of each other), but the study concluded that footprints are unique, in that no other person in the sample shared the same outline shape. However Robbins notes that walking (dynamic) footprints are

subjectively different to stationary (static) ones. The paper provides a commentary on the author's own observations but does not include any statistical analyses.

The design of this study is questionable as the prehistoric footprints are retrieved by a method of plaster-of-Paris casting, producing three-dimensional impressions that are inappropriately compared with two-dimensional inked footprints. The study sets out to further knowledge concerning a group of prehistoric people known to have resided in caves in Tennessee. In turn, the study aims to investigate the individuality of human footprints. These aims are unfulfilled as its results and conclusions do not refer to the prehistoric footprints, reflecting on the reduced extent to which content and construct validity are explored. The paper concludes that a person's footprints are unique which is a sweeping statement considering the complete absence of reported data and analysis. This would be of a level 5 grade using the OLE system and scores 1 point according to OCPM grading for validity.

In a later article, footprint measurements are further clarified with the use of labelled diagrams, descriptive statistics and right/left foot outlines, barefoot print/foot outlines, stature and weight correlations (Robbins, 1985). The author concluded that the uniqueness of footprints may be found in a print's overall shape, rather than in its measurement variables. The data from the 1978 study was used to investigate further the association between foot dimensions with the height and weight of a person (Robbins, 1986) and can be graded level 4 material according to the OLE system, as it was a retrospective study with an absence of sensitivity analyses, and used a heterogeneous sample. Robbins' published work has since been openly criticised regarding the studies' designs and mathematical errors in the interpretations of results (Tuttle, 1986; Kahane & Thornton, 1987; Giles & Vallandigham, 1991; Gordon & Buikstra, 1992). Regardless, her footprint measuring methods continue to be used in some areas and adapted (Barker & Scheuer, 1998; Krishan, 2008a).

2.4.2.3 Gunn Method

In a case report describing a comparison method between a crime scene bare footprint and a suspect's print, a sheet of acetate is utilised to facilitate the evaluation process as previously described by Smerecki & Lovejoy but

modified to incorporate a grid for ease of reference (Gunn, 1991). The author of this paper, Dr Norman Gunn, is considered to be the pioneer of forensic podiatry (Vernon, 2007) and appears to be the first podiatrist to have undertaken actual casework. The article describes the recovery of footprints found in a sandy substrate by a river near to the body of a murdered toddler. The author used plaster-of-Paris to take casts of the adult imprints and photographed this evidence from several angles for further analysis. Socked footprints were made by the suspect whilst in police custody in the same substrate at the site and also bare footprints made in a 'puddle-cast' (page 8). These were also recovered using plaster-of-Paris and photographed in the same way. The socked and bare footprints were then compared. Additionally a set of socked imprints were made at the site by a third party carrying a weight of approximately thirty five pounds to simulate the weight of the child, and again casts taken and photographed from different angles for analysis. Finally, the author himself made a positive cast of his socked left foot whilst carrying the same weight, and a puddle-cast made of his bare foot whilst semi weight-bearing and photographed. The photographs were assembled accordingly and an acetate graph of the characteristics of the impression produced for each cast. This acetate was then placed over the insole impression recovered from the suspect's shoe and the similarities noted using the overlay method described by Smerecki & Lovejoy (1985). Furthering the approach, a series of lines were constructed on the acetate from the base of the heel print to the tips of each of the five toe prints, amongst other measurements.

The similarities between the exemplar impressions from the suspect and the question prints convinced the author that they belonged to the same person as there appeared to be no differences between the measurements. This is despite the fact that 'allowances were made for the thickness of the sock' (page 8) in the comparison analysis of the bare foot impressions with the socked impressions. Descriptive data regarding the other two subjects (including the author) are not reported. Discrimination between the suspect's prints and those of the other two subjects would be highly probable due to the small comparison sample size, threatening internal validity of the case study and allowing for a false identification. In the absence of correlation analyses, the construct validity of this published case is also questionable. A separate analysis in the article

uses the application of photogrammetry, whereby two-dimensional and three-dimensional images are used to create an exact copy of the casts using computer modelling. This experiment was done to compare the suspect's bare foot cast with his socked foot cast. The result was that the two casts 'left absolutely no doubt that the same individual belonged to both casts' (page 11). The author's reference to the 'great accuracy' (page 11) of the approach reflects on the misinterpretation of this word in forensic practice. The statement may have inferred that if the measurement or test is accurate, then it is valid. However, accuracy can only be determined if a test is proven valid first and appropriate correlational testing with a suitable standard has been carried out to pre-determine an acceptable level of accuracy (section 2.3.1). The author finally concludes that the paper shows that both approaches for measuring footprints are 'valid and true' (page 11). However, there is no supporting statistical evidence to provide assurances of either reliability or validity of the study. The mixture of comparisons between two-dimensional, three-dimensional, bare foot, socked foot and insole impressions threatens internal validity. If the photogrammetry approach had been applied in order to achieve concurrent validity with Gunn's own method of measurement this would have improved the trustworthiness of the measurement approach. Many of Dr Gunn's linear measurements for evaluating footprint impressions in a forensic capacity are presently employed by forensic analysts (Vernon, 2007). The article by Gunn offers a detailed description of the investigative processes, and the application of software to replicate the footprint offers a scientific approach to the research. It describes a simple test to try and replicate the circumstances of a crime and as such, account for any variables that may have affected the bare footprint. However, it is of a case-study nature and does not include a statistical analysis. It can therefore be rated as level 5 evidence and scores 0 using the OCPM method for grading the validity of an article.

2.4.2.4 Rossi's Podometrics System

Many anthropometric, orthopaedic and podiatric studies involve measuring the feet from various populations in order to categorise into types, such as pes planus (flat feet), pes cavus (high arched feet), pronated feet, etc. (e.g. Randall et al., 1951; Cobey & Sella, 1981; Staheli et al., 1987; Welton, 1992; Hawes & Sovak, 1994; Mathieson et al., 1999; Wearing et al., 2004; Stavlas et al., 2005;

Nikolaidou & Boudolos, 2006). This categorisation can be used for application in a biomechanical, anthropological, forensic, clinical, and also footwear design context. Pertaining to foot categorisation in this latter field, Rossi (1992) devised a system named 'podometrics' for 'mapping' the foot, for determining foot type in terms of shoe design (page 301). Podometrics involved a cartographical system of foot typing rather than classification based upon physical measurements of the foot as had been used in the past. Rossi acknowledges the large range of 'normal' foot types and his system of podometrics incorporates this variability of the human foot (page 301). The method was to be used as a baseline in the evolution of further studies, and indeed incorporated as a reference measurement in the footprint uniqueness studies undertaken by Kennedy et al. (2003; 2005). Rossi's design involved a variety of measurements taken from the anatomical proportions of the feet of over eight hundred subjects. The author states that the purpose of the resulting data was to establish the reliability of the method, and not for accumulation of measurement data for general analysis. However, the outcome of the reliability analysis of the method is not reported. This is not surprising as it is a validation, rather than a reliability study (in which a repeated-measures type of study would be expected). The author has possibly used the term 'reliability' to mean validity (page 301). The study also aimed to establish construct validity but without relevant data or indeed a results section, it can be described as commentary only and therefore merits a level of evidence grading of 5 and an OCPM score of 4.

2.4.2.5 Optical Center Method

The uniqueness of the human footprint has been reported in many studies (e.g. Robbins, 1978; Cassidy, 1980; Qamra et al., 1980; Laskowski & Kyle, 1988; Kennedy, 1996; Borkowski, 2002; Kennedy et al., 2003; Kennedy, 2005; Kennedy et al., 2005). One such study conducted in 1986 at the US Federal Bureau of Investigation, examined the inked footprints of three hundred and ninety nine males and one hundred and one females (Bodziak, 2000). The optical centres of the toe pads from the footprints, as well as the optical centres of the heel prints were marked on the impressions. Determination of optical centres in this publication are not properly described; however the author explains that the same methodology was applied in footprint measurement

study later presented by Kennedy et al. (2003). Here, Kennedy identified optical centres using AutoCAD software. Polylines were constructed around each toe and heel print and the software marked out centre points determined from these outlines.

Other unspecified reference points were also noted. A metric grid transparency was then placed over each footprint and the optical centres of the heel and second toe pad lined up vertically. These enabled easier identification of the most lateral and medial aspects of the metatarso-phalangeal joint (MPJ) area and were subsequently marked out. A line drawn between these two points denoted the widest part of the forefoot, and the addition of a line defining the y-axis allowed the main body of the footprint to be divided into four portions and the outlines traced. Forty four different points of reference were recorded, measured and compared using an unspecified software program. A ±5mm bracket was allowed for the linear measurements (e.g. heel to toe length), to 'account for any variations in the impression process' (page 388). The study concluded that only 'five or fewer of the most general characteristics were necessary to either identify or discriminate these footprints from all others in the study' (page 388). The study does not include data to outline the statistical methodology used, nor does it detail the software used to calculate the optical centres, the inter-comparison of the raw data or the tracing of the footprints. The choice of measurement approach is not referred to and the 'tracing' (page 388) of the footprint outline may be subject to error. The study suggests that a person's foot leaves a repeatedly consistent impression time after time. Unfortunately, results from statistical testing to support this statement are not reported. The OLE system of grading literature would categorise this paper of being of level 5 as it is commentary only and does not offer any in-depth analysis of data. It scores 2 in the OCPM method of grading the validity of an article.

The five approaches outlined above appear to be lacking in demonstrable measurement rigour. This may be due partly to the practical environment in which some have been developed.

Literature outside the five popular methods used by forensic practitioners was also considered. It was thought that other footprint measurement approaches

may offer a more robust method and stand up to the rigours of the Law Commission's reliability test, the forensic science regulator's Codes of Practice and Conduct document and the recommendations of the NAS report. For relevancy, the search was confined to the measurement of inked human footprint impressions. Critiques of these articles are summarised in the following tables listing published articles pertaining to footprint measurement with their associated hierarchy of quality. The tables are sorted by measurement type; linear, foot indices, Arch Index, Footprint Angle and Chippaux-Smirak Index. The details regarding these measurement types are discussed in section 3.4.

Table 2.1 Summary of critical appraisal of literature pertaining to footprint measurement (linear measures).

Author(s) (Date)	Application	Strength of study overall (OLE grading 1-5)	Strength of study: validity (OCPM Score Max 24 pts)	Strength of study: reliability investigated? (yes/no)
Jasuja et al. (1991)	Anthropology/ Identification	4	10	No
Kippen (1993)	Clinical	5	4	No
Kennedy (1996)	Identification	5	4	No
Barker & Scheuer (1998)	Identification	4	4	No*
Kennedy et al. (2003)	Identification	5	2	No*
Kulthanan et al. (2004)	Identification & footwear	3	6	No
Kennedy et al. (2005)	Identification	4	4	No*
Oberoi et al. (2006)	Anthropology/ Identification	4	15	No
Krishan (2008a)	Anthropology/ Identification	4	7	No
Krishan (2008c)	Anthropology/ Identification	4	8	No
Fawzy & Kamal (2010)	Anthropology	4	19	No
Atamturk (2010)	Anthropology/ identification	4	11	No
Moorthy et al. (2011)	Anthropology/ Identification	3	7	No
Vidya et al. (2011)	Anthropology/ Identification	4	9	No
Natarajamo-orthy et al. (2011)	Anthropology/ Identification	4	8	No
Kanchan et al. (2012)	Anthropology/ Identification	4	12	No

* Reliability is investigated but inappropriate statistical tests used.

Table 2.2 Summary of critical appraisal of literature pertaining to footprint measurement (foot indices)

Author(s) (Date)	Application	Strength of study overall (OLE grading 1-5)	Strength of study: validity (OCPM Score Max 24 pts)	Strength of study: reliability investigated? (yes/no)
Qamra et al. (1980)	Identification	4	3	No
Laskowski & Kyle (1988)	Identification	4	2	No
Maes et al. (2006)	Clinical	4	8	No

Table 2.3 Summary of critical appraisal of literature pertaining to footprint measurement (Arch Index)

Author(s) (Date)	Application	Strength of study overall (OLE grading 1- 5)	Strength of study: validity (OCPM Score Max 24 pts)	Strength of study: reliability investigated? (yes/no)
Cavanagh & Rodgers (1987)	Clinical	4	4	No*
Staheli et al. (1987)	Clinical	5	10	No
Hamill et al. (1989)	Clinical	4	1	No
Hawes et al. (1992)	Clinical	4	2	No*
Igbigbi & Msamati (2002)	Clinical	4	4	No*
Sacco et al. (2009)	Clinical	4	1	No
Xiong et al. (2010)	Footwear design	4	19	Yes. ICC 0.96

* Reliability is investigated but inappropriate statistical tests used

Table 2.4 Summary of critical appraisal of literature pertaining to footprint measurement (Footprint Angle)

Author(s) (Date)	Application	Strength of study overall (OLE grading 1-5)	Strength of study: validity (OCPM Score Max 24 pts)	Strength of study: reliability investigated? (yes/no)
Clarke (1933)	Clinical	4	9	No*
Forriol & Pascual (1990)	Clinical	4	10	No
Hawes et al. (1992)	Clinical	4	2	No*
Riddiford-Harland et al. (2000)	Clinical	3	15	No
Maes et al. (2006)	Clinical	4	8	No
Villarroya et al. (2008)	Clinical	3	17	No
Sacco et al. (2009)	Clinical	4	1	No

*Reliability is investigated but inappropriate statistical tests used

Table 2.5 Summary of critical appraisal of literature pertaining to footprint measurement (Chippaux-Smirak Index)

Author(s) (Date)	Application	Strength of study overall (OLE grading 1- 5)	Strength of study: validity (OCPM Score Max 24 pts)	Strength of study: reliability investigated? (yes/no)
Forriol & Pascual (1990)	Clinical	4	10	No
Hawes et al. (1992)	Clinical	4	2	No*
Maes et al. (2006)	Clinical	4	8	No
Villarroya et al. (2008)	Clinical	3	17	No
Sacco et al. (2009)	Clinical	4	1	No

*Reliability is investigated but inappropriate statistical tests used.

Tables 2.1 to 2.5 demonstrate overall poor scoring for strength of evidence of the articles appraised. The OLE system grades literature mainly for the purposes of establishing an EBP hierarchy of evidence. As previously discussed, the highest level (1) is assigned often to RCTs. Although examples of RCTs are to be found in articles relating to forensic science, the use of such a design may not be wholly appropriate for identification science. It is apparent that EBP does not rely solely on the OLE grading of literature which is dictated by the type of design used in the experiment. EBP involves many more components including the identification of the most effective methods, and could therefore be regarded not only as a method, but also as an ideology (Bloom et al., 2009). For example, NICE guidelines also incorporate cost-effectiveness within their EBP model. In the absence of peer-reviewed and published empirical testing to determine the probability of outcomes of a given population, often clinical or practitioner judgement is relied upon as the source of the best evidence (Greenhalgh, 2004; Cole, 2007). The above tables show articles

demonstrating an average score of 4 using the OLE system. Although initially considered a poor rating according to OLE, this remains the best evidence in this area to hand, and is therefore acceptable within the context of footprint measurement.

Reliability of the measurement methods was not sufficiently evident in the reviewed articles, except for the arch index utilised in the study carried out by Xiong et al. (2010). Validity scores using the OCPM system showed varied results, independent of a particular measurement approach.

2.5 Summary

Studies that examine the shapes of two-dimensional inked footprints show a lack of certainty regarding the various measurement approaches used, as supported by the low grading of the literature in terms of reliability and validity. For forensic measurement research to be of an acceptable standard for identification purposes, one would expect a reliability analysis with supporting error margins, perhaps in the form of confidence intervals. A study would also be expected to provide results of multiple statistical analyses in order to achieve a better perspective as to its validity.

A review of the literature pertaining to the evaluation of two-dimensional footprints did not bring forth a gold standard measurement approach. In the absence of a rigorously tested baseline method, it was apparent that the basic foundations supporting forensic footprint comparison in both research and in practice may be unsound. Coupled with these issues was the query over the use of multiple methods currently undertaken in the field. It has been difficult for the forensic science and the forensic podiatry communities to explicate a precise methodology in terms of the evaluative, analytical and comparative procedures involved in the footprint identification process. However, custom and practice may outweigh these difficulties. Evolved methods, based on the literature critiqued in this chapter, may be justified as admissible evidence in a court of law, providing the practitioner can demonstrate expertise in the area. This premise is supported by the decision in the Otway v R case (2011). Nonetheless, without a quantitative and objective rigorous tool to determine the nature of crime scene footprints, the development of a scientific basis from

which to justify the use of footprint evidence in criminal investigations is restricted. It seemed that the starting point to this inquiry was to establish a measurement approach deemed both valid and reliable for research purposes and also to answer the criticisms detailed in the UK and US law-driven recommendations and requirements. It was envisaged that the developed approach may also be applicable for uses in forensic practice.

2.5.1 Presentation of the thesis

A pragmatic method for measuring two-dimension bare footprints is proposed. Footprint measurement methods described in the literature will be considered, as these offer the best available evidence. Construct validity, content validity, criterion validity, reliability, accuracy, and consistency of the approach will be examined and established as part of this process. Since the study will extrapolate data from measurements, a quantitative evaluation will be used; the exception to this will be provided in the examination of external validity in which a qualitative method of inquiry will be employed to assess the utility of the measurement approach in the field.

The chapters of the thesis will reflect the outcomes of each analysis starting with the development of the measurement method, separate explorations of its validity and reliability and finally evaluation of the new approach in the form of external validation. Appraisal of literature relevant to each section will be discussed alongside methodologies, results, analyses and conclusions. A synthesis of the research elements will be presented and also a final discussion concerning the main conclusions from the study.

Chapter 3
The Development of a New Measurement Approach

3.1 Introduction

The decision to develop and evaluate a new approach is supported by policy, research and practice, in that without a practical tool underpinned by research (in this case measurement rigour), there is little chance of policy being implemented that is meaningful to practitioners.

The reason forensic individualisation science finds itself in this position is because it failed to establish a scientific basis to measurement which has been questioned and found wanting in respect of case law and underpinning scientific integrity. The drive to resolve these issues from within the forensic fraternity has resulted in policy stating that measurement rigour should be a cornerstone of tools used in practice.

In the previous chapter, which critically appraised the general literature, a stand-alone rigorous method for use as a baseline approach for further research was not revealed; however elements of the various published methods were considered by the researcher for the development of a new pragmatic approach to footprint impression measurement. In a bid to establish content validity, this chapter will explore these elements and rationalise why they were specifically selected by the researcher for the new measurement approach. It will also determine why other measurements and tools indicated by the literature were rejected, procuring the development of a new approach for measuring two-dimensional footprint impressions.

A small pilot study was undertaken in order to test the feasibility of the approach in terms of the collection of the footprints, the scanning process to produce digitised images and the use of the chosen software to record the proposed measurements. Right footprints were collected from a convenience sample of seven consenting adult subjects. These footprints were not included in the database used for the analyses presented in this thesis.

3.2 Footprint collection method

As discussed in section 2.1.1, identification from footprint impressions in a forensic context relies on conclusions made from the comparison of an exemplar and unknown print. Ideally, an exemplar print obtained from the donor should be made on the same surface and use the same printing medium as that of the unknown print. Whilst this would provide a closer model to the circumstances in which the crime scene print was originally conceived, these conditions may not be achievable due to ethical and practical constraints. For example, asking a donor to walk through blood incurs ethical considerations; therefore a similar medium to blood for the formation of a plantar print can be supplied, such as poster paint or ink. Section 2.4.1 considered these options and justified the use of an inkless paper system.

This method of print collection is quick, easy to use, clean, and relatively inexpensive and has been used in previous studies with no reported allergies or issues concerning cross-infection/hygiene (Bodziak, 2000; Kennedy et al., 2003; Kennedy et al., 2005).

In the pilot study, footprints were collected using an inkless paper system and the feasibility of this collection method examined.

Only two-dimensional footprints were required for measuring and therefore a hard surface was used and not a soft surface such as carpet which may have created a non-uniform three-dimensional print (Barker & Scheuer, 1998).

The right footprint only was taken from each participant in the pilot study. Sforza et al. (1998) suggest people demonstrate high symmetry between left and right feet. Landorf (2002) and Menz (2004) warn against using data from right and left feet from one person, as a high correlation will exist in whatever measurements are taken and essentially the same foot is being measured twice. Contrary to these findings regarding the dimensions of the foot, bilateral asymmetry has been observed in studies examining footprints (Oberoi et al., 2006; Krishan, 2007; Fawzy & Kamal, 2010). However, the use of two feet from one person may violate the assumption of independence of statistical testing, increasing the possibility of a type I error (Menz, 2004; Fascione et al., 2012). The thesis aims to develop a new approach to footprint measurement, providing a baseline for further research in this area. For this reason, measurement and analysis of both feet from each participant

was considered to be beyond the remit of this research. Therefore prints from the right foot were chosen for the purposes of the research.

Each sheet of inkless paper used in the footprint collection measured 297mm x 210mm (equivalent of A4 sized paper). This posed problems during the pilot study, regarding complete dynamic footprint capture especially amongst subjects with the longest of foot size. Although the paper could accommodate the length and width of the largest feet, an incomplete print sometimes resulted when the paper was not visually targeted by the subject. Whittle (2003) suggests that visual guidance, or 'targeting' the inkless sheet is 'likely to lead to an artificial gait pattern, as the subjects 'aim' for the platform' (page 149). However, studies investigating the effects of targeting on ground reaction force variability and the temporospatial parameters of gait have shown that there are no statistical differences between non-visually and visually guided steps (Sanderson et al., 1993; Wearing et al., 2000; Grabiner et al., 1995) . The use of a long roll of paper to capture dynamic footprints using alternative inking methods was considered to be potentially problematic, due to under- and over-inking issues, discussed in the previous chapter. The ideal solution to the dynamic footprint collection method targeting problem would be to have a long roll of inkless paper. However personal communications with the manufacturing company dealing with the inkless system failed to satisfy this demand. Anecdotal evidence had suggested a roll of fax (thermal imaging) paper used in conjunction with the inkless pad from the inkless system may prove successful in the capture of successive dynamic footprints from each person. A small-scale study was therefore instigated to examine this idea.

Five metre lengths of fax paper were secured into position on a hard-surfaced floor using masking tape. An inkless pad was placed at the start of the paper. Four subjects from the pilot study were asked to walk on the pad and continue walking along the roll of fax paper to the end. The footprints were then examined and compared with dynamic footprints from the same subjects captured using the inkless system. The footprints captured using the fax paper appeared to bleed beyond their perimeters. In this subjective comparison examination the resultant lengths and widths of the fax paper prints all appeared larger than the corresponding dimensions of the inkless paper footprints. After eight months of storage the fax paper footprints also appeared to fade compared with the inkless paper prints stored alongside. This phenomenon has been reported previously

(Yamamoto & Wiebe, 1989; Dulniak et al., 1996; Farrell et al., 2010). Footprints captured on fax paper were therefore discounted as a data collection method for the research presented in this thesis.

The seven volunteers from the pilot study were asked to walk up and down in their bare feet in the allocated area. This was to allow them to adjust and feel accustomed to walking unshod by the time their footprints were taken, permitting as near to a normal bare foot walking style (Mathieson et al., 1999). Literature evolving from gait analysis studies, suggests various protocols for optimum data retrieval, depending on the design of the study and the pathology of the subjects involved (Whittle, 2003). These studies often set out to capture data at a point of walking that is as 'natural' for that subject as possible, often collected midgait (Whittle, 2003). In a crime scene scenario, an unidentified footprint is devoid of information regarding the donor's walking style or the type of activity the person was undergoing at the time the mark was created by the foot. Therefore producing a comparison footprint to imitate a 'natural' walking pattern for that person is not a major consideration. Indeed the process of footprint identification recommends the collection of multiple exemplar footprints in different states, for example, twisting, walking and standing (DiMaggio & Vernon, 2011). However, as previously discussed, the thesis presents a baseline footprint measurement approach and as such, footprint collection in a consistent manner is the main consideration for satisfactory data collection and analysis.

The collection of footprints from all subjects using a strict walking speed was considered for consistency. This protocol involves calculating the stride length and cadence of each subject (Whittle, 2003). Adherence to a specified walking speed is important in the study of gait cycles but its value is unknown in research involving inked footprints. Bosch et al. (2009) noted with their subjects that a less than natural gait pattern was incurred when asked to walk at a specified cadence. Arif et al. (2004) observed walking speeds differed between subjects placed in groups according to age, with subjects displaying walking instability when asked to follow a sound signal (metronome). Instability during walking may result in a non-typical footprint from that subject, as the foot may move unnaturally in a bid to compensate during footprint capture. Therefore, the collection of footprints using a prescribed cadence was not included in the method for the PhD as it was important to capture a natural baseline. Other

footprint collection protocols were considered and adopted from previous literature that discusses both electronic and inked footprints.
Each participant was asked to stand comfortably at a designated starting point at one end of a five metre walkway, their eyes fixed on a marker placed at eye-level on a wall ahead of them. They were then required to walk normally at their own pace along the walkway, starting on their right foot. The inkless pad was positioned at the side of the walkway adjacent to the position where the right foot tended to land on second contact with the ground (third step). The inkless paper was placed at the side of the walkway where the foot was landing on third contact with the ground, (the fifth step), in accordance with the midgait protocol (Morlock & Mittlmeiser (1992). The participant was asked to repeat the process and the operator adjusted the inkless paper and mat accordingly until the positions of foot strike were confirmed (DiMaggio & Vernon, 2011; Reel et al, 2012).

The mat and inkless paper were then secured in place with masking tape in the previously confirmed positions (DiMaggio & Vernon, 2011; Reel et al, 2012). For the final walk, the participant was advised to look on the floor ahead of them in a bid to capture the whole footprint within the small area of inkless paper. All dynamic footprints were thus captured using this five-step protocol (midgait protocol) in accordance with collection details outlined in the studies by Morlock & Mittlmeiser (1992) and Wearing et al. (1999).

The midgait protocol was chosen over the two-step method recommended by Meyers-Rice et al. (1994), as the midgait method (four steps or more) has been suggested to be more reflective of a person's natural walking style and cadence than the shorter gait protocols (Morlock & Mittlmeiser, 1992; Wearing et al., 1999). Nicholson et al. (1998) noted that by using shorter gait protocols, the number of spoiled trials was reduced by over two-thirds. However, Wearing et al. (1999) observed statistically significantly reduced rearfoot pressures for the two-step initiation protocol. Conversely, these authors also noted significantly reduced forefoot pressures for the two-step termination protocol, and concluded that neither method can be interchanged with the midgait protocol. Literature that examines the associations of variable foot pressures with footprint dimension could not be found. However, in order for the research to incur as few unknown variables as possible, a gait protocol that captures a consistent representation of the stance phase of gait on repeated occasions, is favoured. Thus the midgait protocol was the preferred

method for footprint capture. The five metre walkway allows enough room for the footprint collection to be carried out in this described manner.

Literature that discusses the collection of static footprints for analysis is generally poorly described in terms of protocol. For example, Krishan (2008a) describes how each subject in his study was asked to 'step on to white plain paper' after their feet had been inked (page 94). This is echoed by other authors including Qamra (1980), Robbins (1986) and Fawzy & Kamal (2011). As consistency of the method for footprint collection was the prime concern for data collection purposes in this research, it was most important that the static footprint capture process was repeatable. The following describes the protocol that was followed for the purposes of the pilot study.

Each subject was asked to stand comfortably, hands on hips, with their feet on either side of the inkless pad. They were requested to raise their right foot, place it onto the pad and then place it back to its original position whereby the inkless paper had now been put into position by the operator.

Three footprints in each state, static and dynamic, were initially collected to assess practice or learning effects, in which a better, usually a more positive result is produced for every repeated test (Robson, 2002). Conversely negative results occur when subjects become worse at performing the task, known as fatigue effects (Hicks, 2004). In the collection of footprints, the first print may be unrepresentative due to a misunderstanding of the correct protocol, for example. After collection of a series of footprints from the same person, the participant may be feeling fatigued and produce a smudged print due to foot drag. These effects are examples of unwanted systematic error or bias that occur in an experiment (Portney & Watkins, 2009). Past researchers have argued that three is an appropriate number of times to repeat a procedure in order to expose the effects of learning, and also fatigue effects (e.g. Salthouse & Tucker-Drob, 2008; Lamparter et al. 2011). No literature could be found by the researcher proposing an appropriate number of trials for footprint analysis to counteract practice and learning effects. Therefore, in order to determine the presence of practice effects, each set of three prints from the subjects recruited for the pilot study, were measured and results compared. Descriptive data suggested little variation occurred between prints one to three from the pilot study, for example Calc_A1

(static) displayed a mean value of 237.94mm across three prints for all subjects with a standard deviation (SD) of 0.21. Small standard deviations were noted for other measurements also: Calc_A1 (dynamic) mean 255.29mm, SD 0.56; MPJWidth (static) mean 93.36mm, SD 0.26; MPJWidth (dynamic) mean 93.14mm, SD 0.11. The small differences between measurements observed in the raw data did not appear to be directional. If a significant variation between a subject's measurements does not occur between three trials this would indicate that that the collection method is not altered by learning or fatigue effects, negating the collection of more than one print to counter the effects of systematic variation (McCaffrey et al., 2000). However, in an analysis of data using a larger sample (presented later in the thesis), it was intended that reliability estimates be examined using intraclass correlation coefficients (ICC). Baumgartner argues for the use of three trials in preference to one as the extent of reliability is better reflected in the larger number of trials (Baumgartner 1989). This is supported by the work of Gauch (2006) and Bruton et al., (2000) who suggest three or more trials should be performed to ensure useful ICC results. Therefore for the main body of the research, three static and dynamic footprints from each subject of the larger sample were collected.

3.3 Measurement of scanned images

The resulting footprints obtained from the pilot study were scanned using an Epson scanner (DX4850) set at 150 dots per inch. This was the default setting on this commercially available scanner typical of those on sale at the time, chosen as it remained within the pragmatic parameters of the research aims.

Footprint measurements were automatically recorded by the GNU Image Manipulation Program software (Version 2.6.8) on a Windows XP PC and then entered into a database for further analysis using Statistical Package for the Social Sciences software (SPSS) (Version 17.0 SPSS Inc. Chicago IL). The specific details regarding these measurements will be discussed in section 3.4.

The measured scanned images were stored in JPEG (Joint Photographic Expert Group) format on the PC as this was considered a more pragmatic approach compared with storing in tagged image file format or RAW (uncompressed) formats, which take up considerably more disk space. Tagged image file format

and RAW images are termed 'lossless' because their images can be stored and reconstituted without a loss of digital code. However, tagged image file format and RAW create large file sizes and are slow to transfer. JPEG is lossy (in which redundant information is permanently eliminated) and compressed to save storage space, therefore decreasing transfer time. Riviello proposes this compression of the JPEG has the detrimental effect of degrading or altering the image due to the loss of digital code (Riviello, 2008). Riviello suggests that compression occurs every time an image of a photograph is saved and subsequently closed, cumulatively altering the dimensions of the image on each successive occasion. Therefore the following small-scale study was initiated to assess this process in the context of the research.

A dynamic footprint from the collection was picked at random and an American Board of Forensic Odontology No. 2 Photomacrographic Scale (TBS0121A) displaying one millimetre graduations was placed alongside the print in the scanner, as recommended by Hyzer & Krauss (1988). A length line from the base of the heel to the tip of the first toe print (Calc_A1) and width lines across the heel print (CalcWidth) and forefoot print (MPJWidth) were constructed and measured using GNU Image Manipulation (GIMP) software. Also measured were the calibrations of the American Board of Forensic Odontology scale to check that the millimetre markings were comparable with the measurements recorded by the software. The results of the three measurements on the footprint were recorded. The image was then saved and closed. The same image was subsequently opened, the three lines measured and recorded, the image saved and then closed. The image was opened, measured, saved and closed in this way for a total of ten times and the ten sets of recorded measurements compared for differences. The results of this small test confirmed no differences between the measurements and compressing the image appeared not to be detrimental as previously postulated by Riviello (2008).

The JPEG images used for the research were opened only once for construction and recording of linear and angle measurements, then saved and closed. None of the stored images were re-opened once the measurement data had been recorded. The small-scale study in which the JPEG images were opened, saved and closed ten times is therefore an overestimation of the

process involved for this study. However, multiple openings and savings of each image more than ten times in future research is not recommended.

3.4 Justification of measurement choice

Utilisation of the Ohio College of Podiatric Medicine's literature grading system for validity, uncovered varied score results ranging from 1/24 to 19/24. The only article to have established reliability of the measurement method applied, was the footwear design study by Xiong et al., (2010) This also received one of the highest validity scores of 19/24. The study used various measures to classify foot shape in terms of the arch including an approach involving inked footprints called the Arch Index (Forriol & Pascual, 1990). The article by Villarroya et al., (2008) also scored highly using the OCPM system (17/24). Their study examined foot arch shape and compared the degree of the pes planus condition in normal-weight and obese individuals, using inked footprints and weight-bearing radiographs. Measurements of the inked footprints involved the Chippaux-Smirak Index (Cavanagh & Rodgers, 1987). Since these measurement approaches are prominent in the literature and achieved high appraisal scores, they were considered in the development of the new approach.

The Chippaux-Smirak Index calculates a ratio of the width measurement at the narrowest part of the arch divided by the widest part of the forefoot print at the MPJ area. The Arch Index (AI) is the ratio of the area of the middle third of the footprint to the total footprint area, minus the toe prints. Both these indices are based on the premise that the height of the arch is related to the footprint, although it has been noted that the AI can explain approximately only 50% of the variance in arch height (Chu et al., 1995; McCrory et al., 1997). Wearing et al. (2004) also refute the AI as a measure of navicular height in their paper titled 'The arch index: a measure of flat or fat feet?' It is suggested that the calculation of the AI is dependent on soft tissue variations of the arch and is therefore a longitudinally inconsistent measure, making it an unsuitable measurement for footprint identification or research purposes.

Calculations of Chippaux-Smirak Index (Cavanagh & Rodgers, 1987; Villarroya et al., 2008), AI (Forriol & Pascual, 1990; Xiong et al., 2010) and foot indices

(Qamra et al., 1980; Laskowski & Kyle, 1988; Maes et al., 2006) produce ratios and percentages to describe footprint dimensions. These types of estimations have the effect of generalising the data. Qamra et al. (1980) explain that this method is utilised in order 'to overcome the faults of registration, recording and observation' (page 146) and does not allow for the examination of actual footprint dimensions. Therefore, these methods were not considered further for the purposes of the research presented in this thesis.

Additionally, literature describing the methodology for these methods considers measurement from static prints only. It is unlikely that crime scene footprints are those of a static nature only and the new approach incorporates the evaluation of measurements from both static and dynamic footprints.

Studies that have compared foot shapes captured in both static and dynamic states have noted that the measurements for the latter are larger than the former (Kippen, 1993; Barker & Scheuer, 1998; Mathieson et al., 1999; Tsung et al., 2003). For this reason, it was decided that footprint measurements capable of disseminating length and width information should be employed to enable an analysis of discriminant validity. Measurement approaches used in footprint evaluation, critically appraised in the previous chapter, were all considered as possible tools capable of disseminating this information. This is despite low scoring of these articles in terms of measurement validity and reliability. In the absence of a gold standard footprint measurement method, the existing literature (the best available evidence) and practitioner expertise in this field were evaluated in order to identify a suitable approach for analysis. This involved a process of abductive reasoning, employing logical inference and supported by anecdotal evidence from practitioners. For example, the selection of random points at the base of the heel from which to draw the length measurements appeared to produce varying results in the pilot study, noted also by Kennedy et al. (2003) who reported that 'numerous instances were found where the precise pixel to choose for the heel point of the print was ambiguous', adding that 'each alternative [pixel] led to different measurements' (page 57). Therefore alignment of the scanned footprints was deemed preferable, prior to the determination of this designated point of the heel print, to allow for consistency (Reel et al, 2010). Using the chosen software discussed further in section 3.5, the inner and outer tangents of the footprint were identified (Rossi, 1992; Kennedy et al., 2003; Kennedy et al., 2005) and bisected to

create the central axis as described by Kennedy et al., (2005). A grid was placed over the image, which was subsequently rotated so that the central axis was vertically aligned. A horizontal mark was then introduced to traverse the most proximal pixel of the heel in this new alignment (Reel et al, 2010). From the point where the central axis and the heel line intersected, a series of five lines were drawn to the apices of each toe, as suggested by Gunn (1991). Also included were lines drawn using the software to highlight the widest parts of the heel and the ball of the footprint, as indicated by the outer and inner tangents (Reel et al, 2010). These width measurements have been described in previous footprint studies (Robbins, 1985; Gunn, 1991; Bodziak, 2000; Kennedy et al., 2003; Kennedy et al., 2005).

Although there exists literature pertaining to the anatomy and function of the toes (Mann and Hagy, 1979; Hughes et al., 1990; Endo et al., 2002; Menz et al., 2006), articles exploring variations in individual toe print distribution could not be found. Therefore various angles were constructed over the toe area for further analysis. In the pilot study, it was noted that for two subjects the fifth toe would fail to make contact with the ground. For this reason, toe angles incorporating and excluding the fifth toe were included for analysis. The footprint angle was included because some studies that have employed this method of arch estimation have shown to score highly for validity using the OCPM system described in the previous chapter.

A summary of all chosen width, length and angles drawn on the footprint images from the pilot study are summarised in Figures 3.1 and 3.2 below.

Linear measurements as depicted in Figure 3.1 are referred to using the following abbreviations:

Calc_A1 Base of heel print to apex of first toe print.

Calc_A2 Base of heel print to apex of second toe print.

Calc_A3 Base of heel print to apex of third toe print.

Calc_A4 Base of heel print to apex of fourth toe print.

Calc_A5 Base of heel print to apex of fifth toe print.

MPJWidth Widest part of the forefoot print

CalcWidth Widest part of the heel print

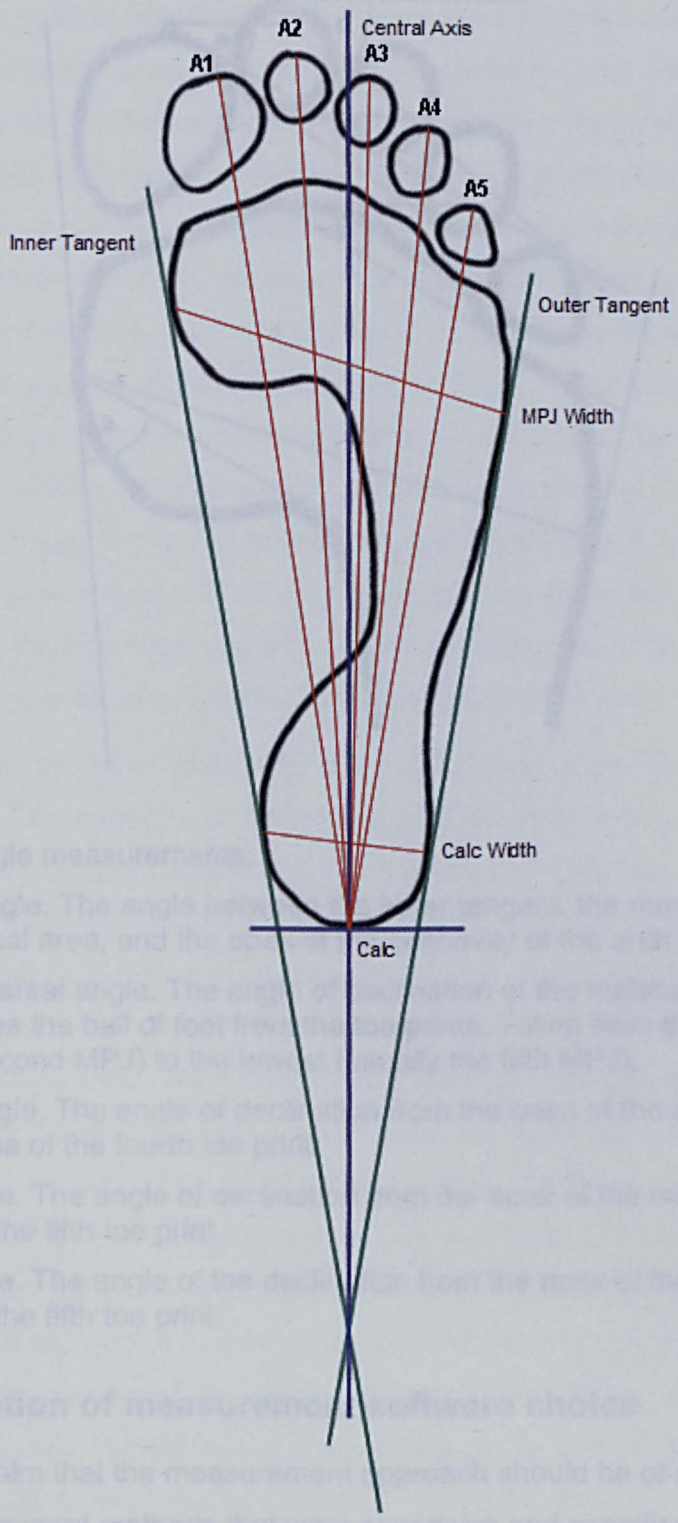


Figure 3.1 Length and width measurements

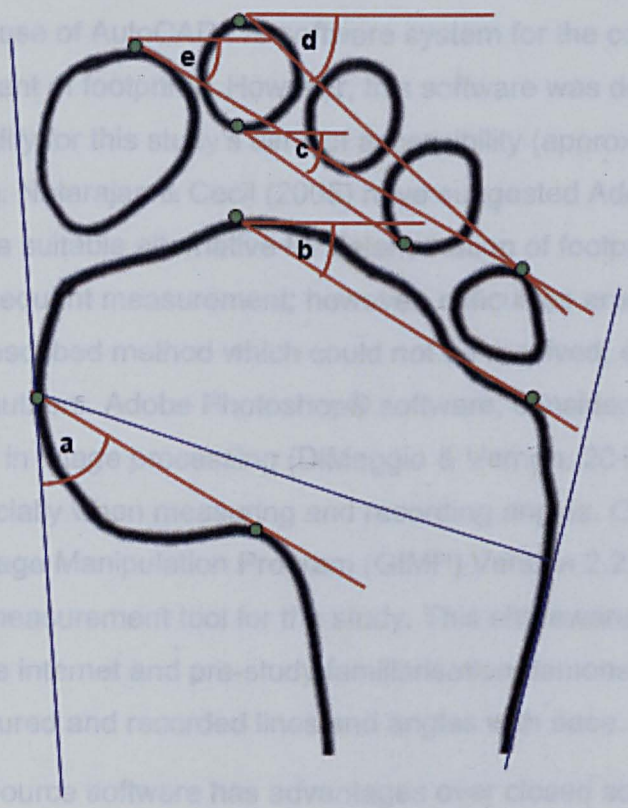


Figure 3.2 Angle measurements:

a) Footprint angle. The angle between the inner tangent, the most medial point of the metatarsal area, and the apex of the concavity of the arch of the footprint.

b) Distal metatarsal angle. The angle of declination of the metatarsal ridge which separates the ball of foot from the toe prints. Taken from the highest ridge (usually the second MPJ) to the lowest (usually the fifth MPJ).

c) 2-4 base angle. The angle of declination from the base of the second toe print to the base of the fourth toe print.

d) 2-5 toe angle. The angle of declination from the apex of the second toe print to the apex of the fifth toe print.

e) 1-5 toe angle. The angle of toe declination from the apex of the first toe print to the apex of the fifth toe print.

3.5 Justification of measurement software choice

In fulfilling the aim that the measurement approach should be of a pragmatic nature, measurement methods that were expensive and complicated to complete were not considered. Automatic measurement of the outlines of scanned footprint images was favoured over manual methods in order to reduce random and systematic error. Various measuring software packages were sampled, as recommended by the pertinent literature. For example, the Optical

Center Method (Bodziak, 2000; Kennedy et al., 2003; Kennedy et al., 2005) incorporates the use of AutoCAD r13 software system for the construction and linear measurement of footprints. However, this software was deemed too expensive to qualify for this study's aims of accessibility (approximately £3000 at time of writing). Natarajan & Cecil (2005) have suggested Adobe PageMaker® as a suitable alternative for determination of footprint optical centres and subsequent measurement; however, difficulties arose when following their prescribed method which could not be resolved, even by direct contact with the authors. Adobe Photoshop® software, considered the forensic industry standard in image processing (DiMaggio & Vernon, 2011), also proved problematic especially when measuring and recording angles. Open source software GNU Image Manipulation Program (GIMP) Version 2.2.17 was finally singled out as a measurement tool for the study. This shareware was free to download from the internet and pre-study familiarisation demonstrated that the programme measured and recorded lines and angles with ease.

The use of open source software has advantages over closed source software such as Adobe Photoshop®, in that by sharing its source code, rapid advancement and innovation of the product ensues, leading to greater stability and richer functionality. Developers cooperate under a continuously peer-reviewed model, resulting in parallel debugging (Perens, 1999; Koch and Schneider, 2000). The GNU general public licence allows the freedom to use, copy and distribute software and is also multi-platform. The cost-effectiveness, stability, and the user-friendly aspects of GIMP result in pedagogic, economic and technical advantages over other measurement packages. In an article by Brian Carrier, the Daubert criteria is employed to critically appraise differences between open and closed source software (Carrier, 2003). He concludes that open source tools meet the guideline requirements for legal admissibility more comprehensively than closed source tools arguing that having access to the tool's source code facilitates improvement of the quality of testing and establishment of error rates. Peer review is a constant requirement of open source software and the large numbers of developers and users involved with the products establishes general acceptance in the community (Koch & Schneider, 2000).

An account of the steps taken during forensic image analysis including measurement is an essential requirement if used as evidence in a court of law, or for method replication for research purposes. A disadvantage of GIMP software is its inability to record an audit trail, or history log during use, as opposed to Adobe PhotoShop® which has this capability. This shortcoming has been examined by Chen et al. (2011) who proposed an algorithm for revision control for binary files used in software such as GIMP. They developed a prototype system built upon GIMP. Four computer software postgraduate students and three professional illustrators competent in using Adobe PhotoShop® were recruited to evaluate the GIMP prototype system developed by the authors. The results of the initial qualitative inquiry led to various amendments to the system. The final evaluation concluded that the revision control forming the audit trail was pragmatic and useful. The authors have subsequently released the source code in the public domain; however this is yet to be officially released for use with GIMP at the time of writing. US guidelines relevant to the use of digital image processing in forensic applications require that 'documentation of quantitative image analysis steps should be in sufficient detail to enable another comparably trained person to repeat the steps and produce the same conclusions' (SWGIT, 2001, page 14). European guidelines dictate similar guidelines requiring that when using commercially available software tools for forensic purposes, all steps taken should be documented 'in order to produce a process that could be repeated, if necessary, by someone else and give the same result as the original' (ENFSI, 2006, page 14). A manual created by the researcher with supporting CD demonstrating moving screen shots of each step taken in the construction and recording of measurements on the footprint images, may stand up to these requirements. This manual will be discussed further in the next section.

3.6 Development of the manual

A manual with supporting DVD and CD was created as a practical guide to collecting and measuring footprints using the method developed by the research. The main body of the manual was comprised of annotated screen shots demonstrating the method of line construction and measurement using GNU Image Manipulation Program Version 2.2.17. It also included health and

safety guidance (e.g. maximum length of time seated before a computer, advice regarding posture, etc.), a section on measurement concepts including reliability and validity, and a brief summary of a reliability study carried out at York St John University (to be discussed further in Chapter 7) with a simplified table of results for comparison. The DVD demonstrated visually how to collect the footprints using an inkless paper system. Suggested protocols in terms of obtaining dynamic and static prints were explained in the DVD and also discussed in the manual. The included CD depicted moving real-time screen shots of how the measurement software should be manipulated.

The researcher incorporated a variety of learning approach options within the package for the reader to assimilate. For example, the DVD and CD allowed the new user to be taught through visual and auditory learning mechanisms, whilst the manual aided the reading learner and guided the kinaesthetic learner as he/she used the software concurrently with the CD.

It has long been accepted that there exists a variety of learning styles, involving educating methods appropriate to each individual (Kolb, 1984; Honey & Mumford, 1992; Sternberg, 1997). A popular categorisation of the different ways of learning is Fleming's VARK model (visual, auditory, reading/writing, and kinaesthetic) based on neuro-linguistic programming (Fleming & Mills, 1992). However a systematic review focussing on thirteen of seventy-one separate learning-style models has suggested that constitutionally based approaches including the VARK model have not been rigorously tested for reliability and validity. More flexible learning models are instead recommended, for example those which concentrate more on personal factors such as cooperation and motivation (Coffield et al., 2004) The authors of this detailed systematic review cite Hermann (1996) and Allinson & Haynes (1996) as introducing the most credible learning models in terms of evaluations of internal consistency, test-retest reliability, predictive validity and concurrent validity. These two models incorporate *brain learning* and emphasis is placed on the cognitive style of the learner. Other models are dependent on assessing personality traits, though empirical evidence as to the effectiveness of these approaches is limited (Allinson & Hayes, 1996). Despite conflicting theories as to the best method of learning styles and pedagogy, it is agreed that students' learning styles are widely varied (Allinson & Hayes, 1996). It was considered that inclusion of

statistical evidence to support the theory detailed in the manual would be helpful to Herrmann's 'upper left brain' students who learn by applying logic and analysis, and like to acquire and quantify facts as part of that process. These types of students according to Herrmann (1996), respond to data-based content. In contrast, 'upper right brain' students learn by self-discovery and prefer to synthesise content; these learners respond to free-flow and spontaneity, and it was hoped that the structure of the package allowed these types of learners to 'dip in and out' should they so wish. Additionally, an inclusion of a description and summary of results of the measurement method's reliability tests provided the user with feasible error margins.

The newly developed manual, CD and DVD detailing the footprint measurement approach, was now dependent on external evaluation to ascertain validity. Multiple users of the approach following the steps provided by the guide could establish a degree of reproducibility if the same measurements using the same footprint images were collated and analysed. The manual together with supporting CD and DVD is utilised further in two remaining chapters of this thesis dealing with aspects of evaluation and reliability of the measurement approach.

3.7 Conclusions

This chapter has explained how the measurement approach was developed from the available literature, in turn establishing a degree of content validity. The feasibility of the approach was explored in a small pilot study and a manual produced as a practical guide. The following chapters will examine the extent of reliability and validity of the approach, using a larger sample for appropriate data analysis, in a bid to prove the concept.

Chapter 4
Establishing Evidence of Convergent and Discriminant Validity

4.1 Introduction

In the previous chapter, the extent of content validity of the new measurement approach was sought. This explained the process by which the method was developed from both the existing literature and methods used in practice, a definition of what was to be measured by the tool, how that definition was derived, and whether important measurements were omitted or irrelevant ones included.

In this chapter, efforts are made to establish construct validity in the forms of convergent and discriminant validity. As previously discussed in Chapter 2, construct validity is established if the theoretical concepts, or constructs, are accurately portrayed by the operational concepts (Stevens, 1993). Determination of this type of validity is important to understand the degree to which inferences can be made from the operational procedures to the concepts on which the procedures were based. For example, intuitively a difference between the lengths of static and dynamic footprints from the same person could be expected. This is supported by the literature (Barker & Scheuer, 1998). Thus, linear measurements which take into account the length of both static and dynamic footprints were developed and included in the new approach for assessment of this particular theoretical construct.

The construct-related validity examined in this chapter can be sub-divided into two types; convergent and discriminant validity. Convergent validity employs correlational procedures as a methodology to explore the extent of construct validity. According to Wood (1989), moderate to high associations ($r \geq 0.51$) measured utilising Pearson product-moment (PPM) correlation coefficients between the same or similar constructs would be an indication of evidence of convergent validity. Complimentary to this, the extent of discriminant validity is realised when differences are established where in theory, the concepts being measured would show those differences. Differences in theoretical concepts as

suggested by a prior appraisal of the relevant literature can be determined statistically by dependent t-testing (Field, 2005).

Bryman & Cramer argue that it is more important to establish evidence of convergent validity rather than discriminant validity (Bryman & Cramer, 2005) although others argue it is best practice to include both types of validity in an analysis of its construct variables (Campbell & Fiske, 1959; Douglass, 1979; Wood, 1989). This chapter will examine associations and differences between length, width and angle measurement data from three hundred and sixty six footprints in a bid to determine the extent of both convergent and discriminant validity.

By exploring convergent and discriminant validity, it was hoped that further understanding could be gained of the footprint's potential to yield information within an identification context. Such information could possibly include the effects of motion, sex, height, weight, body mass index, age and ethnicity upon footprint shape. The following section (4.2) will examine evidence of previously published interpretations of these effects upon the dimensions of the human footprint.

4.2 Literature Review

Literature regarding the associations and differences between and within footprints was sought in order to inform the method designs and analyses for the examination of validity for this part of the research. As well as an exploration of inter-relationships between the width, length and angle measurements, the effects of motion, weight, sex, height, body mass index, age and ethnicity upon subjects' footprint measurements was also considered. Predictions as to which measures show similar traits and those that illustrate differences under differing circumstances can only be attempted after a search of the relevant literature as part of a content validity analysis. Databases including MEDLINE, AMED, ZETOC, Science Direct, CINAHL, PsycARTICLES and SPORTDiscus were selected and the following criteria were included in the search with no limitations. For example, all languages were selected and the search was not prohibited by type of publication or date. Terms that were excluded from the general search were 'carbon', 'gene*', 'child*' and 'hormon*'.

Specific searching then ensued to gain a list of articles pertaining to each particular theme. For example, in order to identify papers examining the relationships between footprint dimensions and height, the following terms were used; 'footprint* AND human AND adult AND height OR stature'. Serendipitous searching and grey literature proffered additional articles. Literature pertaining to electronic footprints, for example pedobarometric systems, was excluded as there is evidence to suggest that inked footprints and electronic footprints are different (Urry & Wearing, 2001), discussed in Chapter 2. A search strategy table adapted from British Medical Association Library Seeking Evidence (MEDLINE course materials, 2005) is displayed in Appendix C.1. Differences in footprint dimensions within subjects are most prevalent in the literature when comparing static and dynamic footprints. Relationships between individuals' footprint dimensions are examined when correlated with variables such as age, weight, sex, height and ethnicity. The following appraisal of the literature will examine these variables in turn, commencing with differences between static and dynamic footprints.

4.2.1 Motion

The article 'Predictive value of human footprints in a forensic context' (Barker & Scheuer, 1998) investigates the differences between actual foot dimensions, static inked footprint dimensions and dynamic inked footprints dimensions. The aim of the study was to analyse the probability of determining a person's sex, shoe size and stature from a partial or whole footprint. Inked footprints were collected from fifty six female and forty nine male subjects from a UK population and the prints measured using a method similar to that employed by Robbins (1985; 1986).

Construction lines and measurement parameters from the footprints were tested for reliability. Results from their intra- and inter-rater assessments using data from the heterogeneous sample from three different observers suggested reliability fell within 'acceptable' limits (page 341), although a definition of acceptability is not offered. They report SDs for intra-rater tests of the footprint parameters ranging from 0.224mm – 0.447mm and a standard error (SE) from 0.100mm to 0.200mm. Inter-rater SDs are reported to have ranged from 0.724 – 1.934 with SEs from 0.296mm – 0.790mm. Here the use of the SE to determine the extent of intra- and inter-rater reliability has been estimated by calculating

the SD of the sample means. This calculation indicates how the sample represents the population. A resultant large SE (relative to the sample mean) would imply great variability between the means of different samples. In this situation the sample under examination might not be representative of the population. A small SE would indicate that most of the sample means display similarity to the population mean thereby inferring that the sample is likely to accurately reflect the population. The SEM however, examines multiple measurement results, such as a repeated measures design, from one person, rather than the overall means of a set of scores from a group of people. The SD of these measurement results can be calculated and thus the error (variation) of the measurement (Brown, 1999). In this particular study it would have been more appropriate to have analysed the data for strength of reliability in terms of the SEM rather than the SE because the former estimation allows for reliability interpretations of the measurement approach for each person's footprint dimensions, rather than the group of subjects' footprints as a whole.

Further results regarding Barker & Scheuer's experiment suggested that the walking length of a footprint (dynamic) exceeded the stick length of the foot (recorded using a foot-measuring last on the actual foot rather than the print) which in turn exceeded the static footprint. Highest correlations were illustrated between the actual foot length and static footprints from the right foot ($r = 0.98$). Actual forefoot width correlations with static and dynamic prints were also good (0.68 to 0.73); the highest correlations here were shown with right dynamic footprints. Measurements of actual heel width with static and dynamic heel width print measurements determined moderate correlations (0.55 to 0.66). Using the OLE scoring for an overall grading of the quality of the literature, this paper is of level 4 standard and scores 4 when rated using the OCPM system.

The podiatrist Kippen investigated the differences between static and dynamic inked footprints employing his own method for measuring a series of prints from one male subject (Kippen, 1993). The measurement method of the inked footprints involved the linear distance from a point at the base of the heel to the apex of the third toe (representing the length of the foot), from the base of the heel to the centres of the second, third and fourth toe apices and finally an angle measurement from the base of the heel to the second and fourth toe apices to record differences in foot breadth.

Static and dynamic footprints were examined for differences using a student t-test for both the foot length and foot breadth. The author concluded that there was a significant difference between the static and dynamic lengths of the footprints ($t = 6.4828$, $df = 7$, $p < 0.001$) and between static and dynamic footprint breadths ($t = 1.2632$, $df = 12$, $p = 0.023$). This part of the study has poor validity as the static inked print dimensions used for comparison with the dynamic prints were taken using a pedograph. The use of two different mediums to collect static and dynamic footprints is a source of error and the use of only one subject limits any inference to the wider population. It can be described as a preliminary study only and is of level 5 quality according to OLE, and scores 4 points using the OCPM system.

Hamill et al. (1989) determined Arch Index values displayed statistically significant differences between static and dynamic inked footprints in a dependent t-test comparison ($p < 0.05$), in their assessment of the relationship between certain static and dynamic lower extremity measures, although the t-value is not stated. The dynamic and static arch index values demonstrated high correlation coefficients ($r = 0.95$). The authors suggest differences occur in the AI values between static and dynamic states because of the influences of differing orientations of the lower leg and increased weight during walking. They also theorise that increased muscle activity in the medial longitudinal arch whilst walking would result in smaller arch index values. The twenty four subjects used for this study were split into three groups; normal, flat or high arched, the allocations based on prior inked footprint evaluation. It is not clear how many subjects were assigned to each homogenous grouping, but it is possible that further statistical investigations, such as those used for estimating reliability, were not undertaken due to inadequate sample numbers. The study is of a survey design, and therefore merits Level 4 scoring using the OLE grading. Due to the poor design of the study it scores 1 using the OCPM system for grading validity.

4.2.2 Sex

Literature pertaining to the determination of the sex of a person from their footprint is sparse. Studies that examine male and female actual foot shape and those using electronic footprints, mainly in the field of footwear design, suggest

there are differences in foot shape between the sexes (Wunderlich & Cavanagh, 2001; Sen & Ghosh, 2008; Luo et al, 2009). This assumption of male/female differences in foot size was the basis of Oberoi et al.'s development of 'the standard footprint length' (page 4) in which a value is calculated from the mean of one hundred male and one hundred female footprint lengths (Oberoi et al., 2006). Footprint length was defined as being from the base of the heel to the longest toe print (either the first or second). Footprints falling on the smaller side of the standard print length were classified as belonging to female subjects and those on the larger side to males. The application of sensitivity and specificity tests demonstrated that sex prediction using this method was 80% accurate. The remaining 20% of values occurred around the mean value, indicating measurement lengths in which male and female footprints overlapped. These results infer that anthropometric differences can be attributed to the prediction of sex, rather than sex differences. On average males are taller and heavier than females; factors which have an effect on footprint length. The study offers a simple method of sex determination from footprints using raw data with no inferential statistics or confidence intervals reported to support the conclusions. It is of OLE level 4 and merits a score of 5 out of a maximum score of 24 using the OCPM system to rate the validity of the study.

Igbigbi & Msamati (2002) found there were non-significant differences between male and female footprints when arch heights were analysed using the arch index. The Malawian sample used for the study included a mixture of ages ranging from thirteen years to seventeen years. Analysis of footprints from this particular sample included the growing feet of children through to fully-developed adult feet and the design of the study did not adjust for this heterogeneity. PPM correlation coefficients were employed to establish reliability, which reflect the extent of linear relationships between the footprint parameters, and not the extent of reliability. In other words, it is a measurement of association and not of agreement (Bland & Altman, 1986; Rankin & Stokes, 1998; Baumgartner, 1989). Therefore it is possible to obtain a strong degree of correlation when in fact agreement is weak between two variables. Bland & Altman were so irked by the incorrect use of PPM correlation coefficient values they suggested journals should recall all articles using this statistic for

assessing reliability and reanalysed using more appropriate statistics (Bland & Altman, 1986). The study is an orthopaedic case report of a group of subjects and is therefore of OLE level 4. It scores 4 using the OCPM grading system for validity.

Atamturk (2010) collected the footprints from five hundred and six males and females from a Turkish sample ranging from seventeen to eighty two years of age. The method used for footprint collection was not supported by any previous literature. It involved asking the subject to wet the surface of the foot by stepping into a tray of water and then onto a piece of paper. The edges of the footprints were identified and highlighted using a pen 'before the papers dried' (page 22). This method may be subject to potential error; not only by the manual measurement technique but also the use of watery footprints on paper could introduce a wicking effect altering the actual dimensions of the footprint. Evaporation of the prints may suggest shrinkage of the dimensions also. Although one effect may cancel the other out, there are too many unknown variables that exist by using this method, the effects of which are not fully understood. The longest footprint length and the widths of the forefoot and the heel were drawn onto each footprint and measured. The resultant data were analysed for accuracy in the prediction of sex with the conclusion that footprint widths and lengths are not useful for this purpose and that shoe measurements and other body parts are more successful for this determination. Grading the paper for the purposes of presenting a hierarchy of evidence in terms of the literature, an OLE score of 4 is given. Despite the concerns regarding the method of footprint collection, the comprehensive statistical analyses and description of sample choice warrant an OCPM score of 14 out of 24.

The article by Kanchan et al. (2012) is the only study to suggest that statistically significant differences exist between male and female footprints. Their sample consisting of fifty male and fifty female Indian subjects with ages ranging from twenty to twenty five years had their inked footprints taken and manually measured using a method adapted from the Gunn method (Gunn, 1991). Data from recorded length measurements were analysed as separate male and female groups and the differences between groups calculated using 'z-values' (table 5, page 3). This statistical test compares each sample mean (for example

the male group) with the whole population (males and females combined) and is perhaps a blunter tool than using t-tests which compare the results of the two independent samples (Field, 2005). They found statistically significant differences between the male and female footprints ($p < 0.001$), suggesting that male footprints are larger for this particular sample, corroborating with the results from previous studies that examine actual foot dimensions. Kanchan et al.'s paper scores 4 using the OLE system for grading literature and scores 12 points altogether for validity. It also offers height predictions for the sample and uses appropriate statistical tests for this part of the study.

4.2.3 Height

The strong association between actual foot length and stature has been recognised by anthropologists for many years, remaining unchallenged since the work of Anderson (1966). It can be assumed foot and footprint dimensions are inextricably linked as the latter is simply a two-dimensional impression of the former, as observed by Barker & Scheuer (1998). The growing foot has been noted to be disproportionate in stature and therefore height calculations necessitate the measurement only of adult feet or footprints (Anderson et al., 1956; Klementa et al., 1973). The literature searched for this section of the chapter focuses on the relationship between stature and footprint dimensions of adult donors. These articles will be appraised in depth in Chapter 6 which provides an analysis of height estimation from the footprint data generated by the research.

Krishan (2008a) examined the dimensions of footprints from one thousand and forty adult males and concluded that length measurements from the base of the heel print to the tip of the longest toe print were the most strongly correlated with subjects' statures ($r = 0.82 – 0.87$) in agreement with Robbins (1986). Width measurements were also positively correlated with stature ($r = 0.52 – 0.66$, $p < 0.01$) but the toe angle of declination, adapted from the Robbins method displayed no statistically significant correlations. Krishan's article scores 4 using the OLE system and 7 points on the OCPM scale.

Using the same measurement method as Krishan, Fawzy & Kamal (2010) did not achieve as high PPM correlation coefficient value. In their study examining the inked static footprint dimensions from fifty male Egyptian medical students,

they found the strongest correlations between footprint and stature were for the footprint length measurements ($r = 0.40 – 0.58$, $p < 0.05$). The best coefficient value was determined by the Calc_A5 length. This is classed as a case report study and scores level 4 using the OLE system. It scores 19 using OCPM rating for validity.

Oberoi et al. (2006) examined the associations between stature and longest footprint length of their sample of one hundred men and one hundred women from India. They found that for their static footprints measurements, there was a strong positive relationship between the two variables for males ($r = 0.70$), females ($r = 0.74$) and the combined group ($r = 0.85$). The study is of a good design as it analysed data from a sufficient number of subjects, suitable for the statistical tests utilised. Also, the inferential and descriptive statistics used appeared appropriate and the study produced informative error rates in the form of the standard error of estimate. This article scores 4 using the OLE system for grading papers and 15 for validity of this part of the research using the OCPM system.

In a similar study conducted by Vidya et al. (2011), the static inked footprints from fifty eight females and forty five males were collected and, as in previous studies, strong correlations between height and footprint length were determined. Using the longest footprint length, the higher correlations were seen in the right footprints compared with the left. PPM correlation coefficients recorded r values of 0.88 for males and 0.82 for females. Critical appraisal of this study resulted in the following hierarchical scores: OLE 4, OCPM 9.

Natarajamoorthy et al. (2011) also established high correlations between height and footprint length. Using a sample of the inked footprints from one hundred and seven Malaysian subjects, the authors determined PPM correlation coefficients for the right foot of $r = 0.74$ (males) and $r = 0.73$ (females). The paper scores 4 using the OLE and 8 for validity using the OCPM system.

As previously discussed, the article by Kanchan et al. (2012) not only presented an exploration of the differences between male and female footprints, it also discussed a study estimating height values of the sample with footprint length measurements. Using a method adapted from the Gunn method (Gunn, 1991) manual measurements were recorded from the heel print to the toe prints of fifty

male and fifty female inked static footprints. Correlations ranged between r = 0.451, R^2 = 0.216, p < 0.01 (Calc_A5 measurement, left footprints, female group), to r = 0.628, R^2 = 0.395, p < 0.01 (Calc_A1 measurement, right and left footprints, male group). These relatively low correlations are in accordance with the results presented in the study by Fawzy & Kamal, discussed earlier. The combined group offered stronger associations between height and footprint lengths, ranging from r = 0.709, R^2 = 0.503, p < 0.01 (Calc_A5, left footprint), to r = 0.787, R^2 = 0.619, p < 0.01 (Calc_A1, right footprint). Stronger correlations of the pooled sample exist because of the larger sample number used for the analysis. The article scores level 4 using the OLE system and achieves 12 points using the OCPM system for grading validity.

The ethnicity of the sample from each of these studies examining the relationship between height and footprint shape may potentially be an influencing factor on regressive outcomes. More studies of this nature using different populations are required for a better understanding of the effects of ethnicity upon the relationship between stature and footprint shape.

4.2.4 Weight

Robbins (1986) measured right and left footprints of five hundred subjects. In this sample, the width measurement across the ball of the footprint displayed the highest correlation with weight (r = 0.72). Descriptive statistics only are presented with no significance values reported for this article. A ratio method for calculating weight was devised, in which subjects' footprint widths were divided by their weights and multiplied by 100. For an adult male right footprint this produced an index figure of 60.48% for the sample. Robbins used the example that if an adult male displayed a footprint width of 100mm, the calculated weight using the index figure would be 75kg. The author then explained that using an absolute number as such would be inappropriate and suggested allowing for an arbitrary ±4.5kg margin. It does not, however, explain how the error margin is formulated and coupled with the mathematical error displayed in the above equation, there remains uncertainty as to the findings relating to footprints and weight associations in this article. Kahane & Thornton (1987) are also critical of this paper, noting that Robbins presented 'unsound and potentially misleading' data (page 9).

Fawzy & Kamal (2010) demonstrated how the forefoot width measurement was the most strongly correlated with weight, concurring with the findings from Robbins' study. The sample involved footprints from fifty adult Egyptian males ranging in age from eighteen to twenty five years. PPM correlation coefficients for the right and left footprint widths were 0.49 and 0.52 respectively. The authors presented regression equations to further their study, affording standard error of estimates (SEE) of 4.05 to 5.28kg. They found that all width and length measurements were statistically significantly correlated with weight except the big toe lengths. This case series study is of level 4 evidence and scores highly on the OCPM validity rating system (19) due to the good design of the study and the quality of its subsequent analysis. However, most of the introduction section of the text appears to follow the opening section of an article written by Krishan (2008a), published in the same journal. In a similar incident in which two paragraphs from an introduction displayed great similarity from articles by different authors, the paper was retracted by the same journal (Roig, 2010).

Krishan also found significant positive correlations with weight and footprint measurements ($r = 0.38$ to 0.75) especially for the Calc_A1 measurement ($r = 0.74$ and 0.75 for left and right feet respectively). The author used inked footprints from fifty adult male Gujjars from North India, aged between eighteen and thirty years (Krishan, 2008c). The study is designed to examine differences in individuals' footprint measurements when the subject is carrying different loads; 0kg, 5kg and 20kg. Non-significant differences were determined between the non-load group and the 5kg load group using t-tests ($p < 0.01$). However, some significant differences were displayed between the means of the footprint measurements of the sample non-load bearing compared with the same measurements from the sample carrying a 20kg load each. These differences were noted for the Calc_A1, Calc_A4, Calc_A5 and the width of the forefoot measurements. Mean errors for the regression calculations of each of the measurements taken ranged from 3.05kg (right foot, Calc_A1) to 4.10kg (right foot, big toe-pad length). The error estimates in this study are smaller than the arbitrarily chosen 4.5kg error margins predicted from the previously described study by Robbins (1986). The footprints in Krishan's study were taken using cyclostyling ink and measurements drawn manually; the latter process potentially incurring a source of error. The author does not report SEEs. His use

of the mean error when reporting error values could be considered a rather blunt statistical instrument as regression calculations are based on the derivation of central tendencies. Applying mean errors to central tendency statistics compounds the values and can result in misleadingly small error margins. This can be described as a case report of OLE level 4 grading and an OCPM score of 8.

As with the studies examining the relationship between height and footprint dimensions, ethnicity may provide an extraneous variable that could limit the conclusions of these studies which explore the associations of weight and footprint shape.

4.2.5 Body Mass Index

No articles could be found by the researcher pertaining to the effects of body mass index (BMI) values upon inked footprint dimensions. A further search uncovered a study by Thompson & Zipfel (2005) considering the relationship between BMI values of their subjects in relation to ethnicity and actual foot dimensions, rather than footprints. This article concluded that there may be a relationship between a high BMI and a larger forefoot width. Although this study considers the dimensions of feet rather than footprints, it may add to the knowledge of footprint behaviour, since the foot is thought to be closely associated with the footprint, as previously discussed in section 4.2.1. A further discussion of the article by Thompson & Zipfel (2005) can be found in section 4.2.7.

4.2.6 Age

No results were found when the search terms 'footprint*', 'foot impression*', measure*, 'dimension*' 'difference*', 'estimat*' and 'age*' were entered into the aforementioned databases. In the absence of relevant literature, the researcher sought information pertaining to the relationship between age and actual foot dimensions instead. One such study by Atamturk & Duyar (2008), examined the feet of five hundred and sixteen subjects split into five separate adult age groups, from eighteen to eighty three years. Descriptive data suggested there was a greater difference for the sixty plus age group, than for all the other more precisely determined age groups. Although this study does not investigate the

estimation of age from footprint measurements, it may bear some relevance on the outcomes of such a study.

4.2.7 Ethnicity

The previously discussed paper by Igbigbi & Msamati (2002) not only explored the data from their sample of three hundred and five Malawians for sex differences as determined by the Arch Indices of inked footprints, but also ethnic differences. The results of the incidences of flat feet in their sample were compared with those from other published studies involving Caucasian samples. These classified the presence of a flat foot in their Caucasian samples using methods other than the AI, including categorisation by visual, subjective means. This could be considered a threat to the validity of this part of the study since the studies were not comparable at baseline, therefore offering little information regarding ethnic differences of footprint shapes.

Although no conclusive literature could be found to examine differences in footprint dimensions between ethnic groups, studies of actual foot dimensions relating to ethnic variations, have been carried out primarily in the field of footwear design. These studies are now briefly explored as they may help to inform the current research pertaining to footprints and its ensuing analyses.

Thompson & Zipfel presented a study to explore the hypothesis that children with a history of walking barefoot rather than shod would display wider forefoot dimensions in adulthood, than those with a predominant history of shoe-wearing (Thompson & Zipfel, 2005). In order to examine the theory, the authors selected sixty urban adult South African females divided into two groups; Caucasoid descent (shod in childhood) and Black African tribal descent (unshod in childhood). However, extensive structured interviews defined both groups had histories of unshod childhoods (80% for black females and 83% for white females) therefore separate groupings were inappropriate for the initial study design since the original thought regarding the shoe-wearing habits of the groups appeared a misconception. However, using paired t-tests, significant differences were reported between the groups when forefoot widths were compared ($p = 0.0484$). Despite this apparent difference in foot dimensions between the two groups, this result may not be solely attributable to ethnicity. The authors report that 40% of black female subjects had a BMI greater than

thirty, compared with the white female subjects (23.3%) and 8.3% of the former group had BMIs over forty. Literature discussed previously in section 4.2.2 has determined that higher body weights in subjects tend to display wider forefoot measurements (Robbins, 1986; Fawzy & Kamal, 2010).

Hawes et al. (1994b) examined the right feet of seven hundred and eight Caucasian North American and five hundred and thirteen Asian (Japanese and Korean) adult male subjects, using measurements taken by a digital caliper. The distance between the base of the heel and the fifth toe for the sample were compared with the maximum foot length and expressed as a percentage. This resulted in 82.60% for the Caucasian group and 85.00% for the Asian group. The authors also discovered that the second toe was longer in comparison to the large toe in 23.91% of the Caucasian group and 49.20% of the Asian group. Ridola et al. found in their sample of ninety seven Italian subjects, 16% displayed a longer second toe (Ridola et al., 2001). Kusumoto et al. observed that the main foot shape difference existing between a group of Japanese female subjects (n = 40) and a group of Filipino females (n= 34) occurred in the prevalence of hallux valgus deformity for the Filipino group (Kusumoto et al., 1996). The authors attribute this pathological difference to the fact that Japanese women wear a style of footwear that has altered since the World War II, whereas Filipino females wear footwear that has remained unchanged since that time. The authors suggest that the prevalence of the deformity in Japanese females has been corrected by better-fitting footwear. Ashizawa et al. (1997) found that in both sexes, Javanese subjects presented a wider foot than subjects from Japan for the same length measurement and in addition, Javanese female feet were relatively wider and longer compared with Japanese male feet regardless of BMI.

4.2.8 Summary of findings from literature review

The literature review infers that in an analysis of footprint data collected from a sample of adults, significant differences may be expected between static and dynamic footprints from the same subject in length, width and footprint angle measurements. Establishing differences between static and dynamic footprint measurements would carry important implications in the practice of forensic footprint identification. For example, Kennedy describes a case in which a footprint impression was found in dust at a crime scene (Kennedy, 2005). Two

suspects were identified; the wife and sister of the murdered victim. Inked impressions were taken of the two women whilst standing and compared with the crime scene print. The crime scene print appeared to match the inked impression from the victim's wife and she was subsequently found guilty of his murder. The match was made despite the fact that the dust impression may have been formed in the dynamic state (the image on page 408 in the publication depicts apparent 'flaring' at the apices of the toe prints, indicative of dynamic footprint capture) yet comparison prints were captured in the static state. If the discriminant analysis establishes significant differences between static and dynamic footprints this must be accounted for in practice.

The literature review also suggested length and forefoot width measurement may display correlations with subjects' weight when the research data are analysed. The longest footprint length measurement is expected to exhibit the strongest correlation with subjects' stature values. The footprint dimensions of the male subjects may demonstrate significantly longer and wider measurements than those of female participants. If the sample were to be separated into groups reflecting different ethnic backgrounds, differences in footprint dimensions may be apparent, although this was not clarified by the literature review as there was an absence of these types of footprint studies.

In order to explore the constructs proposed by the relevant literature, footprints were collected from volunteers at York St John University in 2007.

4.3 Research ethics

Prior to the collection of data for further investigation, a detailed proposal explaining the research was presented to the ethics board at York St John University and scrutinised. Ethical approval was finally obtained from York St John University Research Ethics Committee in April 2007 (Appendix D.1).

A small-scale pilot study was completed to ensure that the chosen method of collecting and measuring inked footprints was appropriate and manageable, previously discussed in Chapter 3.

4.4 Sample

Ideally a representative sample of the population being studied should be gathered (Hicks, 2005). For example, in order to study the effects of a treatment for elderly diabetics, the sample obtained for investigation would consist of elderly people with diabetes. In this way, inferences to the greater population of elderly diabetics can be deduced. It might be argued that a representative sample for the research presented by this thesis would include footprints from crime scenes. These are not available due to the constraints of the legal system, therefore a convenience sample of the general population was chosen instead. This sampling approach was judged to be appropriate to establish the baseline rigour of the actual measurement approach.

The central limit theorem can be used to prove a normal distribution in sample sizes of thirty or more (Landauer, 1997). According to Cohen (1988), given a medium to large effect size, a sample number of thirty will allow for approximately 80% power, the minimum amount of power suggested for an ordinary study. Relevant literature suggested there may be differences between footprints collected from male and female volunteers, necessitating the analysis of homogenous groupings. Therefore recruitment for the final sample deliberately sought a suitable size of both male and female subjects. Flyers distributed around the campus at the University briefly explaining the research and requesting the footprints of volunteers, successfully obtained a convenience sample of approximately forty people. Although convenience sampling has been criticised due to the incorporation of unspecific influences and biases within the sample (Robson, 2002), this did not seem to be problematic for this particular study which sought to establish the rigour of a measurement approach, as opposed to the effects of a drug intervention, for example. Further snowball sampling through word of mouth, increased the sample size to sixty one. The sample was made up of thirty females and thirty-one males with an ethnic composition of 95% Caucasian, 3% Black and 2% Asian with ages ranging from twenty to seventy two years (Table 4.1).

Each interested participant received details of the study explaining its aims, method and expected involvement (Appendix D.3). Supplied consent forms explaining that participants could withdraw from the study at anytime if necessary were signed

(Appendix D.4). Participants were asked to volunteer information regarding their perceived racial background as previous studies have suggested that foot and footprint shape differs between ethnic groups (section 4.2.7). The statures of the volunteers involved in the study were taken at York St John University using a Class III SECA (SE001) Leicester Portable Height Measure, meeting the current Department of Health standards. Measurements were taken according to the method described by Weiner & Lourie (1969). Height values were recorded in centimetres by the main researcher and verified by a research assistant in attendance. Weights were recorded in kilograms by way of a Tanita WB 100 S MA portable floor scale, Class III (in accordance with Non-Automatic Weighing Instruments Directive, 2000). The instruments used for the collection of weight and height data were kindly leant to the researcher by the nutrition and dietetic department at Harrogate District Hospital. Body mass indices were calculated using the following formula for each subject, as suggested by Dowling et al. (2001): weight (kg) / (height (m)) ^{2}. A summary of descriptive statistics for the sample is displayed in Table 4.1.

Table 4.1 Descriptive statistics for male and female subjects

Sex		Age (yrs)	Height (cm)	Weight (kg)	BMI (kg/m²)
Male	Mean	42.35	176.90	81.77	26.13
	SD	14.42	5.98	11.24	3.38
	N	31	31	31	31
Female	Mean	37.77	163.43	65.67	24.59
	SD	9.56	6.73	13.50	5.03
	N	30	30	30	30
Total	Mean	40.10	170.28	73.85	25.47
	SD	12.39	9.27	14.73	4.30
	N	61	61	61	61

SD Standard deviation

Exclusion criteria included;

- Persons under 20 years of age, ensuring the foot was of full size (Tortora & Grabowski, 2003)
- Insensate feet
- An inability to walk independently
- Persons with a known foot pathology e.g. arthritic conditions, surgery, recent trauma including partial loss of foot tissue
- Persons with a known foot infection or open wounds on the foot, e.g. heel fissures, ulcerations, fungal infections, verrucae, to prevent cross-infection.

4.5 Method

The right footprint of each of sixty one volunteers was captured three times using an inkless paper system supplied by Crime Scene Investigations Ltd, in both static and dynamic states, described in section 3.2. All footprints were collected in Temple Hall at York St John University in June 2007.

Each footprint was coded for anonymity and prints that were too faint, smudged or extended beyond the borders of the paper were excluded. The footprints were

scanned using an Epson scanner set at 150 dots per inch. Lengths, widths and angles were constructed and measured using the GNU Image Manipulation Program (GIMP), previously described in section 3.4. All recorded values were stored on a secured computer.

4.5.1 Statistical analysis

Statistical analysis of the results from this and all further explorations of the data were carried out using SPSS software.

The mean of three values for all measurements from the static footprints of each subject and similarly from the dynamic prints, were compiled in a condensed data set using SPSS. These measurements were then utilised in an exploration of inter-footprint measurement relationships as well as an investigation of how other variables such as height and age would possibly affect the footprint measurements in both static and dynamic states. If the data is parametrically supported, PPM correlation coefficients (r) are recommended for the calculation of the strength of relationship between two variables. The resulting coefficient from this calculated value falls between -1 and +1 (Field, 2005). A correlation coefficient of 0 would indicate no linear relationship exists. Positive correlations may be interpreted as suggested by Innes & Straker (1999), cited by Reneman et al. (2002) and are as follows; $r \leq 0.5$ ($R^2 \leq 25\%$) little similarity or poorly correlated, r 0.51 – 0.75 (R^2 26-56%) some similarity or moderately correlated, $r \geq 0.76$ ($R^2 \geq 75\%$) substantial similarity or highly correlated. Correlation values reported alongside associated p-values, allow the reader to infer the likelihood of a repeated occurrence of the correlation, if a further experiment were to be performed at another time.

The literature review for this section revealed tests to determine not only associations as derived by correlation but also differences between variables. Differences have typically been explored using paired sample t-tests and analysis of variance (ANOVA). Paired sample t-tests compare the means of two variables. The difference between the two variables for each case is calculated, and tested to see if the average difference is significantly different from zero. The effect size of resultant t-test values can be calculated by converting into an r-value as suggested by Rosnow & Rosenthal (2005) using the following equation;

$$r = \sqrt{(t^2/(t^2 + df))}$$

The effect size informs the researcher as to whether the t-value is substantive or not, in practical terms. Effect sizes may be interpreted in the following manner; r = 0.10 small effect, r = 0.30 medium effect and r = 0.50 large effect (Cohen, 1988; Cohen, 1992; Field, 2005).

For correlation, paired t-testing and ANOVA calculations, all variables were tested for normality, the assumption of which is a necessity for these statistical tests to produce meaningful results. Kolmogorov-Smirnov (K-S) tests were used for this part of the research, as the mathematical conclusion is deemed suitable for sample sizes over fifty (D'Agostino, 1971). K-S tests determined all footprint measurements were non-significant ($p > 0.05$), except the length measurement from the base of the heel to the apex of the fifth toe (Calc_A5) in the dynamic state ($p = 0.03$), suggesting a deviation from normality (Appendix F.1). Field (2005) advises the K-S test has its limitations and recommends additional plotting of data and to make 'an informed decision about the extent of non-normality,' (page 93). Upon further analysis of Calc_A5 data, a Quantile-Quantile (Q-Q) plot demonstrated some deviation from normality at both extremities (Appendix F.2). However, no outliers were displayed in the box-whisker plot for this particular measurement, and the histogram appeared to have a normal distribution (Appendix F.2). This is in contrast to the footprint angle which demonstrated an 'S-shaped' Q-Q plot, several outliers in the box-whisker plot and kurtotic distribution in the histogram (Appendix F.3), despite K-S tests suggesting normality for this angle measurement (Appendix F.1). In a further descriptive analysis of these two footprint measurements it was shown that the Calc_A5 measurement data had a large variance compared with the angle measurement, possibly because the length measurement is calibrated in millimeters, whereas degrees are used for the angle measurements – a much smaller unit. As the data from the Calc_A5 measurement appeared to be normally distributed through a range of normality tests, it was decided that this length measurement data should be regarded as normally distributed, along with the other footprint measurement data and therefore suitable for PPM correlation tests.

The direction of relationships and differences between variables can be predicted from the information gleaned from the relevant literature. For example, a positive correlation would be expected between the footprint length measurements and height values for the sample. Using the guidance from previous literature, the experiments carried out throughout this thesis involve one-tailed testing for significance, meaning that only one end of the normally distributed results is examined. This allows for a more appropriate test of statistical significance because more power is provided to detect an effect in one direction by not testing the effect in the other direction (Field, 2005).

4.6 Results

Descriptive statistics for sixty one subjects are displayed in Table 4.1.

4.6.1 Inter-relationships between footprint measurements

The length, width and angle measurements taken from both static and dynamic footprints of sixty one participants displayed moderate to high PPM correlation coefficients (r = 0.64 to 0.97) when each static measure was paired with its dynamic counter-measurement (Table 4.2).

Table 4.2 PPM correlation coefficients of static and dynamic footprint measurements for sixty one subjects

Static and dynamic paired linear measurements	Correlation	Static and dynamic paired angle measurements	Correlation**
Calc_A1	0.94**	Footprint Angle	0.90**
Calc_A2	0.96**	Dist. Met. Angle	0.64**
Calc_A3	0.96**	1-5 Toe Angle	0.90**
Calc_A4	0.97**	2-5 Toe Angle	0.81**
Calc_A5	0.96**	2-4 Base Angle	0.86**
MPJWidth	0.96**		
CalcWidth	0.91**		

** $p < 0.01$

In a multiple correlation analysis involving all static and dynamic footprint measurements, it was determined that the length and width measurements were all moderately to highly correlated ($r = 0.68$ to 0.95, $p < 0.01$); however although statistically significant, the angle measurements were poorly correlated with the length and widths, and of these, some were negatively correlated ($r = -0.26$ to 0.26, $p < 0.05$).

In order to illustrate the poor relationship between the linear and angle measurements, a simplified correlation table using just the length measurement Calc_A1 and the angles was constructed (Table 4.3). As the lengths and widths were so highly associated, choosing just one measurement out of this group was considered to be appropriate. Since paired dynamic and static measurements were strongly associated (Table 4.2) the simplified correlation table represents the static measurements only.

Table 4.3 PPM correlation coefficients of static Calc_A1 measurement and all other static footprint measurements

Static linear measurements (n=61)	Calc_A1 (static) PPM correlation	Static angle measurements (n=61)	Calc_A1 (static) PPM correlation
Calc_A2	0.95**	Dist. Met. Angle	0.01
Calc_A3	0.95**	1-5 Toe Angle	0.17
Calc_A4	0.93**	2-5 Toe Angle	0.17
Calc_A5	0.91**	2-4 Base Angle	0.06
MPJWidth	0.69**		
CalcWidth	0.78**		

** $p < 0.01$

In a further investigation of the relationships between static and dynamic width and length measurements, paired sample t-tests (Table 4.4) demonstrated statistically significant differences between the static and dynamic pairings,

except for the MPJWidth measurement (df (60), t = -1.32, p = 0.19, effect size = 0.05).

Table 4.4 Paired sample 't' test for static and dynamic footprint linear measurements

	Paired differences					df 60
				95% CI of the difference		
	Mean (mm)	SD	SE	Lower	Upper	t-value
Calc_A1 (D)–(S)	17.41	5.87	0.75	15.91	18.91	23.17**
Calc_A2 (D)–(S)	12.59	5.16	0.66	11.26	13.91	19.05**
Calc_A3 (D)–(S)	10.87	4.62	0.59	9.68	12.05	18.39**
Calc_A4(D)–(S)	9.30	4.00	0.51	8.28	10.33	18.16**
Calc_A5 (D)–(S)	9.61	5.02	0.64	8.33	10.90	14.95**
MPJWidth (D)–(S)	-0.23	1.34	0.71	-0.57	0.12	-1.32
CalcWidth (D)–(S)	0.68	1.73	0.22	0.24	1.12	3.08**

(D) Dynamic (S) Static, SD Standard deviation, SE Standard error, CI Confidence interval, df degrees of freedom, **p < 0.01

Bivariate correlations between height and the footprint measurements in both static and dynamic states showed moderate to high associations with the linear measurements; r = 0.60 to 0.84, p < 0.01. The strongest correlation (r = 0.84) was seen in the Calc_A5 measurement. The angle measurements displayed non-significant coefficient values.

Using ANOVA a significant interaction effect of height on the length and width measurements was noted; df (1, 24), F values ranging from 2.06 to 6.92, p <

0.01, effect size ranging from 0.30 (CalcWidth dynamic) to 0.71 (Calc_A5 static) for all linear measurements. There was a non-significant effect of height on the angle measurements.

Descriptive data suggested differences between the male and female footprints existed. Table 4.5 illustrates the mean values for static footprint linear measurements.

Table 4.5 Mean measurement values for three static linear footprint measurements

Measurement	Males (n = 31)		Females (n = 30)	
	Mean (mm)	SD	Mean (mm)	SD
Calc_A1	251.62	12.95	223.70	11.22
MPJWidth	98.96	5.04	87.59	4.31
CalcWidth	52.51	3.99	45.35	2.95

In a factor analysis, multivariate tests suggested that there was a significant difference between the static and dynamic length measurements ($p < 0.01$) but there were no significant differences concerning the sex factor ($p = 0.48$). In other words, sex has no significant influence on the static/dynamic differences.

In tests between footprint measurements effects, a general linear model factorial ANOVA determined that there was a significant main effect of sex upon the length and width measurements in both static and dynamic states (df (1,30), F values ranging from 84.51 to 119.28, $p < 0.01$, effect size ranging from 0.44 to 0.59 for all linear measurements). There was a non-significant main effect of sex on the angle measurements (df (1, 30), F values ranging from 0.01 to 1.12, $p = 0.29$ to 0.95, effect size ranging from 0.01 to 0.02 for all angle measurements).

The error bar graph (Figure 4.1) illustrates these differences between male and female footprint length measurements in both static and dynamic states.

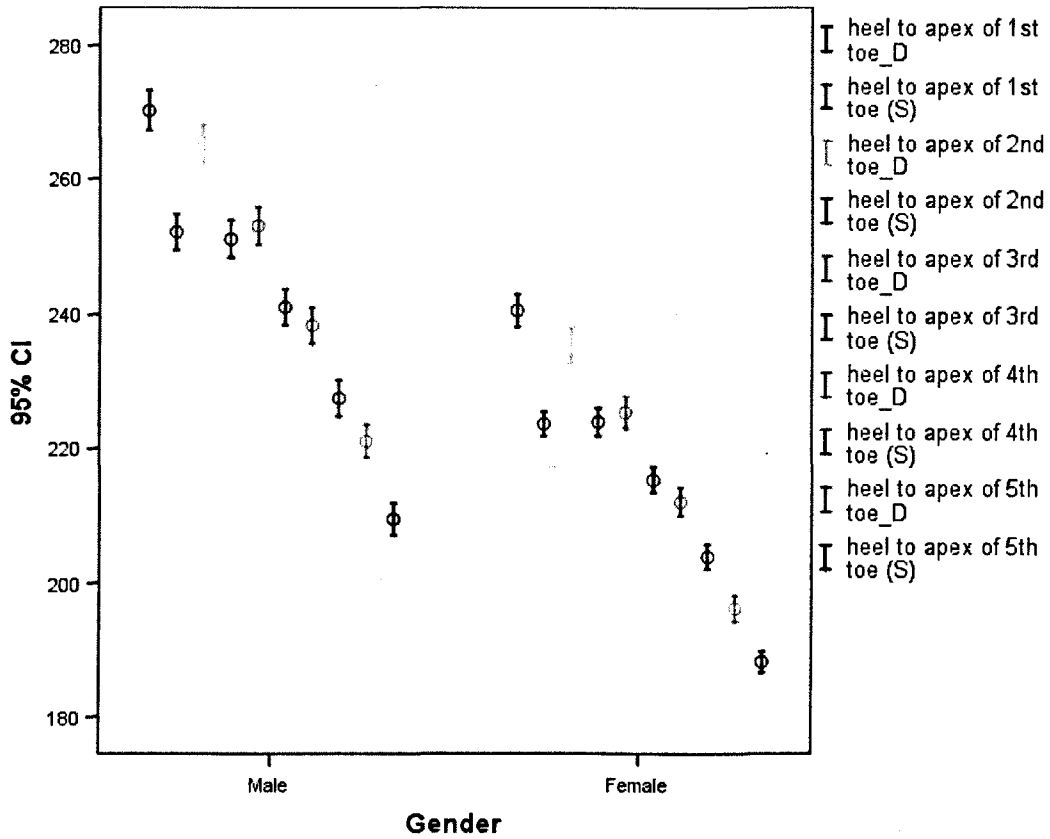


Figure 4.1 Graph illustrating differences between static and dynamic length measurements for male (n = 31) and female (n = 30) footprints

(S) = Static, D = Dynamic

In the above graph, the error bars represent the scores expressed in the context of their respective 95% confidence interval of the length measurements from the base of the heel to the apices of each of the five toes in both static and dynamic states. The means of these scores are denoted by the circle along the bar. The graph highlights the significant differences between the static and the dynamic length measurements, the static lengths being shorter. It can also be seen that male footprints behave in a similar trend to female prints, but there are differences in the magnitude of the measures between the sexes. The male print lengths suggest a greater range for each score and, as expected, were larger than the female prints.

There was a mean paired difference between the dynamic and static Calc_A1 lengths for the male footprints of 17.90mm (SD 5.39; 95% CI 15.92 to 19.87). For the same length measurement of the female footprints, the static and

dynamic mean paired difference was 16.91mm (SD 6.37; 95% CI 14.52 to 19.28).

An investigation of the interaction of ethnicity and age suggested non-significant effects for both static and dynamic measurements in both male and female footprints. For example the effects of ethnicity resulted in effect sizes ranging from 0.01 to 0.07 ($p > 0.05$). Correlation coefficients between the linear measurements in both static and dynamic states with weight were statistically significant but only moderate in terms of strength of association (r ranged from 0.44 to 0.63, $p < 0.01$). Out of all of the linear measurements, the MPJWidth measurement in the static state displayed the highest correlation with weight ($r = 0.63$). BMI displayed little similarity to the static and dynamic linear measurements (r ranged from -0.01, $p > 0.05$ Calc_A5 static; to 0.28, $p < 0.05$ WidthMPJ static).

4.7 Discussion

Coefficients resulting from correlating the paired static and dynamic measurements determined that they were highly and positively related to each other, in accordance with Mathieson et al.'s findings (Mathieson et al., 1999). Multiple correlations between the length, width and angle measurements showed that the angle measurements were highly correlated with other angle measurements. Length and width measurements were more strongly correlated to one another, but there was poor correlation between the angles and the linear measurements. A further exploration using multiple correlation analyses between each individual static and dynamic measurement with the other measurements displayed poor inter-relationships between angles: most were non-significant and some were negatively correlated, in other words, associations were seen to travel in different directions. Physiologically, linear measurements should behave in the same way between static and dynamic stances, but due to extraneous factors such as rotation and twisting, angle measurements possibly do not behave as similarly between the static and dynamic states compared with the linear ones. In Chapter 2, it was determined that the literature pertaining to toe angle measurements was unsupported. Despite support from orthopaedic publications concerning the footprint angle discussed in Chapter 2, the new measurement approach is underpinned by the

methods offered by the previous literature which concern the linear measurements only. Therefore the angle measurements do not fit the traditional model of forensic footprint measurement and therefore less emphasis is placed on these measurements in subsequent analyses within this thesis.

Static and dynamic linear measurements displayed statistically significant differences for all pairings except the MPJWidth measurement which appeared not to vary between the states of standing and walking. The discovery holds important implications for footprint identification as it possesses a predictive potential for the remaining footprint length and width measurements. In forensic evaluation of footprints, the extent of variability of a bare foot impression from a person is currently unknown (DiMaggio & Vernon, 2011). The amount of variation is potentially increased if the footprint is captured on a rough or absorbent surface, or if the person is twisting, running or turning when the impression is formed. A measurement which remains constant between varying states, as exemplified by the MPJWidth length, may act as a common point of reference, the basis of which can be used to seed further research.

Differences between static and dynamic length measurements have been established for the sample used in this validity study, agreeing with other studies from the literature in this area (Kippen, 1993; Barker and Scheuer, 1998). Understanding variation (error estimates) that can be incurred when a print is made in either or both of these two states is helpful. However, a consistent measurement as seen in the analysis of the MPJWidth would further support an evaluation of this kind. The measurement across the widest part of the forefoot does not alter between standing and walking, possibly because of its fibro-elastic architecture which becomes firm and tense in anticipation of ground reaction forces (Erdemir et al., 2004). Also, Weijers et al. suggest the soft tissue volume within the forefoot is displaced dorsally during peak loading in the gait cycle, and not laterally and medially, as one would expect (Weijers et al., 2003).

Bivariate correlations determined the Calc_A5 measurement was most strongly correlated with height, agreeing with the results of Fawzy and Kamal's study, but in contrast to other studies in this area which suggest Calc_A1 is the most strongly correlated with the height variable. This interesting phenomenon will be

examined in greater depth in Chapter 6 (Establishing Evidence of Predictive Validity).

Differences between the footprint shape of males and females was supported by the previous literature, reviewed in section 4.2.2. An analysis of descriptive data of the sample revealed females displayed shorter, narrower footprints than males. It may be no coincidence that males on average were taller and heavier than females (Table 4.1) and therefore anthropometric differences rather than differences between the sexes can be attributed to the difference in footprint measurements, supporting the study by Oberoi et al. (2006). In other words, it can be postulated that the differences between male and female footprints are more likely to be explained by height and weight differences and not sex differences.

The investigation into the effects of weight values on footprint shape variation resulted in statistically significant moderate correlations with MPJWidth in both static and dynamic states. This is in accordance with the findings of Fawzy & Kamal (2010).

Non-significant associations of age and ethnicity with the static and dynamic measurements were seen in this sample, contrary to literature in the relevant area. Splitting the subjects into groups by age range as demonstrated in Atamturk & Duyer's study would have resulted in an analysis depleted of data leading towards a Type II error in which differences between groups that may have been present are not detected (Field, 2005). Similarly the ethnic composition of the participants used for the purposes of the research presented in this thesis did not involve groups large enough for adequate analyses; therefore this is an area requiring further research.

4.8 Conclusions

This chapter set out to explore the differences (discriminant validity) and associations (convergent validity) between footprint measurements from the sample and also between footprint measurements and other factors of interest within the sample, such as weight and ethnicity. These relationships were predetermined using information gleaned from an appraisal of the relevant literature and statistically tested to verify the extent of construct validity. High

correlations were determined between height and footprint length measurements, particularly the Calc_A5 measure and will be investigated further in the thesis. This was supported by the literature. What was not supported however, were the non-associations between footprint linear measurements, particularly width and the weight variable. Ethnicity, age and BMI factors also did not contribute to the analysis. Statistically significant differences were established between static and dynamic footprints from the same person. Differences were also noted between male and female footprint dimensions supporting the previous literature. Error estimates relating to differences and associations in the dimensions of a foot impression from a person and between people, validate the use of the measurements for forensic identification purposes. As described in Chapter 2, establishing validity is a desired element of securing trustworthiness of a new test or technique. To simply state there are differences and associations between these measurements would support Black's definition of validity of a technique in a law context, in that it should demonstrate 'sound and cogent reasoning' (Black, 1988, page 599). The scientific definition of validity goes further and requires knowledge of error estimates relating to these relationships in the dimensions of a foot impression from a specific person and between people. Only when the extent of variability has been calculated using an appropriate sample, can the footprint measurements be said to have construct validity. In forensic identification, understanding differences between static and dynamic footprints from the same person goes some way in explaining footprint dimension variation, overlooked by forensic practitioners, exemplified by the work of Kennedy (2005) described earlier in section 4.2.8 of this chapter, in which a static print was used for comparison with a dynamic print. Here, Kennedy's adoption of Black's definition of validity in this 'real-world' case may be considered to be inappropriate.

Chapter 5
Establishing Evidence of Concurrent Validity with supporting Reliability Analysis

5.1 Introduction

This thesis offers a new footprint measurement approach. Establishing validity by examining the appropriateness of the interpretation of the resultant measurement data is a key element of the study. This chapter will present an exploration of criterion-related validity of the approach in the form of concurrent validity. For clarity, the newly developed approach will be termed the Reel method from this point on.

5.2 Concurrent validity explained by the relevant literature

The approach to test validation in this chapter examines the accuracy and relevancy of measurement scores, and has been described as an on-going evaluative process (Wood, 1989). Accuracy and relevancy can be determined through correlational procedures whereby the new approach is compared with existing methods (Safrit, 1989). The previous critical appraisal of the general literature (Chapter 2) uncovered several methods for evaluating two-dimensional footprints. In Chapter 3, the researcher exposed the preferred methods of footprint assessment used in the field by practitioners. Measurement data afforded by the Reel method are expected to relate well to scores resulting from other footprint evaluation methods that measure the same characteristic. Criterion-related evidence quantifies the relationship between two or more different tests or techniques. A form of criterion-related validity is concurrent validity, defined by Cronbach and Meehl as existing 'when one test is proposed as a substitute for another or a test is shown to correlate with some contemporary criterion' (Cronbach & Meehl, 1955, page 281). Concurrent validation tests often involve the comparison of the new technique or test under investigation with a gold standard criterion test (Norton & Ellison, 1993; Portney & Watkins, 2000; Leard et al., 2004; Souza & Powers, 2009). Correlations between the resultant sets of scores allow for an analysis of the degree of comparability (an indication of accuracy). The criterion measure selected is

assumed to be an established and valid indicator of the variable of interest (Portney & Watkins, 2000). For example, Leard et al. (2004) demonstrated that a figure-of-eight method of measuring oedema in the hand was highly correlated with the traditionally employed gold standard volumetric measurement method. The figure-of-eight method utilises measurements recorded by a standard tape measure when wrapped around the afflicted hand between the base of the fingers and the wrist, crossing under and over the thumb. The volumetric method appeared more costly and time-consuming compared with the new figure-of-eight method; the resultant high correlations from the study allowed the authors to recommend the use of the latter instead (Leard et al., 2004). A gold standard is traditionally thought to be a test that has previously been proven to possess high levels of reliability, validity and accuracy (Pereira-Maxwell, 1998). Portney & Watkins (2002) recognise that the selection of an appropriate criterion measure can be an onerous task, especially if there is an absence of a universally accepted gold standard as is the case in footprint measurement. In this thesis, it has transpired that all currently accepted methods used in the field do not have adequate evidence of reliability or validity; however they are assumed to measure similar constructs in that they all evaluate footprint shape in some way. Claassen (2005) argues that a gold standard test is usually chosen as it is the best available tool that can be used for comparison. In support of this, Sechrest (2005) demonstrated that traditionally accepted gold standard tests such as blood pressure syphygmomanometry have not established robust evidence of validity or reliability. The previous suggestion, that concurrent validity can be examined when the quality of the new test is compared with a gold standard test, is therefore thrown to conjecture. A general search of the literature using the database MEDLINE and search terms 'concurrent AND validity' exposed four hundred and eighty six articles. Of these, only forty five mention the expression *gold standard* in the related text and many compared a new test with several tests, exemplified by Reneman et al.'s study (2002). This article describes the comparison of four well-known instruments used in the field of clinical rehabilitation to assess disability performance and self-reported disability. Poor to moderate correlations were demonstrated between the four instruments and

the authors of the research concluded that this exhibited poor concurrent validity.

5.3 A literature review of concurrent validity studies within the forensic identification sciences

Using the search terms 'concurrent AND valid*' a literature search of the forensic identification science journals failed to retrieve any articles. Selected journals included those with the highest impact factors in the area such as Forensic Science International (Impact Factor: 1.821) and Science and Justice (Impact Factor: 0.966). A total of nine forensic science journals were content-searched including articles within the fields of DNA, ballistics, odontology and fingerprint evidence. A lack of relevant article retrieval is of no surprise, given the current criticisms of the paucity of empirical testing of forensic methods as debated by authors such as Saks & Faigman (2007) and Cole (2008). This has also been one of the key issues in the NAS report (2009) and the document detailing the codes of practice and conduct in the forensic sciences (Rennison, 2011). These criticisms have been previously discussed in Chapter 2. A broader search of forensic literature outside the identification sciences using the same search strategy, proffered many articles pertaining to psychometric analysis, for example those comparing a mental assessment tool with another established method for categorising offenders (e.g. Douglas & Webster, 1999; Strand et al., 1999). Those forensic science studies that evaluate the concurrent validity of a method or a test fall mainly into the category of psychological assessment tests. This is perhaps not surprising, given that measurement research in the social sciences widely cites the seminal work of Cronbach & Meehl (1955) in which the various concepts of validity in psychological testing are discussed. However, Sechrest (2005) argues that Cronbach & Meehl's definitions of construct, content and criterion-related validity are appropriate for all sciences affiliated with measurement, for example, the measurement of blood pressure. Therefore, it is appropriate to explore concurrent validity of a new footprint measurement approach.

5.4 Choice of tests for comparison

According to Vernon (2007) and DiMaggio & Vernon (2011), the Gunn method, described in section 2.4.2.3 and the Optical Center Method (OCM) (section 2.4.2.5) are the most commonly utilised approaches for measuring footprints for identification purposes. Research in the field of forensic anthropology refers to footprint measurement using the Robbins method; however, further appraisal of these articles suggests it is the Gunn method that has been used (e.g. Krishan, 2008a; Fawzy & Kamal, 2010; Kanchan et al., 2012). In addition to these two approaches, the Kennedy method (Kennedy et al., 2003; Kennedy et al., 2005) was considered a suitable comparison test for a concurrent analysis as it incorporates elements from both the Gunn method and OCM.

5.4.1 Gunn Method

The Gunn method (Gunn, 1991) has been described as 'one of the primary methods selected by the examining podiatrist in the footprint comparison process' (DiMaggio & Vernon, 2011, page 59). In this method, five length lines are drawn from the rearmost aspect of the heel print to the uppermost part of each toe print. In addition, a width line is drawn across the area of the ball of the footprint. The six lines are then measured for comparison. Line construction and measurement is often achieved by means of manual methods using a pen and ruler, both in practice as described by DiMaggio & Vernon (2011), and for research purposes (Krishan, 2008a). A search of the literature uncovered some technologies that measure physiological features both manually and digitally. A comparable area of study is cephalometry, applied for clinical purposes (dental, maxillofacial surgery, orthodontics) and also for forensic identification. Traditionally cephalometric analysis has been carried out by measuring lengths, widths and angles of radiographs with a pen, ruler and protractor (Polat-Ozsoy et al., 2009) and is comparable with the Gunn method of evaluating footprints. The manual method of cephalometric analysis is considered time-consuming, introduces a higher degree of operator error and the radiographs require greater storage capacity; therefore an automated digitised method is preferred (Polat-Ozsoy et al., 2009; Thurzo et al., 2010). In a similar cephalometric analysis study, Thurzo et al. employed 95% limits of agreement to show higher agreement and greater accuracy for the scores from the digitised and

automatically measured lengths, widths and angles compared with the manual measurements (Thurzo et al., 2010). Therefore exploration of evidence of concurrent validity with manual measurements using the Gunn method was not considered for this part of the study. However, the Gunn method has further been developed by using graphics editing software such as Adobe PhotoShop® for the construction and measurement of the lines (Vernon, 2006). As the cephalometric studies suggest higher reliability is afforded by digitised methods compared with manual methods, the Reel method incorporates the digitised version of the Gunn method. The researcher opined that the subjective decision as to the appropriate rearmost pixel from which to construct length lines could increase error values therefore resulting in lower reliability estimates compared with the Reel approach. Findings relating to this particular aspect of the Gunn method will be discussed further in this chapter.

5.4.2 Optical Center Method

According to the relevant literature, the OCM employs the software application AutoCAD for assessing two-dimensional footprint images (Bodziak, 2000; Kennedy et al., 2003; Kennedy et al., 2005). The process can also be carried out manually using an overlay of concentric circles drawn onto acetates and then placed directly on the prints, described by Winklemann (1987) in a published case-study. As the Reel method involves the analysis of digitised footprints using a software application, the use of the manual circle overlay for the OCM was discounted as a means of establishing evidence of concurrent validity, the software application AutoCAD favoured instead. Comparisons with the OCM using AutoCAD were included in the concurrent validity analysis.

5.4.3 Kennedy Method

In addition to the more popular methods for evaluating footprint dimensions, the Kennedy approach was also investigated for this part of the study as it not only incorporates optical centres as part of its methodology, but also an alignment process employing a central axis (Kennedy et al., 2003).

This method defines the centre of the heel not from the optical centre, but instead from the intersection of the line of the central axis with the line connecting the inner and outer tangents across the image of the heel prints. Measurements from this central heel point to the optical centres of the toe prints

are favoured in these studies of footprint individuality (Kennedy et al., 2003; Kennedy et al., 2005).

5.5 Method

5.5.1 Reliability analysis

Before any instrument is implemented for the evaluation of footprints, it is desirable to examine its measurement properties for the extent of reliability. Reliability, as previously discussed, is also a form of validity. A reliability analysis of a measurement tool will provide estimates of both random and systematic error (Hicks, 2005). Intra-rater reliability more specifically refers to the consistency of scores when a footprint is measured on different occasions by the same observer (Robson, 2002). Statistical analyses from these types of studies can explore both relative and absolute reliability. According to Baumgartner (1989), relative reliability refers to the degree to which a person maintains their position (ranking) in terms of their footprint measurements in a sample over repeated measurements. Relative reliability can be determined by the use of correlation coefficients such as the ICC. Absolute reliability reflects the degree of variation that occurs between repeated measurements for each person's specific footprint measurements results, in other words, the less the measurements vary, the higher the reliability (Baumgartner, 1989). Absolute reliability can be measured by way of the SEM and 95% LOA (Bruton et al., 2000). The advantage of these statistics over indicators of relative reliability is that it is easier to extrapolate the results for the comparison of reliability estimates between different measurement tools (Atkinson & Nevill, 1998). Therefore assessments of relative and also absolute reliability of the chosen measurement approaches (Gunn, OCM, Kennedy and Reel) were considered important components in support of the analysis of construct validity.

5.5.2 Validity analysis

In a previous analysis, all length and width values were highly correlated with one another (Chapter 4, section 4.4). Therefore it was considered unnecessary to evaluate all measurements for the following analysis of concurrent validity and reliability of the different approaches, and only one measurement was constructed and measured for each method. The measurements involved the

heel print and the large toe print for all methods, chosen above other length lines as it was noted that some of the lesser toes failed to print for some footprints. The measurements thus consisted of the line from the aligned base of heel print to apex of the large toe print (Calc_A1) in the Reel method; the optical centre of the heel print to the optical centre of the large toe print (OCC_OC1) in the OCM; the base of the heel print to the apex of the large toe print with no prior alignment of the image (Calc_A1_NCA) in the Gunn method; and the bisection of the central axis in the heel print to the optical centre of the large toe print (CAC_OC1) in the Kennedy method (Figure 5.1).

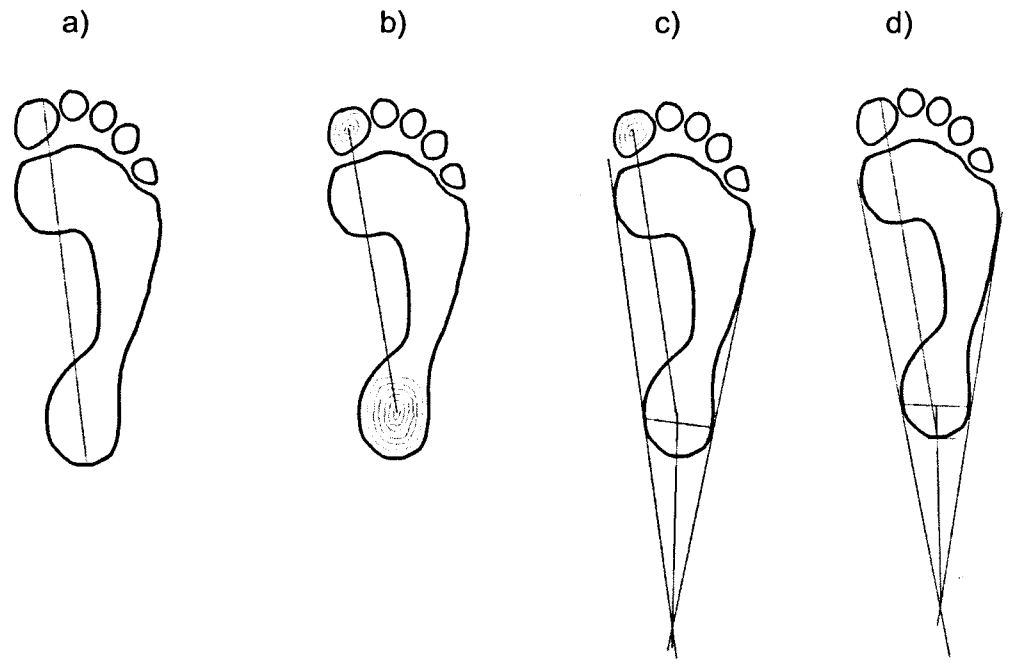


Figure 5.1 Measurements from the heel print to the large toe print for different methods used in forensic evaluation. From left to right: a) Adapted from the Gunn method (Gunn, 1991), b) adapted from the OCM (Bodziak, 2000), c) adapted from the Kennedy method (Kennedy et al., 2003) and d) the Reel method.

Optical centres were derived using AutoCAD®2010 software installed onto a Windows PC. The 'spline' option in AutoCAD Classic drawing mode allows for the creation of optical centres on certain features of the imported footprint image, such as the heel and the toe prints. Instead of creating many inwardly concentric circles to locate the central point of a toe/heel print, the 'through' command avoids the necessity to generate multiple offsets and immediately draws a central location point from which to start and end the construction of measurement lines. The central axis and length lines were created as previously described using 2D drawing tools offered by AutoCAD and subsequently measured in millimetres up to two decimal places using the same software.

5.5.3 Data analysis

All measurements were constructed and recorded twice on separate occasions by the same rater for a repeated-measures style analysis.

Difference data were assessed for normality using K-S tests, Q-Q plots and histograms. Intra-rater reliability was calculated in all measurement methods using a one-way random effects model ICC based on values provided by prior analyses of variance. The one-way model deems that all variance detected by the statistical test, is assumed to be measurement error (Fleiss, 1986, Baumgartner, 1989). Accounting for total error in this way has the effect of lowering the ICC value. This is in contrast to the two-way random effects model which partitions resultant variance into systematic and random error (Baumgartner, 1989). In this way, only one type of error is accounted for and may allow for a raised ICC. The conservative one-way model is therefore considered more rigorous. Using the former ICC methods of nomenclature, the one-way analysis would be the equivalent of $ICC_{3,1}$ (Portney & Watkins, 2000). Ninety-five percent confidence intervals were calculated for all ICC values. ICCs were interpreted using the following reliability criteria as suggested by Shrout: 0.00-0.10, virtually none; 0.11-0.40, slight; 0.41-0.60, fair; 0.61-0.80, moderate; 0.81-1.00, substantial (Shrout, 1998). As ICC scores are susceptible to sample heterogeneity, SEM and graphs of 95% LOA (Bland & Altman, 2003) were constructed to investigate absolute reliability estimates. The SEM is the amount of error to expect in any single footprint's measurements according to the method used. It is calculated by the following equation: $SEM = SD\sqrt{(1- ICC)}$,

where SEM = standard error of measurement, SD = standard deviation of the sample and ICC = the calculated intraclass correlation coefficient (Baumgartner, 1989; Thomas et al., 2005). For a true value within 95% CI limits, the formula 1.96 x SEM was applied.

Concurrent validity of footprint measurement was assessed by examining the correlations of one measurement between the Gunn, Kennedy, Optical Center and the Reel Method, using PPM correlation coefficient (r). According to Reneman et al. (2002), strict criteria to establish concurrent validity has yet to be identified in the literature but regardless, a similarity of measurement results displaying a strong correlation would provide good evidence of this type of criterion-related validity. In other words, the measurement test results should all display a high degree of shared variation as determined by R^2. Therefore the coefficient of determination (R^2) was also calculated by squaring r to determine the proportion of variance in one variable explained by the second variable (Wood, 1989). Correlations were interpreted as suggested by Innes & Straker (1999), cited by Reneman et al. (2002) described previously in section 4.5.1.

5.5.4 Sample

The calculation of the ICC for estimating reliability is based on a prior analysis of variance. A requirement of the one-way repeated measures ANOVA is that the dependent variable follows a normal distribution. Normal distribution can be confirmed by the use of the central limit theorem in sample sizes of thirty or more (Landauer, 1997). According to Cohen (1998), given a medium to large effect size, a sample number of thirty should provide approximately 80% power, acceptable for this type of study.

Differences have been noted between male and female footprints (section 4.4.1). Therefore fifteen digitised footprint images from male volunteers and fifteen from female volunteers were randomly selected using SPSS software from the database with the following constraints: Within each male and female group, seven images were captured in the static state, seven in the dynamic state, plus one other random print (static or dynamic). The heterogeneity of the sample ensures that the homogeneity of variance is challenged allowing for a more rigorous analysis. The same collection of thirty footprint images was used

in the analysis of concurrent validity and reliability of three established methods (OCM, Kennedy and Gunn) and the Reel method.

5.6 Results

Histograms and Q-Q plots determined that parametric analysis was supported as all variables were shown to be normally distributed. K-S statistics were non-significant for scores from all measurement methods ($p = 0.20$ for all methods).

A summary of descriptive statistics is shown in Table 5.1

Table 5.1 Summary item statistics

Approach	Mean	Min	Max	SD	Range
Gunn	244.46	202.50	276.60	18.16	74.10
OCM	193.58	164.29	220.26	13.51	55.97
Kennedy	201.18	169.23	227.12	14.76	57.89
Reel	243.84	203.00	275.20	18.03	72.20

5.6.1 Results from reliability analysis

A summary of results from the reliability analysis is shown in Table 5.2.

Table 5.2 Summary of results from the reliability analysis for all methods

Approach	ICC (95%CI)	LOA (Upper Lower)	SEM (mm)	95% SEM (mm)
Gunn	0.99	1.79	0.57	1.13
	(0.99-0.98)	-1.40		
OCM	0.96	8.68	2.71	5.31
	(0.92-0.98)	-4.76		
Kennedy	0.99	3.48	1.14	2.23
	(0.98-0.99)	-3.31		
Reel	1.00	0.61	0.05	0.10
	(1.00-1.00)	-0.41		

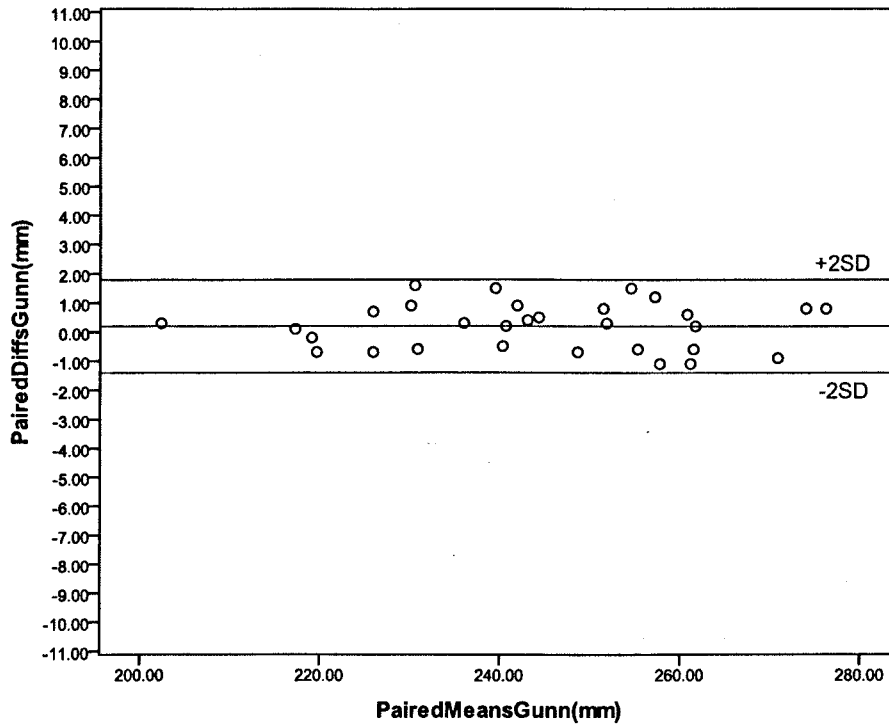


Figure 5.2 LOA graph repeated Gunn measurement

Graphs of limits of agreement are shown in Figures 5.2 – 5.5. The red line in each graph represents the mean difference between the repeated tests, and the blue lines define the limits of agreement (±2SD).

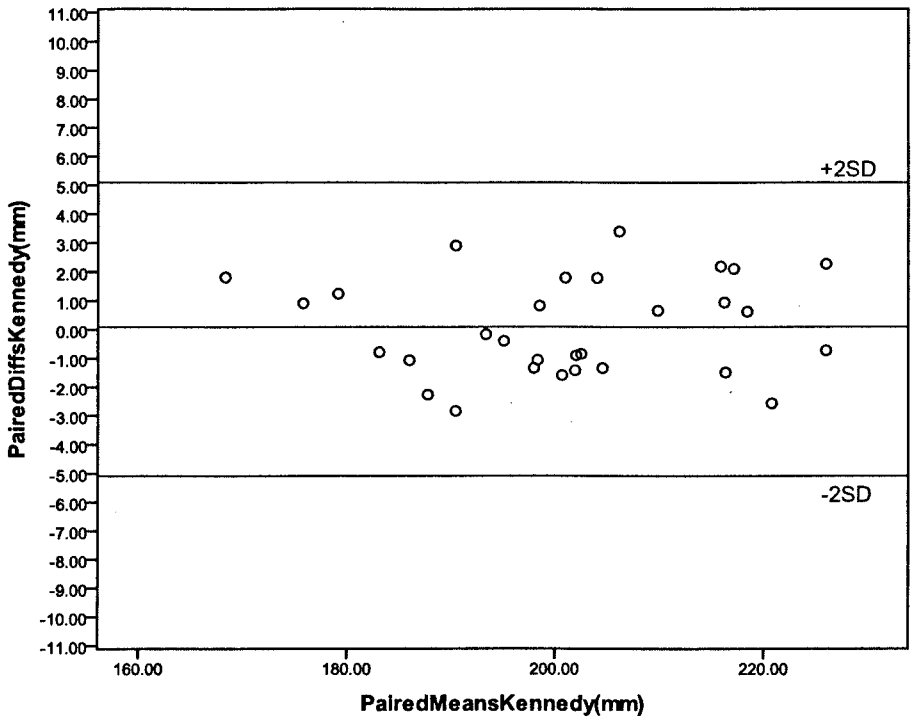


Figure 5.3 LOA graph repeated Kennedy measurement

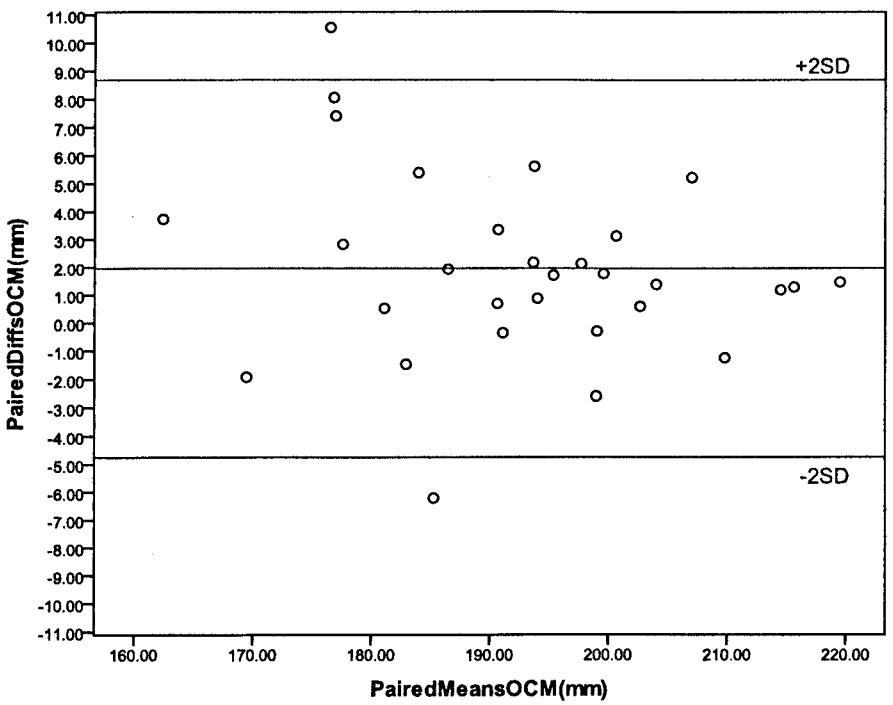


Figure 5.4 LOA repeated OCM measurement

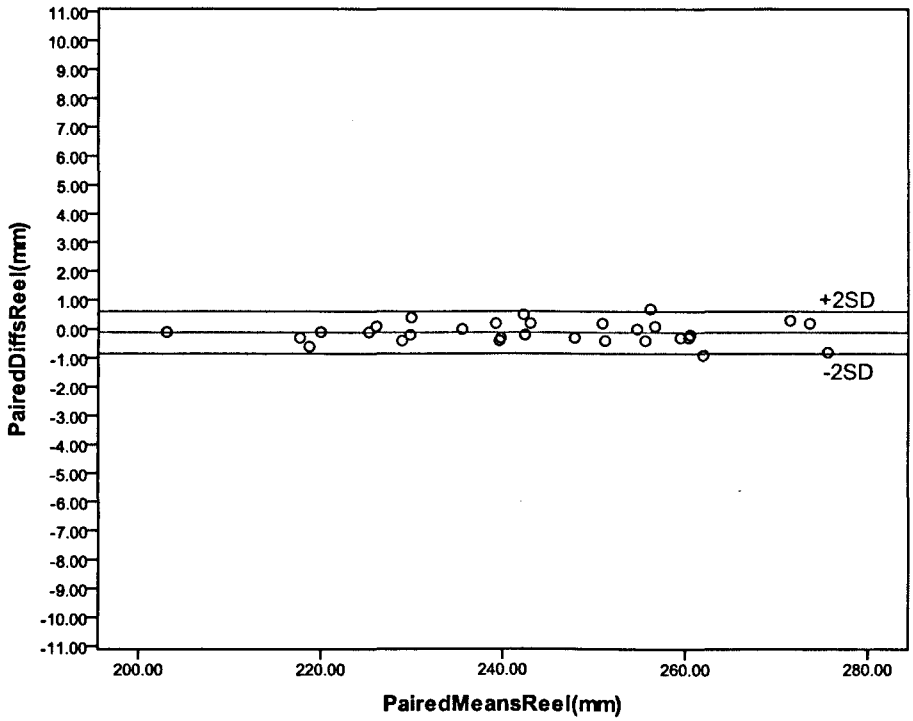


Figure 5.5 LOA repeated Reel measurement

5.6.2 Results from validity analysis

Measurement scores were correlated with scores derived from the Reel method to investigate levels of concurrent validity. Of the three methods investigated, all displayed strong positive correlations with the Reel approach and therefore substantial similarity. A summary of the correlation analysis is shown in Table 5.3.

Table 5.3 Results of correlation analyses

Approach	Reel	
	r	R^2
Gunn	0.99**	0.98
OCM	0.87**	0.76
Kennedy	0.89**	0.79

** Correlation significant at the 0.01 level

r = Pearson product-moment correlation coefficient

R^2 = Coefficient of determination

The substantial associations of the three methods under scrutiny with the Reel method are illustrated in the following scatterplots (Figures 5.6 to 5.8).

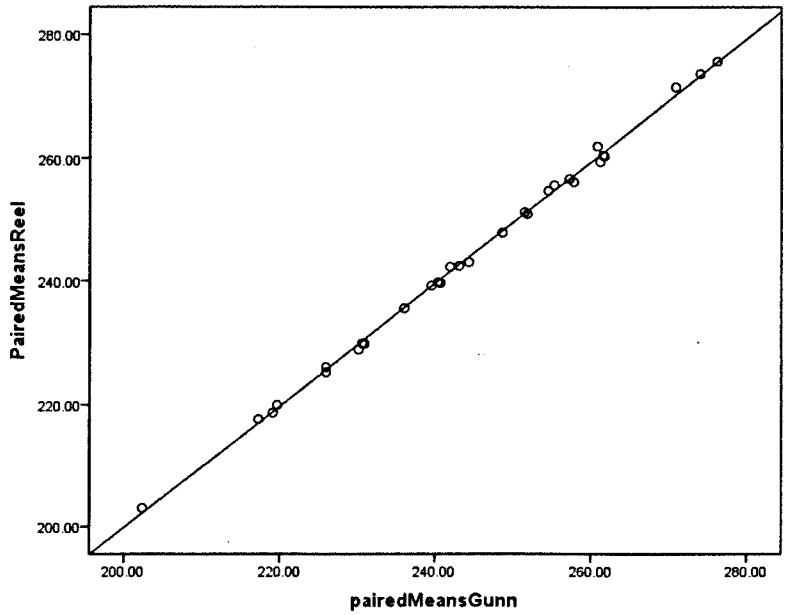


Figure 5.6 Scatterplot of Gunn method paired mean measurements with Reel method paired means

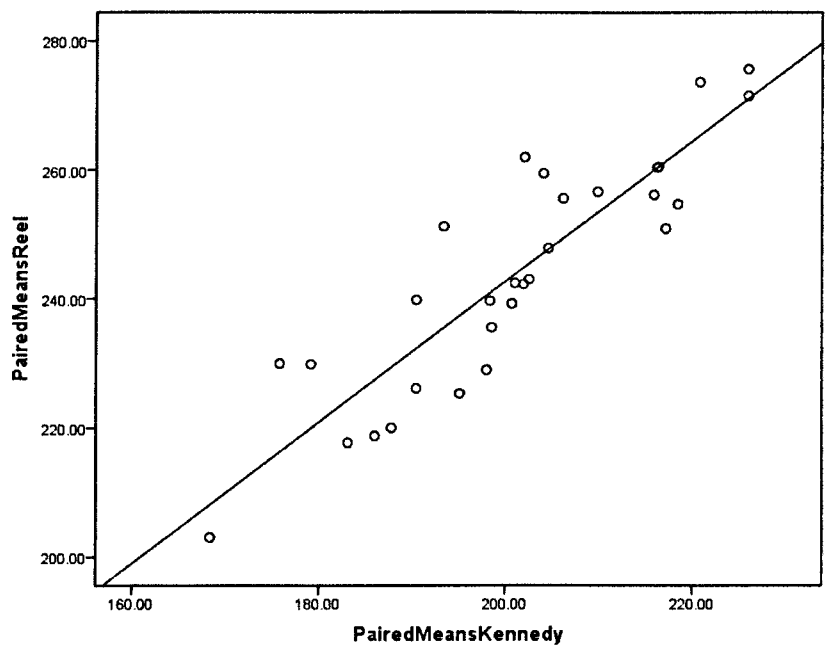


Figure 5.7 Scatterplot of Kennedy method paired mean measurements with Reel method paired means

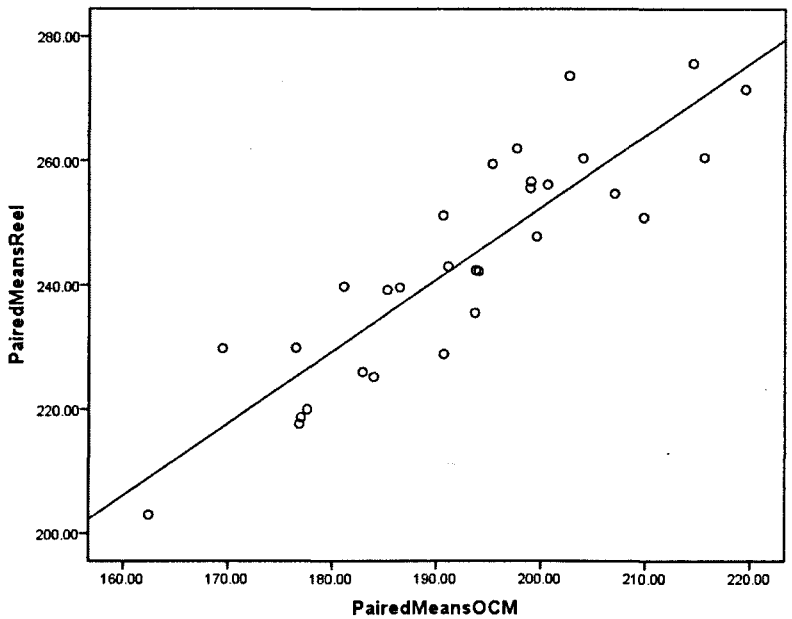


Figure 5.8 Scatterplot of Optical Center Method paired mean measurements with Reel method paired means

5.7 Discussion

Descriptive statistics illustrate the measurement differences between the approaches using full footprint lengths (Gunn, Reel) and those measuring to and from centres of footprint features (Kennedy, Optical Centre method), in that the latter group have smaller values. Associated standard deviations are reflected accordingly. The reliability analysis demonstrated that the ICC was substantial for all approaches, the highest value afforded by the Reel approach (ICC 1.000) and the lowest by the OCM (ICC 0.962). Graphs illustrating 95% limits of agreement reflect absolute reliability and visually describe the agreement between two sets of measurements. They reflect the relationship between the mean value with the variance of the measures and can identify outliers and bias. Results are recorded in the form of two values that lie within the 95% limits; one occurring below the mean value and one above the mean value (Bland & Altman, 2003). In order to interpret findings from the LOA, the values must be taken in relation to the range of recorded measurement values. For example, a variance of –1.403 to +1.796 millimetres when repeatedly measuring a characteristic for a sample that measures say an average of 244.365mm (as demonstrated in the measurements using the Gunn method), demonstrates high agreement between tests and therefore high reliability.

Upper and lower boundaries of the LOA graphs were derived from standard deviations calculated from paired sample t-tests. The graph for the OCM demonstrated a greater heteroscedasticity of scores compared with other approaches suggesting relatively poor consistency. Paired differences were the furthest away from 0.0 suggesting relatively poor reliability for the OCM compared with other methods. The Reel method displayed the closest paired differences to the value of 0.0 and the narrowest band between upper and lower limits of agreement, suggesting this method demonstrates greater repeatability and consistency compared with the other methods analysed (Figure 5.5). Scores from repeated tests using the Reel method fell within 1mm of one another. The graph indicates the presence of two outliers, explained by the close proximity of the upper and lower limits.

SEM values, expressed in millimetres, also reflect absolute, rather than relative reliability. According to Atkinson & Neville (1998), SEM includes only 68% of the variability rather than the conventional 95% criterion used in confidence intervals. Considering the intra-rater reliability SEM of 2.71mm as demonstrated by the reliability analysis for the OCM approach, approximately 95% of the time, the true value of measurement length should fall within ±5.31mm of the measured value (Portney & Watkins, 2000).

A summary of method comparisons for reliability estimates is shown in Table 5.4. This table displays a simplified summary. If a high estimate was achieved for a particular statistical test, a positive sign (+) was allocated to the method in question. If the test produced a low score overall for the method, a negative sign (–) was allocated. Agreement was determined by evaluating the ratio of positives to negatives to create a verbal expression.

Table 5.4 Summary of reliability comparisons

	Gunn	Kennedy	OCM	Reel
LOA	+/-	-	-	+
ICC	+	+	+	+
SEM	+	+/-	-	+
Agreement	high	close	low	very high
Rating	good	mild	poor	excellent

The calculations of PPM correlation coefficients and coefficients of determination demonstrated substantial correlations for all methods compared with the Reel approach, although the OCM displayed the weakest relationship ($r = 0.89$, $R^2 = 0.79$, $p < 0.01$). This is not surprising as this method differs the most from the Reel method in its construction of measurements; however the association between the two methods is still considerable as demonstrated by these results. The highest association occurred between the Reel approach and the Gunn method ($r = 0.99$, $R^2 = 0.99$, $p < 0.01$).

5.8 Conclusion

The new approach was seen to be most highly associated with the Gunn method of measurement, of little surprise since the two methods only differ in the choice of the rearmost pixel. Kennedy et al. (2003) noted variations of measurement occurred when a subjective selection of the rearmost pixel was made from which to construct the five footprint lengths (section 3.4). High correlations between the Reel method and the Gunn method do not appear to support this observation by Kennedy et al. However 95% SEM values were smaller for the Reel method by a difference of 1.03mm compared with the Gunn method possibly reflecting a weakness of this subjective pixel selection. All reliability estimates for repeated measures of the Gunn method were substantial and acceptable; however the Reel method demonstrated higher estimates. Daubert criteria and recommendations from the Law Commission and the NAS report demand new technologies or tests to have rigorous scientific foundations before admissibility in a court of law. In this study, the Reel method offers the best reliability estimates compared with the other approaches investigated and therefore qualifies as a primary consideration of measurement choice in the forensic analysis of footprint impressions. The extent of reliability of the Reel method will be discussed further in Chapter 7.

Substantial PPM coefficient values and coefficients of determination confirm that concurrent validity of the Reel approach is strongly supported. The forensic community-established measurement procedures were highly correlated with the Reel method. Collectively, this supports the utility of the Reel method as an alternative estimate of footprint measurement.

Chapter 6
Establishing Evidence of Predictive Validity

6.1 Introduction

For the Reel approach to be useful in the field and for research purposes, it must demonstrate acceptable levels of rigour (Daubert v Merrell Dow Pharmaceuticals Inc., 1992; The Law Commission, 2009; National Research Council, 2009). In the previous chapter, substantial levels of evidence of concurrent validity, a type of criterion-related validity, were ascertained. This next section of the thesis sets out to explore evidence of another type of criterion-related validity of the new footprint measurement method. This type of validity aims to quantify relationships between scores from two separate variables and is known as predictive validity (Wood, 1989). The primary objective of this chapter will be to establish evidence of criterion-related validity of the Reel method in terms of predictive validity.

6.2 Predictive validity

In establishing the predictive validity of an experiment, typically both regression and correlation are used in the design (Wood, 1989). Predictive validity can be defined as the assessment of the new measurement approach's ability to predict another variable it should theoretically be able to predict, such as stature (Safrit, 1981). Measurements (test scores) derived from the footprint images using the Reel approach could theoretically be used to predict directly, through regression equations, height values (criterion scores) for the sample. This in turn validates the use of such scores (Wood, 1989).

In the forensic arena, previous studies have explored and supported the belief that the shape of the human footprint is unique (Cassidy, 1980; Kennedy et al., 2003; Kennedy et al., 2005). Currently, these studies remain unchallenged therefore footprints could be regarded as evidence to eliminate or link a suspect to a crime scene (Gordon & Buikstra, 1992; Krishan, 2008a). Prediction of the height of an unidentified person, with known error margins, from a single footprint, is thus of interest in the field of forensic identification. In the absence

of other factors such as age, sex or race, the predictor variables for height estimation are accrued from the two-dimensional footprint measurements, in other words, lengths, widths and angles (Robbins, 1985; Krishan & Sharma, 2007; Fawzy & Kamal, 2010).

An initial examination of the data for general trends revealed the footprint length measurement from the base of the heel to the smallest toe print (Calc_A5) resulted in the highest correlation with stature for the sample (section 4.6). This relationship also demonstrated the largest R^2 value. Prior knowledge attained from the appraisal of published articles regarding the relationship between height and actual foot length suggested this may have been an unusual finding; most studies had investigated the longest foot length with stature to calculate regression equations and associated error margins (e.g. Sanli et al., 2005; Zeybek et al., 2008; Rani et al., 2011). A further literature search of general anthropometrical literature and also in the area of stature estimation from footprints was necessary to inform investigations of predictive validity for this chapter.

6.3 Searching the literature

A review of anthropometrical literature will be presented first explaining human proportionality and the foot in terms of stature. This will be followed by a more detailed critical appraisal of the literature concerning footprint dimensions and stature estimation. The literature for this chapter was explored by using the databases CINAHL, MEDLINE, and AMED. ZETOC alerts were set up by the researcher producing additional, more recent, publications. Serendipitous searching, described previously in section 2.4, also afforded further material for review.

6.3.1 Anthropometrical literature review

The strong association between foot length and stature has been recognised by anthropologists for many years (Anderson, 1966). However the growing foot has been noted to be disproportionate of stature and therefore height calculations necessitate the measurement only of adult feet or footprints (Anderson et al., 1956; Klementa et al., 1973). Boys and girls have been found to display different rates of skeletal growth due to hormonal causes (Stavlas et

al., 2005). Grivas et al. (2008) determined that using regression equations that adjust for sex and age, it is possible to calculate the height of a child from a foot length with the best prediction having error estimates of ±6.03cm. Nevertheless heterogeneous samples using children and adult feet or footprints for the prediction of stature are inappropriate.

Artists, anthropologists and clinicians have been interested in the proportions of the body for many years. Anthropologists examine and compare associations and relationships between different parts of the body to try and understand the influences of ethnicity and lifestyle (Sanli et al., 2005). Richer & Hale (1973) suggested that it was the early Egyptians who first proposed rules pertaining to the proportions of the body. Anderson et al. (1956) observed that the foot grows simultaneously with the rest of the body, and not just to the appended leg suggesting a relationship between the foot and overall stature, rather than the lower limb. Various anthropometric studies have shown that relationships between body parts and stature vary between populations due to differences in levels of nutrition, physical activity, climatic changes and familial variation (Malina et al., 1983; Ashizawa et al., 1997; Katzmarzyk & Leonard, 1998; Krishan, 2008b; Sanli et al., 2005; Bogin & Varela-Silva, 2010).

A study by Lamm et al. (2006) determined that the foot achieves maturity much earlier than the femur or tibia. This has important implications for stature estimation from footprint dimension studies, as it suggests that height predictions from footprints cannot be achieved with acceptable probability until full skeletal maturity has been gained. Indeed measurement values taken from immature feet (less than twenty years) appear to skew data leading to questionable regression calculations. This phenomenon is illustrated in the articles by Robbins (1986), Giles & Vallandigham (1991) and Gordon & Buikstra (1992). For example, in their study involving a large sample of army recruits, Giles & Vallandigham (1991) acknowledged large overestimations of height values using regression analysis. They suggested the reason for this discrepancy may have been due to the young age of the sample in which full maturity and size may not have been attained.

It has been observed that males do not reach skeletal maturity up to the age of twenty years (Trotter and Gleser, 1952; Hertzog et al., 1969; Tortora & Grabowski, 2003).

A decline in stature has been noted in people after the age of thirty, especially in females (Knight, 2004). However, the study presented by Kanchan et al. (2008) found that age did not statistically significantly affect the prediction when using foot lengths even though their sample, split into relevant age groups, included ages up to eighty years. These results regarding the age variable are supported by the findings of the study by Sen & Ghosh (2008) which examined height predictions from foot lengths from a sample ranging in age from eighteen to fifty years.

It is thought that human growth responds to the overall quality of living conditions during developing years, making it a highly plastic phenomenon (Bogin, 1999). For example, iodine deficiency during childhood may result in reduced lengths of the tibia, femur and foot (Anderson, 1966). MacDonnell (cited by Giles & Vallandigham, 1991) was possibly the first to publish a study noting variations in stature amongst populations. This study examined the relationships between foot length and stature and involved three thousand male prisoners from the British Isles. As part of the conclusion to this early study, MacDonnell wrote of his concerns of possible population variance affecting results – the average height measurement for his sample of prisoners was 8.4cm less than a sample of one thousand university students at that time. Conclusions obtained from this underpinning anthropometrical literature search are that estimations of height from the dimensions of the foot are population variant, an important consideration in the interpretation of results.

6.3.2 Studies examining the estimation of stature from footprint dimensions

In comparison to studies that examine the relationship between actual foot dimensions and stature there appears a paucity of literature observing the associations of footprint dimensions with stature. This may be due to the musings of previous authors who have suggested that calculating stature from footprint dimensions is unachievable and unnecessary (e.g. Gordon & Buikstra, 1992; Barker & Scheuer, 1998). This is contrary to the opinions of those who

have researched this area (e.g. Robbins, 1985; Krishan, 2008a; Fawzy & Kamal, 2010). Four articles were uncovered using the aforementioned databases with the search terms 'footprint*, AND stature OR height, AND predict* OR estimate*'. A further five publications were found after serendipitous searching as described in Chapter 2, and also through ZETOC alerts. These nine articles will now be critically appraised and scored using the hierarchical OLE and OCPM systems for quality and validity of the published research. Some of the articles have previously been introduced in a review of the literature regarding the influence of height variation upon footprint shape, in which discriminant validity of the measurement approach was sought (section 4.2.3). Since the present chapter explores predictive validity, the focus of appraisal centres on the predictive elements of the selected articles, rather than the affects of height upon footprint shape, as was discussed in Chapter 4.

American anthropologist Louise Robbins correlated height with right footprint length data from a sample of five hundred and fifty subjects (Robbins, 1986). In this study, the length measurements Calc_A1 and Calc_A2 showed the greatest correlation with height ($r = 0.84$) although the other three length measurements also displayed good correlations ($r = 0.83$). The author determined that dimensions of the arch width and toe prints offered the lowest correlations with height (e.g. arch width, $r = 0.25$; toe pad width, $r = 0.44$). Robbins presented a scattergram illustrating the positive correlation between height and footprint length (Calc_A1). The reader is invited to predict height by choosing a footprint length on the y-axis and by using a ruler to define the intersection of the slope, subsequently obtain the appropriate height on the x-axis. However, interpolation is difficult because of the unusual units used (14mm units for foot length, 98mm for height). The author then suggests making an allowance for variance by including 'a plus-or-minus factor' in the height estimation 'of ±25mm' (page 201). There is no statistical support of this reported error estimate, conveniently a measure equivalent to one inch. Regression equations are not submitted in this publication as the author explains in a following associated publication that the method for achieving these calculations is 'unduly complicated' (Robbins, 1986, page 147). Robbins' sample of five hundred and fifty subjects included three hundred and thirty one volunteers between the ages of fifteen and twenty

years, threatening the study's internal validity as previously discussed. Robbins' data analyses of all her studies relating to foot and footprint research have been highly criticised by others (Tuttle, 1986; Giles & Vallandigham, 1991). The study can be classified as a case report using the guidelines set out by the Oxford centre for evidence-based medicine levels of evidence, and is therefore graded at level 4. It scores 3 out of 24 using the Ohio College of Podiatric Medicine grading system for validity.

Strong, positive correlations were found to exist between the longest static right footprint length and the heights of one hundred men and one hundred women from eighteen to twenty six year old Indian students (Oberoi et al., 2006). This study, previously reviewed in section 4.2.3, determined that PPM coefficient correlations were 0.70 for males, 0.74 for females, and 0.85 for the combined group. Statistical significance associated with these results is not reported. Linear regression analyses revealed formulae resulting in SEE values of 4.66cm (males), 4.58cm (females) and 4.77cm (combined group). The study scores 4 using the OLE system for grading papers and 15 for validity of this part of the research using the OCPM system.

Atamturk & Duyar (2008) used a Turkish sample of three hundred and sixteen volunteers and calculated correlations and regression equations of stature with footprint dimensions. The authors found the longest footprint length measurement demonstrated the highest correlations with stature ($r = 0.734$, males; $r = 0.663$ females, $p < 0.01$). The regression equation incorporating the factors age, sex, foot length, footprint length and breadth demonstrated the best error estimates (R^2 0.81; SEE 4.43cm). The method in which the footprints were collected may have threatened the validity of the study's design, as subjects were requested to 'wet their soles totally in buckets of water and then step on tracing paper, so as to facilitate measurements' (page 1297). There are no references cited to support this method and one would imagine a wicking-effect of the wet feet on the paper to occur, possibly confounding the footprint dimension measurements. It warrants an OLE rating of level 4 and scores 11 for validity using the OCPM system.

Using a sample of one thousand and forty subjects, Krishan examined footprints and foot outlines to estimate stature (Krishan, 2008a). The author explains the

importance of bare footprint studies in relation to identification at crime scenes, especially in developing countries where residents tend to walk unshod, citing Qamra et al. (1980) and Sharma (1970) in support of this. Footprints were collected using cyclostyling ink and jute bags. Static prints were collected and whilst weight-bearing, the foot of each subject was drawn around using a pen. The prints were measured using Robbins' method (Robbins, 1985); however the figure illustrating measurements recorded for the study appear more akin to the Gunn method (Gunn, 1991). The author used t-tests to ascertain bilateral asymmetry and PPM correlation coefficients were applied to examine relationships between foot measurements and stature. The division factor method was utilised for the calculation of stature from footprint measurements. Here, the mean foot length of the sample is divided by the mean height of the sample, then multiplied by 100. In a later paper by the same author, the division factor method is cogently argued to be a poor indicator of stature estimation from body parts and footprints compared with linear regression analysis (Krishan et al., 2012). Results of division factor method and regression analyses were accompanied by mean errors. These were determined by calculating the differences between the estimated stature and the actual stature (resulting from the division factor or regression calculation). This is really the paired differences between the results and could be considered to be a descriptive analysis only, as opposed to the standard error of estimate which is a measure of unexplained variation. Highest correlations with stature were seen in the footprint length measurements (r ranging from 0.82 to 0.87, $p < 0.001$). Length measurement Calc_A1 displayed the highest correlation in the left footprint and Calc_A2 length showed the best correlation in the right footprint ($r = 0.87$, $p < 0.001$). The regression equation with the smallest mean error for the left footprint was for the length measurement Calc_A1 (stature = 3.689 x Calc_A1 length + 84.013, mean error 2.12cm). For the right footprint, the measurement with the lowest mean error was for Calc_A2 (stature = 3.361 x Calc_A2 length + 91.303, mean error 2.15cm). Despite these low mean errors, Krishan states that 'the precise prediction from an individual's footprint or foot outline may be an unachievable and unnecessary goal and there would always be an estimation error of a few centimetres,' (page 98) echoing the sentiments of previous authors (Robbins, 1986; Barker & Scheuer, 1998). Finally the author compares

the actual stature with estimated stature values in his sample of one thousand and forty adult genetically isolated males. Mean values suggest a five millimetre difference between the actual height and estimated height for the sample. Such apparent accuracy of the estimation method using regression equations may be as a result of the large number of subjects involved in the study. Regression analysis is dependent on the derivation of central tendencies. The central limit theorem decrees that the larger the sample, the more closely the sample means will be distributed about the population mean (Rowntree, 1981). Estimating mean values for the actual and estimated stature values could have resulted in the magnification of these averages, obscuring variation about the means. Using the OLE scoring for an overall grading of the quality of the literature, this paper is of level 4 standard and scores 7 when rated using the OCPM system.

Fawzy & Kamal (2010) collected static inked footprints from the right and left feet of fifty subjects. Nine measurements were then taken on each footprint by drawing over them with a pen and ruler. The authors acknowledge Robbins (1985) in terms of the measurement method; however the illustration (page 885) appears to depict the method as described by Gunn (1991). Measurements included the five lengths from the base of the heel to the tips of the five toe prints, the widest part of the heel, the widest part of the forefoot, big toe breadth and big toe width. The highest correlation of height and footprint dimensions was found to exist between the right foot Calc_A5 measurement ($r = 0.58$, $p < 0.05$). This measurement on analysis also had the smallest SEE (3.52cm) and largest R^2 value (0.33, $p < 0.05$). The regression equation for this measurement was calculated to be: Stature = 92.57 + 3.72 x Calc_A5 (right foot). Little reference is made to these findings regarding this particular length measurement. The paper concludes by reiterating other authors' opinions in terms of ethnicity and body size studies, declaring '...it is suggested that similar studies should be conducted in different parts of the world so that the effect of genetic and environment can be investigated in forensic terms,' (page 888).

Despite great similarities between the introduction of this study and that of a publication by Krishan (2008a), Fawzy & Kamal's study offers a comprehensive statistical analysis and interpretation of the collected data. The study is graded level 4 quality using the OLE system, and scores 16 for validity rating using the OCPM scoring system.

A similar study was conducted by Vidya et al. (2011), in which static footprints were collected from fifty eight females and forty five males. The footprint length from the base of the heel print to the longest toe print was recorded and correlations with footprint lengths and breadths versus stature were determined. The right footprint length resulted in the highest correlations with stature ($r = 0.88$ males, $r = 0.82$ females, $p < 0.05$). A regression analysis using all components (right and left length and breadth footprint measurements) showed that the linear regression calculation involving the longest right footprint length of the female group possessed the smallest SEE value of 0.91cm. The calculation for the male group for the same footprint length showed a SEE value of 1.45cm. The error estimates appear to be low and are more akin to the results from the calculation of the standard error, rather than the SEE. There is no explanation in the article as to how the SEE was derived. In the article, a table of results is displayed that is combined with a table that appears to have been taken from the published study of Oberoi et al. (2006) bearing no relevance to Vidya et al.'s results. There is no reference to this table or acknowledgment that it is the work of Oberoi et al. The article is of case study level 4 quality using the OLE system and scores 9 according to the OCPM grading system for validity.

Measurements taken from inked static footprints are used for estimating heights from a Malaysian sample consisting of forty two males and sixty five females (Natarajamoorthy et al., 2011). The longest footprint length was derived from both the Robbins method, described as a parallel method and the Gunn method, described in the study as a diagonal method. Correlations with height were high; $r = 0.874$ for the Robbins method longest length and $r = 0.875$ for the Gunn method longest length measurement in the combined male and female group. For the homogenous groupings, correlations with height were similarly high; for the male footprints, $r = 0.743$ for Robbins method longest length measurement and $r = 0.747$ for Gunn method longest length measurement; and for the female footprints, $r = 0.733$, Robbins method, $r = 0.747$, Gunn method. It is interesting to note that the differences between the correlations using the two methods of measurement are seemingly negligible, although the Gunn method presents slightly higher correlations in all cases. Unfortunately, statistical

significance of the calculated correlation coefficients is not reported in this paper.

The authors offer regression equations with supporting SEEs. For the male footprints the longest footprint length using both the Gunn method and Robbins method predict height with an SEE of 4.8cm. For the female group the SEE was smaller allowing for a range of plus or minus 3.4cm. The combined group of males and females presented regression equations with SEEs of 4.17cm for both length measurements. The small SEE values calculated for the separate male and female groups may have resulted from using homogenous samples with too few data points. For example, the male footprint group consisted of forty two measurements, considered too small to be meaningful when calculating regressive equations (Green, 1991; Miles & Shelvin, 2000). The descriptive data regarding the height variable suggests a small range (144cm to 183cm) with a small associated SD of 8.53 compared with other sample data from published studies. As a consequence of the relatively small range and spread of height values, it is inevitable that smaller SEE values will result, compared with other stature estimation studies. The study is of level 4 using the OLE system as it is a case report of a group of subjects, and scores 8 using the OCPM system for validity.

Inked static footprints from fifty males and fifty females from an Indian population are analysed for prediction of stature (Kanchan et al., 2012). Using the Gunn method, lines are manually drawn and measured from the base of the heel to the tips of the toe prints and ensuing data used for the analysis. The authors argue that the advantages of using the Gunn method are that stature estimations can be used for partial footprints. The sample is divided into homogenous male and female groups and also pooled data is utilised for correlation and regression calculations. Positive correlations of the footprint lengths with height were observed for all measurements, which the authors describe as 'strong' (page 4). These correlations range from $r = 0.407$, $p = 0.001$ for the Calc_A3 length in the left footprints for the female group, to $r = 0.628$, $p < 0.000$ for the Calc_A1 in the left and right footprints for the male group. The pooled group suggested stronger correlations ranging from 0.709 (Calc_A5 left footprint) to 0.787 (Calc_A1 right footprint), $p < 0.000$. To describe the ranges as strongly correlated is perhaps a little misleading, for example

Innes & Straker (1999), cited by Reneman et al. (2002) suggest $r \leq 0.5$ as having little similarity and r 0.51 – 0.75 as having some similarity. The pooled data from the Calc_A1 measurement for the right footprints is the only measurement to possess strong correlations with height values. Probability values of less than 0.000 as stated in Kanchan et al.'s article are inconceivable; a p-value of 0.000 is representative of a value that is less than the decimal places shown, but cannot be zero or less than zero.

Linear regression involving the five length measurements for right and left footprints for male and female groups, demonstrated the Calc_A5 length measurement for the right footprints in the male group produced the smallest SEE in height prediction (4.11cm). For the pooled sample, the Calc_A1 measurement of the right footprints produced the smallest SEE value (6.55cm). The pooled sample produced higher correlations and R^2 values than the smaller homogenous groups.

Multiple regression models using all measurements did not display significantly different SEE values compared with the linear models. In the discussion section of the article the authors make an interesting comment as to the scientific value of bare footprint evaluation in forensic identification. They state that crime scene footprints 'can be scientifically analyzed to establish the biological profile and confirm an association of an accused with the scene of crime' (page 4). This is a statement that insinuates scientific rigour in this area has already been established, encouraging the use of footprint evaluation for identification in criminal cases. The research presented in this thesis aspires to move this area of identification forward, but claims that identity can be established through the use of footprints alone are imprudent. The article scores level 4 using the OLE system for grading the quality of literature and 12 points out of 24 for validity using the OCPM system.

In a recent article by Pawar & Pawar (2012) prediction of heights of adolescents is suggested to have more accuracy when using feet rather than long bones such as the femur. They support this by arguing that the long bones ossify and achieve maturity later than the foot bones. However, full skeletal height would not have been reached for a proportion of this sample type and would skew results, therefore nullifying their argument relating to prediction accuracy.

Despite their justification for using an adolescent sample, they actually determined height estimations from an adult population of one hundred male subjects and one hundred female subjects. Inked footprints were taken of the left feet of each participant and the longest length measurement (from the base of the heel print to the tip of either the first or the second toe print) were recorded using a simple pen-and-ruler method. The width of the footprint was also recorded for each subject, but no data or subsequent analyses are reported for this measurement.

The authors report high correlations between the footprint lengths and height values for the sample, although for the female group a result of $r = 0.55$ ($p < 0.01$) is stated which may be better described as moderately correlated.

Regression equations are reported but it is difficult to compare these with other comparable studies, as supporting error estimates are omitted. The authors place more emphasis on the results of their division factor calculation, argued by Krishan et al. (2012) to be an inferior reflection of height prediction. There does not seem to be any references made to pertinent articles in this subject area of height prediction from footprint dimensions, apart from those by Robbins (1986) and Barker & Scheuer (1998). More recent articles have been overlooked which may have improved the discussion and conclusion sections of the publication, which in its present state, does not appear to add to the literature. The article scores 4 using the OLE system for grading the quality of the literature and an OCPM score of 6 for validity grading.

Table 6.1 below summarises the quality of the appraised papers by assessing validity and also determining whether appropriate supporting error estimates are offered (Table 6.1). The article by Fawzy & Kamal (2010) demonstrates the highest evidence of validity.

Table 6.1 Summary of critical appraisal of literature pertaining to estimation of stature from footprint dimensions

Author (date)	Appropriate sample used?*	Sample size and ethnicity	PPM Footprint length [a] v stature ($p < 0.01$)	Appropriate regression error estimates reported?	OLE	OCPM
Robbins (1986)	No	(M) 224 (F) 284 Ethnicity not described	(L/ All subjects) 0.84[†]	No	4	3
Oberoi et al. (2006)	No	(M) 100 (F) 100 Mangalore Indian	(R/M) 0.70 (R/F) 0.74	SEE (M) 46.60mm SEE (F) 45.28mm	4	15
Atamturk & Duyar 2008)	No	(M) 253 (F) 263 Turkish	(L/M) 0.71 (L/F) 0.68	SEE (combined M/F group) 51.42mm	4	11
Krishan (2008a)	No	(M) 1040 Gujjars	(L/M) Calc_A1 0.87 (R/M) Calc_A2 0.87	No	4	7
Fawzy & Kamal (2010)	No	(M) 50 Turkish	(L/M) Calc_A1 0.54 (R/M) Calc_A5 0.58	SEE (L/M) Calc_A1 36.30mm (R/M) Calc_A5 35.20mm	4	16

* In terms of age. Some articles include participants with still-growing feet and final stature undetermined (less than 19 years) as well as fully-grown feet at skeletal maturity (20 years and over), [a] Longest footprint measurement unless otherwise stated, [†] p-values not reported.

Table 6.1 continued

Author (date)	Appropriate sample used?*	Sample size and ethnicity	PPM Footprint length [a] v stature ($p < 0.01$)	Appropriate regression error estimates reported?	OLE	OCPM
Vidya et al. (2011)	Yes	(M) 100 (F) 100 South Indian	(R/M) 0.88 (R/F) 0.82	SEE (L/M) 16.96mm SEE (R/M) 14.54mm SEE (L/F) 11.01mm SEE (R/F) 9.05mm	4	9
Nataraja-moorthy et al. (2011)	No	(M) 42 (F) 65 Malaysian	(L/M) Calc_A1 0.73 (R/M) Calc_A1 0.75 L/F) 0.73 (R/F) Calc_A1 0.74	SEE (L/M) 49.56mm SEE (R/M) 48.21mm SEE (L/F) 35.31mm SEE (R/F) 34.21mm	4	8
Kanchan et al. (2012)	Yes	(M) 50 (F) 50 Mangalore Indian	(L/M) 0.63 (R/M) 0.63 (L/F) Calc_A5 0.45 (R/F) 0.53	SEE (L/M) 41.61mm SEE (R/M) Calc_A5 41.09mm, SEE (L/F) Calc_A2 55.60mm SEE (R/F) 52.87mm	4	12
Pawar & Pawar (2012)	Yes	(M) 50 (F) 50 Indian	(L/M) 0.94 (L/F) 0.55	No	4	6

* In terms of age. Some articles include participants with still-growing feet and final stature undetermined (less than 19 years) as well as fully-grown feet at skeletal maturity (20 years and over), [a] Longest footprint measurement unless otherwise stated, [†] p-values not reported

The researcher next conducted a predictive study estimating stature from footprint dimensions in an attempt to contribute to the validity of the measurement approach.

6.4 Methodology

The data used for this part of the study were derived from the same sample as described in section 4.3. All the participants received details of the study beforehand and ethical approval was obtained prior to recruitment (Appendix D.1).

The statures of the sixty one participants were taken using a SECA Leicester Portable Height Measure (SE001) described in section 4.4.

The right footprint of each volunteer was captured three times using an inkless paper system in both static and dynamic states. The method of footprint capture, digitisation and measurement approached used, is detailed in Chapter 3. Previously it was established that measurements from three footprints from the same subject in the static stance displayed little variance; the same was true for the prints taken from each subject in the dynamic stance (section 3.2). Therefore static and dynamic measurements from one print from each subject were selected for the analysis.

For correlation, all variables were tested for normality. K-S tests indicated the Calc_A5 measurement in the dynamic state deviated from normality, as was observed in section 4.5.1 in the initial data exploration. As previously described, further investigations as to the normality of this measurement's overall distribution indicated that Calc_A5 was indeed suitable for parametric testing. All other data involved in the predictive validity analysis displayed K-S test values of non-significance ($p > 0.05$), suggesting a normal distribution for all variables. Q-Q plots and histograms supported the test results. The data were therefore considered suitable for parametric analysis.

6.4.1 Statistical analysis

Relationships between variables of stature and footprint measurement were analysed using PPM correlation coefficients and interpreted according to Innes & Straker (1999), cited by Reneman et al. (2002) previously discussed in section 4.5.1. A stepwise multiple regression analysis was carried out to ascertain which of the measurements afford the most influence on the predictor

variable. These results were further analysed to determine R^2 values and standard error of estimates.

Multiple regression equations were formulated by following the equation suggested by Robson (2002):

$y = a + b_1 x_1 + b_2 x_2 + ...b_n x_n$, where

y = stature

a = constant

b_1 = regression coefficient for the first measurement variable, e.g. Calc_A1

x_1 = first measurement variable (mm)

b_2 = regression coefficient for the second measurement variable, e.g. Calc_A2

x_2 = second measurement variable (mm)

b_n = regression coefficient for the *n*th measurement variable

x_n = *n*th measurement variable (mm).

The best model was then determined according to the coefficient of determination (R^2) values or adjusted R^2 ($R^2_{adjusted}$). R^2 is the amount of variation in stature that can be explained by the footprint measurements. $R^2_{adjusted}$ is employed to compare regression models that contain different numbers of footprint measures (Sanli et al., 2005).

SEE was calculated for every regression analysis. This is used to quantify the accuracy of the prediction and is calculated from the residual scores (Jackson, 1989). On a scatter plot, it is the standard deviations of the dispersion of the actual *y* observations from the predicted *y*, as predicted by the linear regression equation. Numerically it can be described as the square root of the sum of the squared errors divided by n - 2 (Jackson, 1989);

$$SEE = \sqrt{\frac{\sum(Y-Y')^2}{N-2}}$$

Where,

SEE is the standard error of estimate,

Σ is the sum of,

Y-Y' is the paired difference of the height estimation – the estimate minus the actual height and

N is the sample number.

Assuming homoscedasticity is met, it can be said that 68% of the actual scores will vary ±1 SEE from the regression line (Y') and 95% of the actual scores (Y) to be within ±2 SEE of the regression line (Giles & Klepinger, 1988).

6.5 Results

Sample information regarding age, height and weight is shown in Table 4.1. Parametric analysis was supported as all variables were found to be normally distributed.

A summary of descriptive statistics for combined static and dynamic footprint measurements is illustrated in Table 6.2.

Table 6.2 Descriptive statistics for static and dynamic footprint measurements (n = 122)

Measurement	Min (mm)	Max (mm)	Mean (mm)	SD
Calc_A1	203.00	298.80	246.23	20.23
Calc_A2	199.80	296.10	243.22	19.20
Calc_A3	191.50	284.30	233.34	18.03
Calc_A4	186.80	266.90	220.26	16.82
Calc_A5	174.80	244.50	202.90	15.51
CalcWidth	38.70	63.00	49.28	5.25
MPJWidth	79.60	108.40	93.16	7.43
Footprint Angle	22.59	59.93	47.11	7.03
2-5 Toe Angle	33.19	57.05	45.52	5.67
1-5 Toe Angle	21.92	42.55	31.59	4.65
Dist. Met Angle	24.40	46.95	35.21	4.44
2-4 Base Angle	16.41	53.64	36.92	7.50

6.5.1 Correlations

PPM correlation coefficients of stature with the width, angle and length measurements of the scanned dynamic and static footprint images from sixty one subjects resulted in the widths and lengths displaying the highest correlations. Of these, the length measurement from the base of the heel to the tip of the smallest toe print (Calc_A5) displayed the highest correlation with stature (Table 6.3, N[a]).

Correlations for dynamic prints of the mixed group (males and females) resulted in the strongest significant associations with stature and length measurements. Of these measurements, Calc_A5 was the strongest correlation with stature (Table 6.3, N[b]).

Again, width and length measurements displayed the strongest correlations with stature when the data from the mixed group (males and females), static prints only, were analysed (Table 6.3, N[c]). However, in this instance, it was the length measurement from the base of the heel to the tip of the fourth toe print (Calc_A4) which displayed the highest correlation.

A smaller dataset (n = 30) analysing the values from the female group, static prints only, displayed weaker correlations than the previous sub-sets. The static print measurements in this set suggest the length measurement Calc_A4 to be the most strongly correlated with stature (Table 6.4, N[e]).

The dynamic print measurements for the female group suggest strong correlations between stature and the length measurements only; the highest correlation occurring between stature and Calc_A4 measurement (Table 6.4, N[d]).

Static footprint measurements did not display as strong a correlation with stature for the male group, compared with other sub-sets (Table 6.5, N[g]). The Calc-A5 length measurement was unobtainable for five prints examined, due to an absence of a 5th toe print, reported as missing values in the dataset. This lowered the sample size for this particular length measurement to n = 26.

Again, missing values were noted in Calc_A5 length measurements of dynamic footprints for the male group (n = 28). However, this length measurement

displayed the highest correlation with stature but the correlations overall were smaller than the mixed sub-sets (Table 6.5, N[f]).

Table 6.3 Stature and footprint measurement correlations (combined M/F)

	Height of subject					
	r	N[a]	r	N[b]	r	N[c]
Calc_A1	0.71**	122	0.80**	61	0.76**	61
Calc_A2	0.74**	122	0.80**	61	0.77**	61
Calc_A3	0.76**	122	0.82**	61	0.78**	61
Calc_A4	0.78**	122	0.83**	61	0.79**	61
Calc_A5	0.78**	113	0.86**	58	0.78**	55
CalcWidth	0.60**	122	0.60*	61	0.61**	61
MPJWidth	0.67**	122	0.66*	61	0.68**	61
Footprint Angle	0.14	120	0.13	60	0.15	60
2-5 Toe Angle	-0.01	115	-0.06	58	0.05	57
1-5 Toe Angle	-0.03	113	-0.17	58	0.12	55
Dist. Met Angle	-.016*	122	-0.20	61	-0.06	61
2-4 Base Angle	-0.07	122	-0.08	61	-0.06	61

**. Correlation is significant at the 0.01 level (one-tailed).

*. Correlation is significant at the 0.05 level (one-tailed).

N[a] Correlations of stature with both static and dynamic measurements (males and females)

N[b] Correlations of stature with dynamic measurements only (males and females)

N[c] Correlations of stature with static measurements only (males and females)

Table 6.4 Stature and footprint measurement correlations (female footprints)

	Height of subject			
	r	N^d	r	N^e
Calc_A1	0.67**	30	0.41*	30
Calc_A2	0.58**	30	0.41*	30
Calc_A3	0.63**	30	0.46**	30
Calc_A4	0.73**	30	0.55**	30
Calc_A5	0.72**	30	0.48**	29
CalcWidth	0.34*	30	0.22	30
MPJWidth	0.06	30	0.24	30
Footprint Angle	0.22	30	0.24	30
2-5 Toe Angle	-0.03	30	0.06	30
1-5 Toe Angle	-0.16	30	-0.06	29
Dist. Met Angle	-0.15	30	-0.12	30
2-4 Base Angle	-0.16	30	-0.13	30

**. Correlation is significant at the 0.01 level (one-tailed)

*Correlation is significant at the 0.05 level (one-tailed)

N^d Correlations of stature with dynamic measurements (females)

N^e Correlations of stature with static measurements (females)

Table 6.5 Stature and footprint measurement correlations (male footprints)

	Height of subject			
	r	N^f	r	N^g
Calc_A1	0.46**	31	0.45**	31
Calc_A2	0.55**	31	0.54**	31
Calc_A3	0.55**	31	0.54**	31
Calc_A4	0.52**	31	0.51**	31
Calc_A5	0.66**	28	0.50**	26
CalcWidth	0.15	31	0.11	31
MPJWidth	0.36*	31	0.34*	31
Footprint Angle	0.19	30	0.21	30
2-5 Toe Angle	-0.06	28	-0.10	27
1-5 Toe Angle	-0.23	28	-0.13	26
Dist. Met Angle	-0.21	31	-0.12	31
2-4 Base Angle	-0.06	31	-0.07	31

**. Correlation is significant at the 0.01 level (one-tailed).

*Correlation is significant at the 0.05 level (one-tailed)

N^f Correlations of stature with dynamic measurements (males).

N^g Correlations of stature with static measurements (males).

The strongest correlations were between stature and the length and width measurements. The angle measurements displayed the poorest correlations and were not included for further analysis in this part of the validity study.

Means were calculated for the combined length measurements Calc_A1;Calc_A2 and also Calc_A4;Calc_A5 and subsequently used to describe the combined values of the medial and lateral borders of the footprints respectively. These were then correlated with stature values. The results

illustrated in Table 6.6 show that the lateral borders were more highly correlated with the statures of the volunteers than the medial borders.

Table 6.6 PPM correlation coefficients between stature and medial/lateral borders of footprints

	Medial border (Calc_A1 & Calc_A2)	Lateral border (Calc_A4 & Calc_A5)
Dynamic footprints	0.81**	0.85**
Static footprints	0.77**	0.79**
Static and dynamic footprints	0.73**	0.79**

** Correlation is significant at the 0.01 level (one-tailed)

6.5.2 Regression Analysis

A multiple stepwise regression analysis was performed on the dataset including static and dynamic prints for all subjects. The large dataset (n = 122) was chosen for this particular analysis as data cases below 100 do not correctly reflect the overall fit nor the behaviour of the individual predictors of the model (Green, 1991). The factors age, weight, width of the calcaneum, width of the metatarsophalangeal area, length measurements from the base of the heel to the tips of the five toe prints and the five angle measurements were included for the multiple regression analysis.

The resultant hierarchical regression analysis demonstrated that the measurement Calc_A5 was the strongest predictor variable for stature for the sample, Calc_A4 measurement second and finally the Calc_A1 measurement. Using this approach, all other variables were excluded. The model summary resulted in an R^2 value of 0.608 ($R^2_{adjusted}$ 0.607) for the length measurement Calc_A5.

A further analysis was undertaken in which linear regression equations were formulated in order to predict statures, using data from static and also dynamic linear measurements.

Table 6.7 Linear regression equations for estimating stature from static width and length footprint measurements (mm) n = 61

Regression equations	SEE	R^2
4.08 x Calc_A1 + 740.7	58.9	0.57**
4.17 x Calc_A2 + 710.7	59.8	0.59**
4.47 x Calc_A3 + 684.3	58.1	0.61**
4.76 x Calc_A4 + 676.7	57.8	0.62**
5.53 x Calc_A5 + 643.1	60.4	0.60**
10.84 x CalcWidth + 1172.7	74.3	0.37**
8.37 x MPJWidth + 92.34	68.4	0.46**

SEE Standard error of estimate, R^2 Coefficient of determination, **p < 0.01

Table 6.8 Linear regression equations for estimating stature from dynamic width and length footprint measurements (mm) n = 61

Regression equations	SEE	R^2
3.84 x Calc_A1 + 725.4	56.0	0.64**
3.83 x Calc_A2 + 748.0	56.0	0.64**
4.16 x Calc_A3 + 710.3	53.6	0.67**
4.49 x Calc_A4 + 693.8	52.1	0.69**
4.70 x Calc_A5 + 747.0	41.6	0.74**
10.48 x CalcWidth + 1182.3	74.6	0.36**
8.28 x MPJWidth + 930.6	70.2	0.44**

SEE Standard error of estimate, R^2 Coefficient of determination, **p < 0.01

Table 6.7 illustrates the measurement Calc_A4 was the better indicator for stature estimation than the other linear measurements for the static footprints. This equation demonstrated the lowest SEE for the static footprints (57.80mm)

and the highest R^2 value (0.62). In the dynamic footprints, the regression equation for the Calc_A5 measurement was the best indicator for stature; SEE 41.6mm, R^2 0.74 ($p < 0.01$) as illustrated in Table 6.8.

6.6 Discussion

Anthropological papers investigating the estimation of stature from foot dimensions argue for sample populations to be separated into their ethnic groups before applying meaningful analyses (Krishan, 2008b; Sen and Ghosh, 2008). Carrying out such studies in this way permits regression equations for the prediction of stature with the smallest of error margins for each ethnic population (Krishan 2008a). These types of studies are invaluable for anthropological research as they offer useful information regarding these specific ethnic populations.

However, the reality in the field of forensic identification is that it is fraught with the *unknown* rather than the *known*. A two-dimensional footprint impression left at a scene of crime cannot presently inform the observer with any degree of certainty if the print belonged to a male or female, whether the impression was left whilst the person was walking or standing, how old the person was or of their ethnic background. Since the research presented in this thesis hinges on the evaluation of a footprint measurement approach for forensic uses and not necessarily for clinical or anthropological applications, it was considered that data from the heterogeneous sample used for this predictive validity study, allowed for real-world inferences of multi-racial countries such as the UK or the US.

Out of the thirteen footprint dimension measurements, the length measurement Calc_A5 had the strongest correlation with stature for the combined group of males, females, static and dynamic footprints ($r = 0.782$, $p < 0.01$). The strength of this correlation increased when the combined group was analysed using dynamic prints only ($r = 0.858$, $p < 0.01$) for this particular footprint length measurement. Given that crime scenes reflect activity-based events, it may be more likely that dynamic prints are more commonly found at a crime scene than static prints; however this has not been documented.

When the sample was split into homogenous groupings of males, females, static prints and dynamic prints, correlations were strongest for the Calc_A4 and Calc_A5 measurement for the static prints ($r = 0.786$, $p < 0.01$) and had an even stronger correlation in the dynamic measures ($r = 0.858$, $p < 0.01$). The mean:SD proportions for Calc_A4 and Calc_A5 lengths were as expected (Table 6.2). This suggests that the lateral border of the footprint is a more stable measure than the medial border in the prediction of height. Combining Calc_A1 and Calc_A2 lengths and Calc_A4 and Calc_A5 lengths further demonstrated the differences of correlations with stature and medial/lateral edges of the footprints. The results exhibited a lower correlation on the medial border than the lateral border (Table 6.6). However, the grouping males/static prints seemed not to adhere to the Calc_A4/A5 measurement trend of strong correlation with stature; in this grouping's case the measurement Calc_A2 was the strongest correlate. For this group of $n = 31$ there were five missing values for the Calc-A5 length measurement. This occurred because the fifth toe of these participants did not make contact with the ground during the process of the development of the static footprints. Although this occurred also in the female group (one missing value) in the formation of the static footprints, the missing values were more significant in the male group where the missing fifth toe footprints accounted for 16.13% of the total footprints. Kulthanan et al. (2004) studied the footprints of athletes and non-athletes and found that 25.3% of male non-athletes and 18.5% of male athletes did not make contact with the ground regarding the fifth toe. In Moorthy et al.'s study of one hundred and forty non-athletes and one hundred and fifty athletes from Malaysia, 16% of male non-athletes and 3.2% of male athletes did not make contact with the ground with their fifth toes (Moorthy et al., 2011). Hughes et al. (1990) studied the electronic footprints of eighty male and eighty female subjects ranging from five years to seventy eight years of age and noted that 8% of the male footprints had toe prints which did not make contact with the ground, observing that this occurred in 'usually the fifth' (page 248). Previous work has suggested the female pelvic girdle tilts forwards in the sagittal plane, as opposed to the male pelvic girdle which tilts backwards (Falls, 1986; Van De Graff, 1988). A study by Opila found significant differences between males and females when examining subjects' centre of gravity whilst standing barefoot (Opila, 1988). Opila found that the line

of gravity passed posteriorly to the greater trochanter in females and anteriorly in males. Thus it can be postulated that males stand naturally with their bodyweight directed over the distal aspect of their feet rather than over their hindfeet. Weight over this position may allow the intrinsic extensors of the toes to fire prematurely in anticipation of gait initiation, contracting the extensor tendons of the lesser toes in preparation for toe-off (Mann & Hagy, 1979; Hughes et al., 1990). The first toe remains largely stable due to stronger extensor and flexor tendons around the interphalangeal joint and a decreased amount of type II collagen in the extensors of this toe prohibiting flexibility, compared with the lesser toes (Milz et al., 1998).

The multifactorial regression analysis including all variables including weight, footprint measurements, age, etc., suggested the regression equation was weighted by three footprint measurement factors; Calc_A5, Calc_A4 and Calc_A1. Footprint evidence retrieved at a crime scene is devoid of additional factors for the completion of a multiple regression equation (e.g. tibial length, ethnicity) and therefore it is fortuitous that the best predictors of stature determined in this study do in fact appear to relate to measurements which can be extracted directly from the footprint itself.

In formulating regression equations, it is good practice to state standard error estimates as this allows an interpretation of the expected margin of error (Jackson, 1989). A SEE of 41.66mm derived from the formulated regression equation (stature mm = 4.697 x Calc_A5 Length mm + 746.96), is comparable with SEEs resulting from other height estimation studies. Sanli et al. (2005), for example, reported a SEE of 44.50mm using the longest footprint length measurement for their combined male and female sample (Sanli et al., 2005).

Regressive equations were derived from the combined sample which included both male and female footprints. Kanchan et al. (2008) reported that the predictive value of actual foot dimensions and correlation coefficients are not affected by sex; stature estimation remained accurate even when sex was unknown.

The Calc_A5 length measurement proved to be the best overall predictor for the stature variable and displayed an associated coefficient of determination of 0.608 (adjusted R^2 0.607). In other words, 61% of the variation of stature is

inherent of the Calc_A5 length measurement. This is comparable with Özaslan et al.'s study which suggested R^2 for leg length for males was 0.55 and for females, 0.63. In other words, 55% (males) and 63% (females) of variation in height was influenced by tibial length in their study's sample (Özaslan et al., 2003).

Studies that examine the prediction of height from footprints more often consider the longest foot length measurement; from the base of the heel to the tip of either the first toe or the second toe (Oberoi et al, 2006; Atamturk & Duyar, 2008; Natarajamoorthy et al., 2011; Moorthy et al., 2011; Vidya et al., 2011; Pawar & Pawar, 2012). Exceptions to this design were the studies carried out by Krishan (2008c), Fawzy & Kamal (2010) and Kanchan et al. (2012) which determine the estimation of stature by including all five heel-to-toe print measurements. The Calc_A5 measurement displayed the highest correlation with height for the female left footprints in Kanchan et al.'s recent study from 2012. Fawzy & Kamal (2010) also concluded that the Calc_A5 length was the best predictor of height, concurring with the study presented here. An explanation of why this particular footprint length is the best predictor of height in the present study and in the other two publications can perhaps be explained by the structure of the lateral longitudinal arch in the foot. This arch, as opposed to the medial longitudinal arch (MLA) is far more stable and also has fewer articulations (Cunningham & Romanes, 1976). The measurement from the base of the heel to the big toe has to contend with the variation of the MLA, which not only 'gives' more than the lateral border during gait due to tendon laxity, its variability is also subject to genetic, ethnic, weight and age factors (Saltzman & Nawoczenski, 1995; Dowling et al., 2001; Thompson & Zipfel, 2005).

The arch of the foot is supported by the plantar fascia, extending from the insertions at the proximal phalanges to the medial tubercle of the calcaneus (Erdemir et al. 2004). This thick band of connective tissue is thought to contribute to changes in the structure of the foot during the stance phase of gait (Hicks, 1954; Thordarson et al., 1997; Sharkey et al., 1999). It is during this phase of walking that the inked print from the plantar surface of the foot is created. Stance phase consists of three stages: contact, midstance and propulsive (Merriman & Tollafield, 2002). The contact stage of stance phase involves the contact of the heel with the ground followed by the rest of the foot

which pronates to allow shock absorption. During this part of stance phase, tibial internal rotation lowers the MLA to enable body balance and absorb shock (Saltzman & Nawoczenski, 1995). The foot then supinates during midstance in which there is total contact of the foot with the ground, followed by propulsion in which the heel and finally the hallux leave the ground (Merriman & Tollafield, 2001). At this stage, the tibia externally rotates and the arch rises. Nester (1997) observed that subjects displaying subtalar joint pronation, indicated by calcaneal eversion, had lower medial longitudinal arches: conversely subjects displaying subtalar supination had raised arches. Thus, the MLA is subject to variation from both extraneous and internal factors.

Krishan's study examining the impact of different loads on the footprint suggest that for the right foot, the length measurement from the base of the heel to the tip of the large toe displayed significant differences in length between normal weight bearing and carrying a 20kg load ($t = 2.51$) (Krishan, 2008c). This is in contrast with the measurement from the base of the heel to the tip of the small toe which suggested non-significant changes ($t = 2.21$). Descriptive statistics suggest there was a mean difference of 0.08cm between the normal and 20kg load bearing states for the longest toe length, whereas there was a mean difference of 0.04cm between the two load-bearing states and the smallest toe length, reflecting perhaps on the role and elasticity of the MLA.

The findings from Krishan's study (Krishan, 2008a) analysing the associations of all five footprint length measurements with stature, did not concur with the findings from the predictive study presented in this thesis, Kanchan et al.'s 2012 study and that of Fawzy & Kamal (2010). Upon further personal enquiry, Krishan explained that he found that 15% of his subjects (males) failed to make contact with the ground with their fifth toes whilst standing (Krishan, 2011). This concurs with the observations regarding non-contact of the fifth toe from previous studies (Hughes et al., 1990; Kulthanan et al., 2004; Moorthy et al., 2011). However, Krishan simply excluded these missing values from the investigation which most likely skewed the final regression analysis of this study. This highlights the importance of recognising and explaining missing values of this nature.

In order to investigate the prowess of the predictive equation calculated for the sample used in the present study, a further analysis was applied. Employing the equation 4.697 x Calc_A5 + 746.96 derived from the linear regression analysis of the dynamic measurements (Table 6.8), a scatterplot of the correlations with actual height was created (Fig. 6.1).

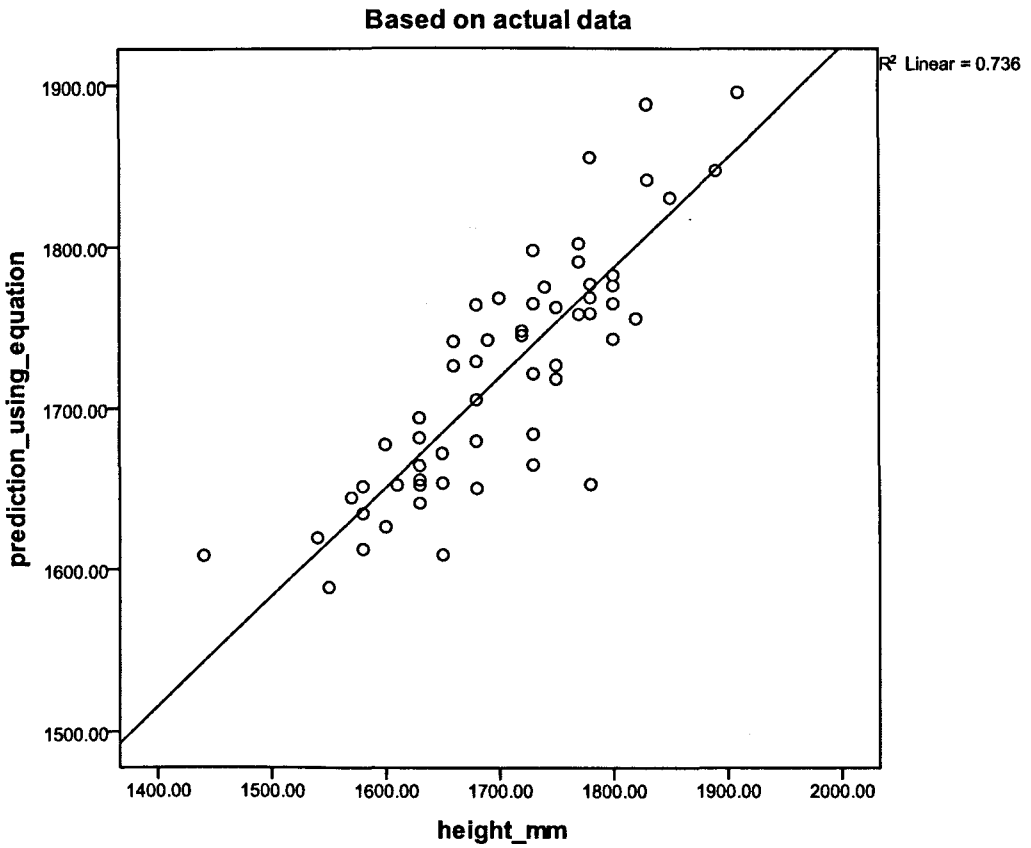


Figure 6.1 Scatterplot illustrating correlations of predicted and actual height values of the male and female subjects using Calc_A5 measurement from dynamic footprints

The coefficient of determination (R^2) derived from the graph (Figure 6.1) suggests that 74% of variability within the dataset can be explained by the regression model. It is a reflection of how likely the predictive equation will hold true for other samples.

In a further more stringent analysis, the same dataset was reduced randomly by 50% to lessen the association between the predictive equation and the sample it was derived from. SPSS software randomly chose sixteen female and

seventeen male subjects from the sample. The predicted and actual height values were correlated and R^2 calculated (Figure 6.2).

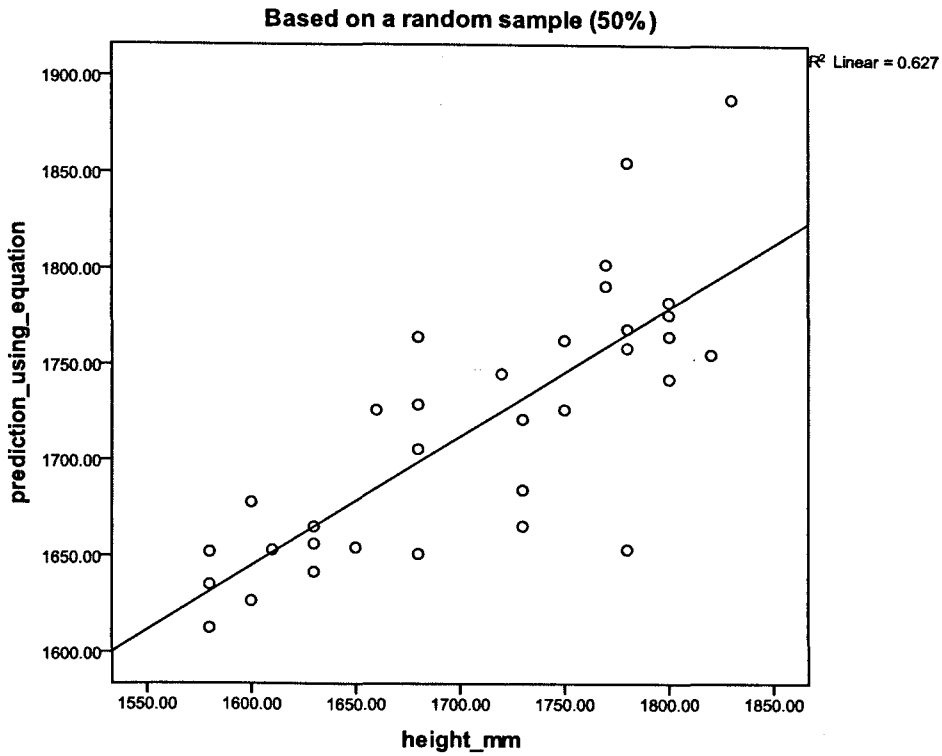


Figure 6.2 Scatterplot illustrating correlations of predicted and actual height values from a random sample (50% of original sample)

This smaller sample also presents a positive correlation between the predicted and the actual heights of the randomly selected subjects. The resultant R^2 calculation indicates that there is a 63% confidence that the prediction is proficient in determining similar results in other samples. However, this has yet to be tested in the field. Further studies involving larger, more diverse populations are required to understand the prediction fully.

6.7 Conclusions

Multiple regression and bivariate correlation analyses have shown that the footprints from the sample included in this part of the research can predict height in the absence of other factors such as age, tibial length or ethnicity. The length measurements Calc_A5 and Calc_A4 are the best predictor variables for

stature. Although differences in the bivariate correlations of measurements for dynamic and static footprints with stature were noted, stronger correlations for the lateral borders of both types of prints compared with the medial borders were apparent. The calculation of regression equations was complimentary alongside correlation analyses between the static and dynamic footprints. Knowledge of the functioning foot suggests the lateral border is less variable in nature and therefore a more stable indicator in the estimation of height from footprint dimensions. The use of regression formulae to predict the stature of a person from their footprint alone is important for forensic identification purposes, especially accompanied by calculated error margins in the form of 95% SEE. The present study confirmed a small SEE of 41.66mm derived from the formulated regression equation from dynamic footprints using the Calc_A5 measurement, compared with SEEs resulting from published height estimation studies using the longest footprint length measurement variable. In practice this implies stature of the relevant population can be predicted with a total error margin of 81.95mm 95% of the time, equivalent to just less than the length of a credit card. It also reports the largest R^2 value reflecting on the Calc_A5 measurement's influence over the sample's height. The study by Fawzy & Kamal (2010) investigating footprint length measurements and height achieves a smaller SEE than the present study (35.20mm) and also use a relatively small sample compared with other similar studies. They too discovered Calc_A5 footprint length measurement to have the greatest correlations, R^2 values and smallest SEEs when associated with the stature variable. The research presented here is the only study to date, to estimate stature from both static and dynamic footprints. The value of predicting stature from a footprint impression at a crime scene has previously been questioned; however, if the calculations produced are associated with such small error estimates, this serves as an additional component of evidence in the process of identification.

The interpretation of analyses presented by this chapter confirms the predictive validity of the measurement approach developed for the thesis, thereby offering criterion-related evidence for the validity of the design as a whole.

Chapter 7
Establishing Evidence of Reliability

Previous chapters of this thesis have explored and established the extent of various types of validity regarding the Reel method of footprint measurement. This chapter will now seek to determine the extent of reliability of the new approach. It will include a discussion of reliability issues regarding similar studies that consider the analysis of data consisting of continuous measurement variables. The variety of statistical tests exemplified in the literature that demonstrate the existence of reliability within a particular setting, will be examined. From this, the most suitable form of testing, appropriate for estimating the reliability of the Reel method, will be identified and utilised in a repeated-measures study design.

7.1 Introduction

Error or variance in measurement is often unavoidable but good measurement tools would account for the extent of error. Forensic science, like all other forms of science, is duty bound to establish the extent of measurement error in the tools it uses. However the forensic literature in the area of study does not demonstrate scientific rigour.

In the scientific context, reliability is considered an essential characteristic of any technique or measuring instrument, and influences the validity of the technique (Baumgartner, 1989). This is because for a test or measuring technique to be valid, it must also be reliable. In simple terms, reliability refers to the reproducibility of scores obtained from a measure. It is the degree to which a measurement technique yields the same result when scores are taken from two or more operators (inter-rater reliability) or on at least two different occasions by the same operator, known as intra-rater reliability (Michels, 1985). But as previously discussed, the term 'reliability' in law-driven policy and in the forensic science disciplines, is often confused with the subjective terms 'trustworthy' and 'relevancy' (Chapter 2). Point three of the law-driven Daubert ruling (Daubert v Merrell Dow Pharmaceuticals Inc., 1992) is perhaps the most

pertinent acknowledgment of the scientific meaning of reliability in the forensic sciences. This directive advocates that the test or technique being used in the investigation of identification is able to relay error rates attributable to the methodology concerned (Faigman et al., 2005). The term 'error' in law and in the forensic individualisation sciences however, is viewed contrary to the scientific definition. Here 'error' is regarded as an incorrect result. For example, in the case of US v Trala in which the admissibility of PCR- (polymerase chain reaction) based DNA was considered, error was explained in court as the extent to which a methodology has not been followed (US v Trala, 2001). Further to this, it was explained to the judge that if properly calibrated instruments are used, the resultant error rate will be zero, implying that in practice, if an operator applies the correct methodology, no error will be incurred. In scientific practice, an amount of error is always assumed, thus nullifying the notion of zero error, and therefore reliability is defined by the amount of acceptable measurement error for the tool to be effective (Atkinson & Nevill, 1998). In a case more akin to bare footprint evaluation, issues arose considering the admissibility of footwear identification techniques (US v Allen, 2002). In this Daubert hearing, the expert testified that a zero error rate resulted from the process of the evaluation. It was inferred by this statement that based upon scientific foundations, the shoe either does or does not make the impression in question. The expert further implied that any error that does occur is 'caused by examiner error in the application of the process or by examiner error in reaching a particular conclusion' (US v Allen, 2002, page 862). This is a rejection of one of the fundamental principles of reliability concepts in test measurement; the acceptance of random error.

The US National Academy of Sciences (NAS) report recognised that although appropriate standards for the competency to carry out calibrations and/or tests have been specified by, for example, International Standards Organisation 17025 in order to establish error rates, these are rarely correctly used or reported in the forensic identification sciences. The NAS report identified that this problem regarding reliability in forensic science is furthered by a paucity of scientific empirical research to validate the basic techniques and principles in the individualisation disciplines, including footwear and bare footprint evaluation. The report addresses this by way of recommendation three: 'Studies should establish the limits of reliability and accuracy that analytic methods can

be expected to achieve as the conditions of forensic evidence vary' (National Research Council, 2009, page 23). Government initiatives and reports from the UK have also criticised the interpretation of scientific reliability in a court of law. The UK Law Commission report of 2009 states, 'the trial judge has been provided with no guidance whatsoever to assist him or her in the determination of evidentiary reliability' (page 22), leaving them to simply guess whether scientific evidence should or should not be admissible (The Law Commission, 2009). The report continues by recommending the adoption of a Daubert-type ruling to objectively assess whether a scientific technique or technology is reliable enough to merit admission in court.

It is of no surprise therefore, the apparently low numbers of articles available that discuss reliability estimates associated with forensic identification techniques. In the critical review of the general literature (Chapter 2), problems concerning the reliability of techniques and tests used in the individualisation forensic sciences were discussed briefly. This next section will expand upon these difficulties, identified from the existing literature.

7.2 Literature review

A literature search was carried out to gather publications specifically in the area of the evaluation of two-dimensional footprint impression evidence in forensic investigation, using the key terms 'footprint*', 'forensic*', and 'measure*'. Databases MEDLINE, AMED, CINAHL, PsycINFO, SciVerse and WestLaw were utilised as well as 'grey' literature searching within other forensic publications. Twenty six articles were found using these three key terms; of these only two papers discussed the concept of reliability and are included in the next section in which the pertinent literature is reviewed. Not wishing to be limited solely within the forensic field, the search was widened to include clinical footprint measurement studies that assessed reliability estimates. This increased the number of relevant articles worthy of further critical appraisal to seven.

7.2.1 A review of footprint identification and footprint clinical literature in terms of reliability

The literature detailing bare footprint measurement was appraised in terms of strength of reliability analysis for either the measurement method employed and/or between-footprint reliability.

The footprint angle or Clarke's arch angle was first described by Harrison Clarke (Clarke, 1933). This preliminary investigation of a new approach to evaluate arch height from footprints, examined the extent of intra-rater reliability of repeated measures from footprints of one hundred and thirty five footprints. Clarke reported high reliability for the angle (0.971) but, typically from an article of this era, PPM correlation coefficients were chosen for these calculations, previously argued to be an inappropriate test for this purpose (section 4.2.2).

Hawes et al. examined intra-rater reliability between five footprint parameters (arch index, footprint index, arch angle, arch length index and truncated arch index) on two separate occasions (Hawes et al., 1992). All measurements were considered to be highly reliable as they displayed coefficients ranging from 0.91 to 0.99. Unfortunately the PPM correlation coefficient was used for these calculations, which is a measure of association and does not appropriately explain reliability (section 4.2.2).

In a similar study by Cavanagh & Rodgers, test-retest intra-rater reliability was explored in an attempt to further validate a method for measuring footprints termed the Arch Index (Cavanagh & Rodgers, 1987). The within-day reliability estimate of the AI resulted in a coefficient of 0.96 and the between-day value was 0.94. Again, the PPM correlation coefficient was employed in these calculations, which is not representative of the extent of reliability for this investigation.

Kippen (1993) set out to establish reliability of seven dynamic and seven static footprints from one subject. In this study, one line was manually drawn from the heel to the centre of the third toe of one selected static inked footprint and measured with a ruler. Following a repeated measures design, the study recorded the measurements three times by five independent raters. The results were statistically analysed and an analysis of variance was employed. High reliability between the measurements were reported by the author although

recommended reliability coefficients were not displayed; the more descriptive yet less informative standard deviation was used to express inter-rater reliability (SD ±0.27mm) and intra-rater reliability (SD ±0.01mm).

Barker & Scheuer (1998) reported their investigations of the predictive nature of inked footprints. As part of this study, the reliability of their chosen measurement approach was also examined but descriptive statistics only are detailed, in the form of standard errors and standard deviations. Their measurement method involved the construction of a linear axis, based on Robbins' method (Robbins, 1985). Created from this axis, heel-arch-ball length, big toe length and the widths of the forefoot and heel measurements were constructed and measured manually (ruler and pen method) by three independent raters. The reported SD and SE were derived from combined measurements rather than a more in-depth analysis of differences between measurements. Bland & Altman (1996) advocate the use of the SEM to express the repeatability of measurements, rather than the SE alone.

Mathieson et al. (1999) report reliability of a measurement method used on their investigation using electronic footprints. Their measurements included the Stahelis Arch Index, the Footprint Angle and the Chippaux-Smirak Index. Their use of the PPM correlation coefficient for the purpose of determining reliability has been heavily criticised as previously discussed. No other reliability tests were discussed.

Kennedy et al.'s extensive study exploring the uniqueness of the human footprint utilised a two-way analysis of variance to investigate the between-print reliability of approximately two hundred measurements between the footprints from each of one hundred and thirty four subjects (Kennedy et al., 2005). However, results of the ANOVA were not reported. The ANOVA used in isolation for examining the extent of reliability, is not recommended because, as with paired t-tests, the detection of systematic error is influenced by a large random (residual) variation (Altman, 1991; Bland & Altman 1996; Atkinson & Nevill, 1998).

Kennedy et al. employed the coefficient of variance (described as the 'standard deviation ratio', page 1074) to report individual footprint measurement variation within the sample (between-print reliability). Their conclusion that there is a one

in 1.27 billion chance of sharing the same footprint dimensions with another is based on the % CV calculation from the between-print reliability study. The % CV is reliant upon the data displaying heteroscedasticity for it to provide meaningful results. The test makes the assumption that the widest test-retest variation will occur in the variable achieving the highest measurement values, for example heel to first toe length (Atkinson & Nevill, 1998). In the case of footprint evaluation, wide variation values for the longer measurements would not necessarily be demonstrated in a reliability analysis compared with variance limits for the smaller measurements. % CV is a statistical method that should be applied to data in which the magnitude of the measured values is an essential factor when assessing the degree of agreement between tests, unlike the SEM which is not dependent on this factor. A difficulty arising from this statistical method is that *x* % of the smallest measurement value will be considerably different to *x* % of the largest measurement value (Bland, 1987). Others are also critical of the use of the % CV for evaluating reliability (Chinn, 1991; Rankin & Stokes, 1998; Bruton et al., 2000).

It is evident that reliability estimates are not explored fully regarding the footprint related articles discussed in this section. A further investigation of the differing approaches of reliability in terms of statistical methods and study design was deemed necessary, by way of an additional literature review outside the realms of footprint measurement.

7.2.2 Critical review of articles pertaining to the reliability of clinical measuring tools

An initial search of the literature was carried out using the databases CINAHL, AMED, and MEDLINE using the following terms; 'rater reliability', 'measure*' and 'quantitativ*'. These key terms were chosen to reflect the type of evaluation considered suitable for establishing reliability of the new footprint measurement approach and elicited one hundred and forty seven results. The search was narrowed further by excluding non-English language and duplicate articles. Articles of most interest were those pertaining to human clinical measurement, particularly those assessing a new technique or measurement test. Measurement tools assessed in the remaining twenty four studies included those in the field of goniometry, sphygmomanometry and photoplethysmography, electromyography, gait analysis systems, radiography,

and echocardiography. The eight articles finally chosen for review in this section were those that would sustain a comprehensive discussion pertaining to choice of analysis from a variety of statistical methods and measurement tools. This would provide supporting evidence as to the choice of reliability test(s) required to examine the extent of reliability in the Reel method.

Clinical measurement involving assessment of joint range of motion proffered many articles in terms of reliability analyses. Joint range of motion is commonly measured using goniometry, considered a widely accepted technique for this type of measurement (Norkin & White, 2003). There are several devices employed for measuring range of motion including digital and manual goniometers. Limitations to the inter-rater and intra-rater reliability of these tools has been previously questioned by researchers, due to device and rater/recorder error and also participant and environmental variations (Carey et al., 2010). The low cost and ease of use, however, ensures popularity within the clinical community. Therefore many of the articles pertaining to goniometry also investigate criterion-related validity to evaluate the instrument alongside a gold standard method, as well as reliability of the goniometer in question. For example, Carey et al. (2010) compared a newly developed digital goniometer prototype with the gold standard universal goniometer. Five physiotherapists each recorded five shoulder and elbow measurements using both devices on eighteen patient models. An inter-rater reliability analysis utilising the ICC was conducted to evaluate the performance of the operators' use of the device. Correlation coefficients such as the ICC determine relative reliability of a test, previously discussed in section 5.5.1. Carey et al. reported coefficients ranging from 0.41 to 0.60. The authors used Landis & Koch's benchmark system for rating reliability values (Landis & Koch, 1977). They commented that the inter-rater reliability results indicated a 'moderate to substantial level of reliability' (page 64); however, Landis & Koch's ICC rating system was designed for nominal data only and therefore this interpretation may have its limitations for Carey et al.'s study. Chinn (1991) recommends that an ICC of at least 0.60 for a measure can be considered useful and Chiu & Sing (2002) recommend any value 0.69 and below as exhibiting poor reliability, contrary to Landis & Koch who argue 0.21 to 0.40 is indicative of fair reliability and 0.41 to 0.60 indicates moderate reliability. No other supporting statistics were used for the reliability

analysis of this study, criticised by Atkinson & Neville who emphasise the difficulties of interpreting ICCs and therefore do not recommend their use in isolation (Atkinson & Nevill, 1998). Rankin & Stokes further explain that the ICC is difficult to interpret as it gives no indication of the amount of disagreement between measurements and also recommend the use of alternative statistics to enhance the reliability interpretation (Rankin & Stokes, 1998). Another problem concerning the use of the ICC in isolation is that the calculation reflects upon heterogeneity of variance, interpreted as the ratio of true score variance (between-subjects variance) to true score variance plus error. Reliability will always offer high estimates if the true score variance is sufficiently high, and vice-versa (Rankin & Stokes, 1998). Problems relating to the sole use of the ICC are also evident in Stone et al.'s study in which the inter-rater reliability of hand-held dynamometry was examined when measuring knee extensor strength in cancer patients (Stone et al., 2011). Here, the authors explained that although high ICCs were obtained for the inter-rater reliability assessment, the limits of agreement displayed large values, reflecting poor reliability. The authors suggested that this was as a result of the subjects' inadequate tester strength, which compromised reliability.

Carey et al. then sought to establish criterion-related validity of the technique used in their experiment. The authors correctly employed PPM correlation coefficients for the interpretation yet incorrectly described this part of the investigation as an 'intra-rater reliability analysis' (Carey et al. 2010, page 59). This study used eighteen 'patient models' and explained that the subjects were not physiotherapists or students of physiotherapy (page 56). The five therapists of varying experience employed for the study as operators, were 'blinded' to the results which were subsequently recorded by physiotherapy students. These two points are important as they allow for an interpretation for general application as bias has actively been reduced for the reliability study.

In another investigation of goniometry, joint range of motions of children presenting with congenital muscular torticollis were examined in a test-retest reliability design in which measurements were taken for a group of twenty three infants and measurements repeated one hour later (Klackenberg et al., 2005). Intra-rater reliability between the two sets of measurements was first analysed to determine any significant mean differences between the two sets employing

ANOVA, which would have reflected systematic bias. Using the between-subjects and within-subjects mean squares from the ANOVA, ICCs were calculated. Because of the problems identified with the ICC as previously discussed in this section, Klackenberg et al. recommended the additional use of the SEM as it remains unaffected by the range of subjects' measurements. The SEM examines the average error to expect in an individual rater's recorded measurement(s), estimated by using the measurements of a group. It assesses absolute, rather than relative reliability, focusing on differences that arise between repeated tests of same measurement (Baumgartner, 1989). The standard deviation of differences was also used in this study to determine the reproducibility of the results for future clinical use in treatment outcome. This is a statistical test prevalent in clinical treatment test-retest designs, first described by Bland & Altman (1986). The study involved measurements taken by one experienced physiotherapist and results suggested that the measurement method employed proved to be acceptable for following treatment effects over time.

In a study investigating the reliability of measures of hammer toe deformity and tibial torsion, goniometry and a three-dimensional digitiser were independently tested for extent of reliability and then compared with measurement results from computed tomography (considered the gold standard) in a further investigation of validity (Kwon et al., 2009). For their intra-rater reliability examination, the authors used ICC and SEM reliability estimates. ICCs were high, ranging from 0.95 to 0.99 with supporting small SEMs (1.42 to 3.35 degrees). The authors discussed the relevance of expressing the SEM in terms of its 95% confidence interval for interpreting the reliability of the true value for a single score. This is helpful in expressing the unreliability of a test score (measurement) in an understandable way. Baumgartner explains how the SEM can be used to identify whether differences in measurements between two participants is actual or due to measurement error (Baumgartner, 1989). He explicates that 'confidence bands' (page 62) based on the error statistic are formed around each individual score. If the bands display an overlap the interpretation is that the difference between the measurements is due to measurement error. Conversely, separate bands indicate real differences between the measurements. The conventional criterion usually applied in confidence interval

comparisons covers approximately 95% of the variability; however, the SEM covers approximately only 68%. Multiplying the SEM value by 1.96 will therefore include error for up to two standard deviations, or 95% of the variability (Thomas et al., 2005).

Perhaps a better statistical interpretation of agreement between two tests or two different methods of clinical measurement is offered by 95% LOA (Bland & Altman, 1986). There are various worked examples of clinical measurement studies employing this method for estimating reliability between two tests or approaches. One such study by Romanos et al. (2011) examined inter- and intra-rater reliability of toe systolic pressure and the Toe Brachial Index as a method for establishing blood supply to the foot in diabetic patients. In this repeated measures study, three podiatrists with varying degrees of experience acted as raters. Toe systolic pressures and brachial indices were taken from thirty patients and again one week later by the raters in an investigation of intra-rater reliability. Using the same group of patients, the raters performed both sphygmomanometry and photoplethysmography toe tests five minutes apart to determine the extent of inter-rater reliability. ICCs for both toe tests in the intra-rater and the inter-rater analysis were moderate to high, ranging from 0.72 to 0.91. The authors pointed out that these results were comparable to other studies examining reliability in this field. Romanos et al. improved their estimations compared with other studies, by incorporating 95% LOA into their analysis in addition to the ICC. The results from the Bland & Altman plots determined that bands were wide relative to the overall measurement in both toe tests, suggesting differences between and within raters existed. This was contrary to the encouraging ICC results and statistically insignificant repeated measures ANOVA results ($p < 0.01$) suggesting no systematic differences between raters existed. Using values calculated for the 95% LOA plots, an actual error measurement pertaining to the sample used was applied to exemplify the margins to expect in a clinical environment using the toe systolic pressure and toe brachial index. The results indicated that to determine a difference in pressure for an actual change and not measurement error, the observed change must be ±28mm Hg for the same rater and ±30mm Hg for different raters. For example, for a score of 70mm Hg, the true score (with 95% confidence) will lay between 40mm Hg and 100mm Hg. In terms of systolic

pressure, this is a relatively large error margin and forced the authors to question the reliability of this method of assessment. Similar results were seen regarding the Toe Brachial Index. This article illustrates the importance of interpreting reliability estimates in the context of the study. The advantage of the LOA over the ICC is that the constructed Bland & Altman graphs display an informative visual representation of the degree of agreement. This facilitates identification of any outliers, bias and relationships between the variance in measures and the size of the mean (Bland & Altman, 1986). The 95% limit on the difference between a pair of measurements means that resultant values would be expected to fall within this distance of each other, 95% of the time. Therefore the smaller the repeatability coefficient produced by the constructed graphs, the greater the reliability.

Selfe et al. examined the agreement between measurements taken by the Peak 5 motion analysis system in an exploration of the extent of the system's reliability (Selfe, 1998). In this repeated measures study, seventeen healthy volunteers had reflective markers attached to their hips, knees and ankles. They were then filmed using the video analysis system and the experiment repeated. Intra-rater reliability of the replacing of markers and the reliability of repeated video analysis of three sections of film was scrutinised. Reliability was evaluated primarily using paired t-tests to investigate differences between the repeated tests. The authors reported no significant differences using the t-test and therefore high repeatability of the outcome measures. The paired t-test examines the degree of statistically significant bias between the tests, but should not be employed as a true assessment of reliability as the t-statistic affords no indication of random variation between the tests (Altman, 1991). The detection of a significant difference is dependent upon the degree of random error in the test. If large amounts of random error are present alongside a large degree of systematic bias, the latter will not be detected by the t-test and will result in an acceptable measurement error (Atkinson & Nevill, 1998). Sim (2001) argues that low agreement between tests can actually result in a non-significant t-test. The use of the t-test for determining the extent of reliability between tests is therefore not recommended.

Electronic footprint capture methods have been evaluated by various authors for exploration of pathological conditions that incur abnormal foot pressures and

gait patterns. One such study by Pomeroy et al. (2004) investigated the intra- and inter-rater reliability of operators when measuring tempo-spatial gait parameters of nineteen stroke patients using GaitMat II, an instrumented walkway. The raters consisted of a medical doctor, a biomechanist, a medical student with no experience in either clinical or laboratory gait analysis and two physiotherapists. ICCs with supporting F values from two-way ANOVA calculations were utilised in the estimation of reliability. Intra-rater reliability was shown to be high, with ICCs ranging from 0.84 to 0.93. However, some variation between raters was apparent in that the mean intra-rater reliability for the inexperienced medical student was lower, and ICCs for the inter-rater test also appeared lower for this rater for seven of the seventeen parameters tested. The results for this rater were then excluded and the data reanalysed, resulting in F values that suggested no statistically significant differences in the variation between raters. Also of interest was that as the gait pattern for various subjects became more abnormal, the amount of inter-rater disagreement increased. These findings from Pomeroy et al.'s study may have implications for the reliability investigation of the Reel method. Toe flaring that was noted in the heel and at the apices of certain toe prints can appear as a faint 'smudge' on the images and therefore present a more difficult determination of the start/end pixel for the raters (Figure 7.1). For these types of images, the greater the extent of flaring and reduction in pixel visibility, the greater the disagreement between raters is anticipated.

Figure 7.1 Toe 'flaring'. The fainter part of the image extending distally beyond the apex of the toe print is included in the evaluation

A study determining the reliability of a newly developed semiautomatic method of measuring adult scoliosis from radiographs was considered to bear a resemblance to the reliability study presented in this thesis (Aubin et al., 2011). This article describes a novel method of two-dimensional linear measurement of scanned images (set at 150 dots per inch) using measurement software that has not been previously reported. Radiographic measurement methods prior to the study have traditionally employed manual approaches of the radiographs. Despite the inclusion of measurement software, the authors commented that the approach still 'relies on identifying anatomic landmarks to calculate measurements. The process therefore becomes subject to human error' (page E781). The parallels with this and the Reel method are apparent. In Aubin et al.'s study, thirty two scanned images of adult patients presenting with idiopathic scoliosis deformity were measured by three raters with varying experience. One rater was a research nurse with experience of adolescent idiopathic scoliosis measurement spanning fifteen years, another a qualified radiographer not experienced in measurement and the third rater a clinical orthopaedic coordinator without experience of either radiographs or measurement. ICCs for inter-rater and intra-rater reliability tests were reported to be good to excellent (0.70 - 0.99) for all measurements except sacral obliquity (rater one, 0.56; rater two, 0.77; and rater three, 0.23). The authors suggested the low ICC estimate recorded for rater three regarding this measurement was possibly due to inexperience. ICCs were supported by SD values described in the text as the 'standard variation'. It may have been more informative to have provided 95% confidence intervals to explain the context of the ICC calculations, or to have provided results from additional reliability tests.

The reliability of a method for evaluating left ventricular volumes and function of the heart was examined in Kleijn et al.'s study (Kleijn et al., 2011). This article described intra-rater, inter-rater and test-retest reliability of measurements of the chamber using three-dimensional speckle tracking echocardiography, from one hundred and seventeen patients. This imaging modality relies on an experienced operator to scan the left ventricular chamber of the heart and set markers on the scanned image (two markers at the edges of the mitral valve ring and one at the apex of the left ventricle). Three-dimensional wall motion tracking software automatically measures left ventricular volumes and strain for

a complete cardiac cycle. Again, there are certain parallels with this and the footprint measurement study presented in this thesis, in that measurement involves subjective identification for the placement of markers on scanned images to facilitate the recording of automatic measures. In Kleijn et al.'s study, two separate observers were involved in the inter-rater reliability test and intra-rater measures were taken one week apart. Relative reliability was assessed by way of ICCs calculated from a one-way ANOVA. Absolute reliability was determined using the SEM. In addition to the examination of absolute reliability, Bland & Altman plots of LOA were constructed to further visualise measurement error. High intra-rater, inter-rater and test-retest reliability was reported for left ventricular volume and ejection fraction (0.85 - 0.99), supporting the use of this type of echocardiography for routine evaluation. ICC values for strain were not as high; inter-rater and test-retest coefficients for segmental radial strain were 0.44 and 0.41 respectively. SEM values for this measurement were consequently high compared with other measurements. Coupled with the evidence of substantial systematic bias, the LOA graphs also presented a visual depiction of the large biases for radial strain measurements between raters, and led the authors to recommend that consecutive measurements should be done by one operator only. The article demonstrates a thorough examination of reliability using multiple statistical methods, the results of which afford much clinical relevance.

In summary, it is clear from these studies assessing clinical measurement tools for reliability in all its forms, a single statistical test of the data used in isolation is not recommended. The above literature review illustrated a variety of statistical analytical procedures that can be utilised in the examination of reliability; however some of these were used inappropriately or were inadequate. It appears that the choice of statistical test depends on the design of the reliability study and furthermore, a combination of suitable reliability tests be should be used in order to fully define reliability estimates.

It seems that reliability studies accompanied by more than one appropriate test are necessary to ensure that the involved amount of measurement error is small enough to detect real changes in what is being measured. The more reliable the

measurements are in themselves and the more reliable operators are in performing them is essential in helping clinicians, practitioners and technicians decide whether or not a particular measurement is of any value.

Atkinson & Neville (1998) suggest comparing the tool in question with similar measurement approaches and then choosing the tool affording the least measurement error value, to inevitably reduce random error. However, as the relevant literature failed to unearth a comparable instrument, the researcher looked to other literature in the area of clinical measurement to examine different statistical analytical approaches for what was considered to be the best models in terms of research design and statistical analysis, in order to test the new measurement approach. The article by Aubin et al. (2011) which determined the extent of the reliability of a novel measuring tool to evaluate scanned images of radiographs, seemed to illustrate a methodology and study design more akin to the research presented by this thesis in terms of reliability testing, than the other reviewed publications. Certain elements from Aubin et al.'s study echo and support the design of the methodology used to examine reliability of the of the Reel method, for example, the number of scanned images used and the number and varied experience of the raters. Kleijn et al.'s study demonstrated through appropriate and thorough statistical testing, measurement error can be defined and transferred to a clinical setting (Kleijn et al., 2011). Their methods of analyses confirm the appropriateness of the choice of statistical testing proposed for this investigation of reliability of the new footprint measurement approach; ICC, 95% LOA and 95% SEM.

7.3 Methodology

A repeated-measures design experiment was initiated to enable the analysis of rater reliability and instrument reliability of the Reel method. A fundamental rationale for examining the extent of reliability of a tool is to establish the amount of error, or unexplained variance, the test incurs. In reliability estimation, error can be divided into random error and systematic error, and measurement error is sum total of these two components (Chatburn, 1996). Systematic error refers to a general trend for measurements to differ in a particular direction (negatively or positively) between repeated tests (Atkinson & Nevill, 1998). This type of error can be considered to be the result of bias. For

example, if a positive direction is noted between tests this may be due to a learning effect being present as illustrated in Coldwells et al.'s study investigating measurement reliability of back strength using a portable dynamometer (Coldwells et al., 1994). If the trend for a retest is shown to be lower than a prior test, this could be due to operator fatigue or a lack of motivation (Atkinson & Nevill, 1998). It is anticipated that systematic error from repeated tests of the Reel method may incur all these types of bias in both intra-rater and inter-rater experiments.

Random error is often due to the measurement tool itself and is therefore more difficult to control, compared with the sources of systematic error. Atkinson and Neville identify random error as constituting the majority component of total error in relation to systematic error, and argue that one way of reducing random error is to choose a measurement tool which demonstrates the least measurement error value compared with similar measurement tools. The authors warn that in this case, the same statistic of measurement error should be consistently applied for all tools used in the comparison (Atkinson & Nevill, 1998). In the Reel method, random error could be attributed to image deformation due to the process of image scanning, although the presence of a horizontal and vertical scale as a reference should control this source of error. A more plausible source of error regarding the new measurement approach may lie in the identification and interpretation of the appropriate pixels from which to commence and end the linear measurements, due to low quality of the scanned image or rater subjectivity.

The collection and measurement of static and dynamic footprints from sixty-one subjects has been described previously (Chapter 3 and 4). The reliability investigations comprised of three separate studies in order to examine, 1) reliability of the measurements between sets of footprints, 2) intra-rater reliability of the measurement approach, and 3) inter-rater reliability of the measurement approach.

7.4 Study 1: Between-print reliability

A search of the literature was devoid of papers that appropriately examined within-subject repeatability in terms of their inked footprint measurements. Due

to human variability factors, it would not be improbable to presuppose that the shape of a person's foot impression will vary between consecutive prints, even taken under clinical conditions. Despite this, studies that have examined the repeatability of electronic prints made by, for example, pressure capacitance platforms, have reported good between-print reliability for a person's footprints captured in this way (Mathieson et al., 1999; Zammit et al., 2011). Although outcomes of electronic footprint studies may influence initial hypotheses of inked footprint studies, Urry & Wearing have suggested inferences are not interchangeable, as electronic footprints are completely different to inked ones and do not incorporate the finite borders of the print for analysis (Urry & Wearing, 2005). Study 1 therefore was designed to determine the extent of variation that exists between someone's inked footprints taken from the same foot.

7.4.1 Study 1: Method

In an investigation of intra-individual variation, an analysis was carried out to examine the reproducibility of two width and five length measurements over three consecutive instances in the dynamic state and also in the static state. The footprints were collected under the same environmental conditions for all subjects, described in section 4.5. Measurements from the subsequent scanned images were recorded for three static prints and three dynamic prints for all subjects (61 subjects x 6 prints each x 7 measurements). The measurements were first assessed for normality, followed by an analysis of homogeneity of variance across three groups of measures (print 1, print 2, and print 3 for both dynamic and static states) and finally an exploration of reliability between the sets of footprints.

Thirty static and thirty dynamic footprints randomly selected using SPSS software from the original database had their lengths, widths and angles re-measured by the researcher using GIMP (Version 2.2.17) and the ensuing data tested for normality. The data were then compared with measurements from the original measurement dataset in an exploration of intra-rater reliability. Chinn (1991) argues for the use of twenty six data points or more in order to adequately reflect inferences of reliability coefficients, supported by Bruton et al. (2000). It is for this reason that thirty or more measurements were analysed in each of the reliability sub-studies examined for the purposes of this thesis.

7.4.2 Study 1: Data analysis

Data for all reliability sub-studies were analysed using appropriate statistical testing employing SPSS software.

For Study 1, tests for normality were analysed using K-S tests (appropriate for fifty or more data points). Normality tests such as the K-S test were included as they ascertain whether the data follows a normal distribution allowing parametric testing to be used. Data analysis methods such as ANOVA depend on the assumption that data were sampled from a Gaussian distribution (Field, 2005). However, D'Agostino (1986) has argued that normality tests in isolation do not provide sufficient information as to the exact distribution of data and therefore histograms and Q-Q plots were calculated to provide further information.

Homogeneity of variance was analysed using Levene's test, appropriate for groups of data. The assumption of homogeneity of variance implies that at every level of one variable, the variance of the other measurements should not change (Matthews et al., 1990). In other words, although the value for the mean may increase for a group of measurements, the spread of the scores should be the same at each level for the other measurement variables.

As the review of the literature has shown, several authors have argued for the use of various tests together, rather than just one single estimate in order to produce a more definitive picture of reliability. This method of triangulation is deemed as best practice (Safrit & Wood, 1989; Rankin & Stokes, 1998; Bruton et al., 2000) and was thus chosen for the purposes of demonstrating reliability of the Reel method.

In order to test the extent of reliability across the six measurements for each subject and between the subjects, $ICC_{3,1}$ was employed (calculated from a one-way ANOVA) and also 95% LOA. The ICC is a test of relative reliability and is a ratio of the variance between subjects to the total variance obtained from the ANOVA calculation. Interpretation of the ICC followed the recommendations of Fleiss as follows: > 0.90, excellent reliability, 0.40 to 0.75, fair to good, and < 0.40, poor reliability (Fleiss, 1986).

95% LOA graphs were developed in the following manner in accordance with Bland & Altman (1986). First, the differences between two tests were plotted

against the mean values of the tests. Then the mean and standard deviation of the differences between the measures were calculated. This allowed the mean difference (±2 SD) to be visually demonstrated by way of a scatterplot to establish 95% limits of agreement. This has the advantage in that it emphasises relationships, any bias or outliers between variance in measurements. The size of the mean is also pictorially illustrated using this test.

7.4.3 Study 1: Results

For all measurements, the K-S statistic was non-significant ($p > 0.05$) indicating no deviation from normality. Histograms and Q-Q plots pictorially demonstrated normality (see Appendices F.4 and F.5 for examples). Therefore parametric analysis was supported as all variables for the reliability analyses were found to be normally distributed.

Levene's statistic for the static and dynamic measures were non-significant ($p > 0.05$), suggesting little variance and therefore homogeneity between the measurements. Normal distributions across all measurements allowed for parametric analysis.

Three static and three dynamic prints collected from each subject had all sets of length and width measurements analysed (2,562 data points). Descriptive data for these measurements are presented in Table 7.1. Absent fifth toe prints facilitated grounds for exclusion for some of the Calc_A5 measurements.

Table 7.1 Descriptive data for static and dynamic length and width measurements

Measurement	Print type	No	N	Mean (mm)	Min (mm)	Max (mm)	SD
Calc_A1	Static	1	61	237.74	203.00	276.20	17.46
		2	61	237.91	203.80	276.20	17.24
		3	61	238.16	203.30	276.20	17.34
	Dynamic	1	61	254.67	209.40	298.80	19.37
		2	61	255.45	208.20	294.60	19.12
		3	61	255.76	211.60	297.60	18.54
Calc_A2	Static	1	61	237.12	199.80	281.00	17.09
		2	61	237.31	200.50	281.50	17.07
		3	61	237.54	201.40	281.20	17.29
	Dynamic	1	61	249.32	208.50	296.10	19.37
		2	61	249.99	211.00	295.40	19.29
		3	61	250.40	209.20	294.40	18.86
Calc_A3	Static	1	61	227.99	191.50	269.40	16.24
		2	61	228.34	190.80	270.10	16.54
		3	61	228.50	192.10	269.70	16.73
	Dynamic	1	61	238.70	197.00	284.30	18.26
		2	61	239.12	199.00	278.70	17.85
		3	61	239.61	198.90	280.30	17.62
Calc_A4	Static	1	61	215.69	186.80	255.69	15.30
		2	61	215.70	186.10	256.30	15.34
		3	61	216.04	185.90	256.20	15.86
	Dynamic	1	61	224.82	191.60	266.90	17.14
		2	61	225.06	191.80	264.30	16.89
		3	61	225.45	193.30	262.30	16.59
Calc_A5	Static	1	55	198.37	174.80	232.20	13.75
		2	56	197.97	175.10	232.90	13.90
		3	55	198.11	174.60	232.60	14.21
	Dynamic	1	58	207.20	179.20	244.50	15.97
		2	59	207.91	180.30	244.10	15.74
		3	59	208.31	182.00	245.30	15.68
MPJWidth	Static	1	61	93.09	79.60	108.40	7.54
		2	61	93.39	80.50	107.00	7.39
		3	61	93.61	80.50	108.20	7.64
	Dynamic	1	61	93.23	79.60	107.50	7.39
		2	61	93.02	80.40	107.30	7.19
		3	61	93.16	80.10	107.10	7.00
CalcWidth	Static	1	61	48.89	39.20	63.00	5.18
		2	61	48.96	40.10	63.90	5.00
		3	61	49.11	40.40	63.10	5.10
	Dynamic	1	61	49.68	38.70	62.00	5.33
		2	61	49.47	38.80	63.00	5.32
		3	61	49.83	41.50	61.30	5.09

The measurements were scrutinised for reliability using ICCs calculated from a one-way ANOVA ($ICC_{3,1}$). All measurements displayed high relative reliability with intraclass correlation coefficients greater than 0.9.

In a previous exploration of the data, (Chapter 4, Establishing Evidence of Convergent and Discriminant Validity), it was determined that length and width measurements were highly associated with one another (r values ranging from 0.95 to 0.97, $p < 0.001$). Therefore, just one length measurement, from the base of the heel to the apex of the first toe (Calc_A1), was chosen for an in-depth reliability analysis. This measurement is predominant in footprint measurement literature as it is considered to be the longest and therefore indicative of total footprint length. This is especially true of anthropometric studies that predict stature from footprint length (Robbins, 1986; Krishan, 2008a; Kanchan et al., 2012). Because of the effects of variation upon the medial longitudinal arch (Reel et al., 2012) it was thought that this measurement may be the least stable footprint measurement when considering all the lengths and widths and would therefore challenge between-print reliability, displaying larger unreliability estimates compared with others.

Small variations in the differences of the means of the footprint Calc_A1 length measurements between individuals are pictorially represented in an error bar graph (Figure 7.2). Static and dynamic differences are also described in this figure.

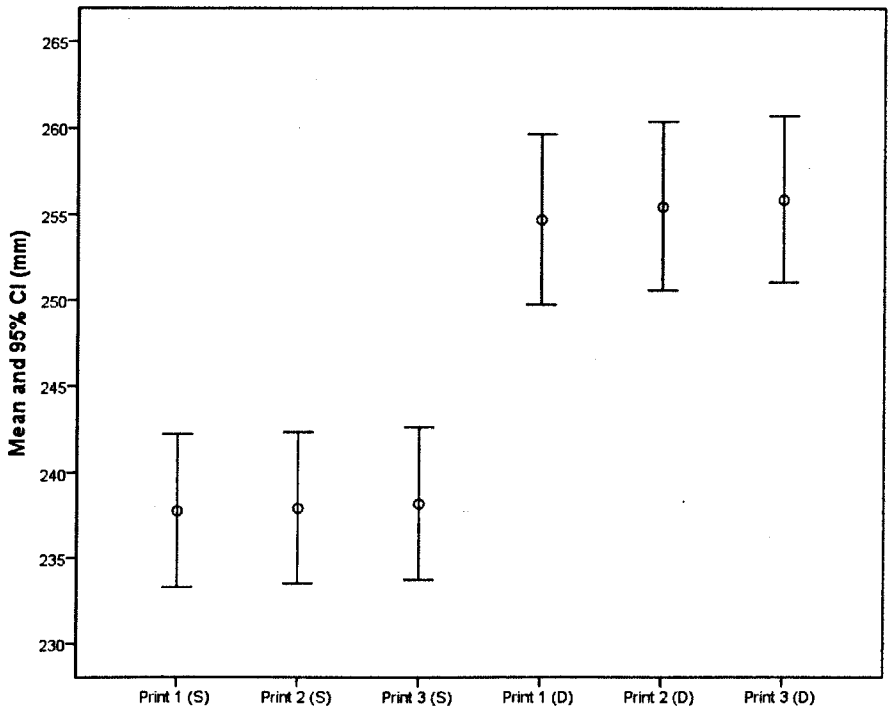


Figure 7.2 Graph illustrating the differences in the means across 3 footprints (1, 2 and 3) from each of 61 subjects, in both static (S) and dynamic (D) states for a given length measurement, (Calc_A1). It also illustrates differences found between static and dynamic length measurements

Calc_A1 measurements from static and dynamic print numbers 2 and 3 were chosen for use in creating the Bland & Altman plots illustrating 95% limits of agreement. Previous literature has shown that this is acceptable as the first of a series of measures is often inconsistent with the rest and therefore discarded from the onset of the experiment (Burnett et al., 2007; Reid et al., 2007). In this study, although the means and SD for all linear measurements in both states were similar, the first measurement for each set of prints was discarded since the means of the second and third Calc_A1 lengths presented the closest values between the three lengths.

The means of the two static measurements (n = 61) were placed against the x-axis and the paired differences of the measurements on the y-axis. The resulting graph presented three outliers outside the ±2SD confidence limits yet pictorially demonstrated how the scores were tightly clustered around the mean, reflecting reliability between the measurements (Figure 7.3)

Despite the existence of three outliers in both the static and dynamic plots, 95% LOA showed that the static scores were more tightly clustered around the paired mean line than the dynamic measurements, therefore demonstrating that the static measurements were more reliable and showed less variability than the dynamic equivalent (Figure 7.4).

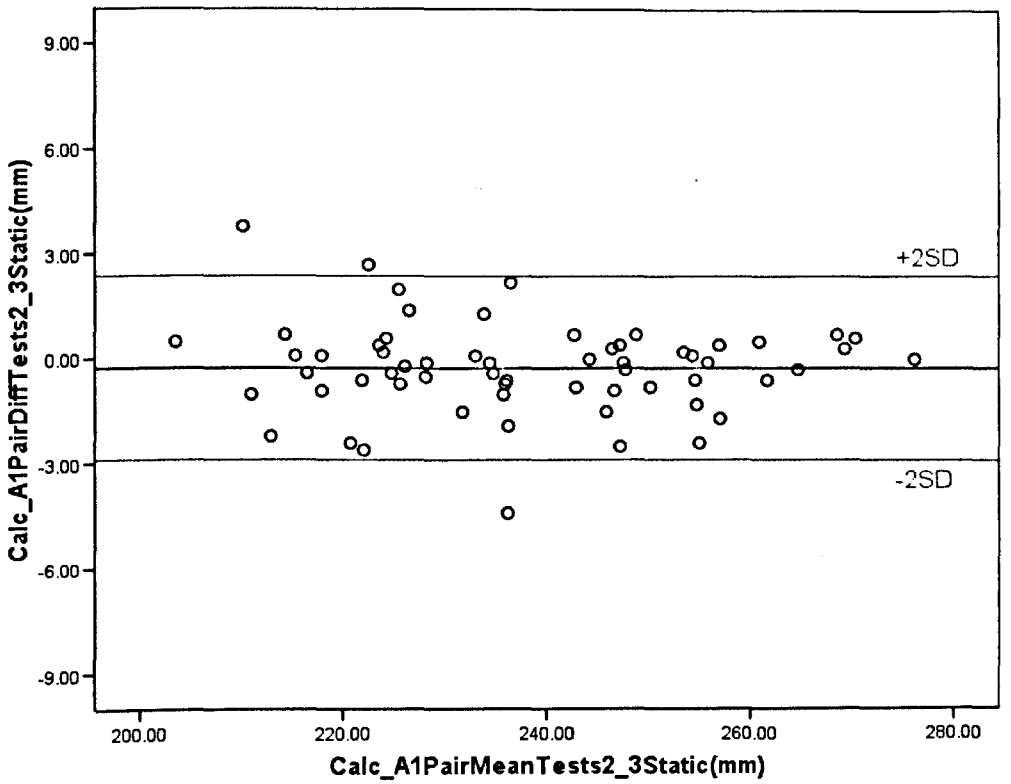
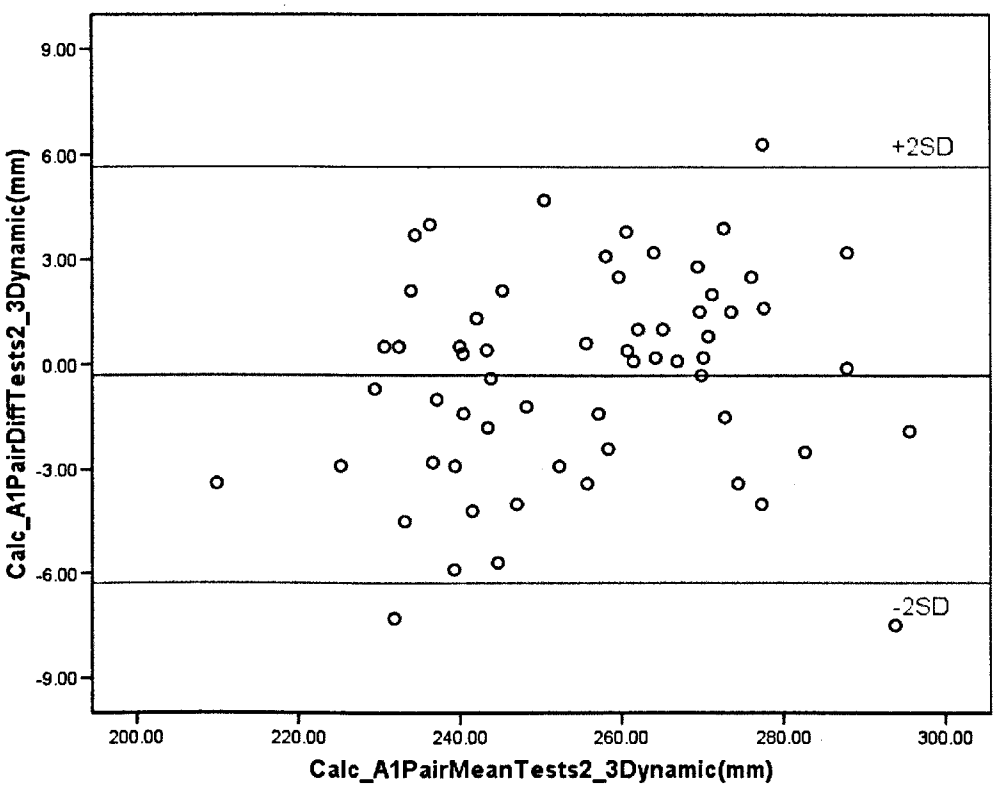


Figure 7.3 Bland & Altman plot of limits of agreement for the paired static Calc_A1 measurements (prints two and three), n = 61. The red line represents the mean difference between the repeated tests, and the blue lines define the limits of agreement (±2SD).

Figure 7.4 Bland & Altman plot of limits of agreement for the paired dynamic Calc_A1 measurements (prints two and three), n = 61



To counter the confounding influence of homogeneity and heterogeneity of the sample on the reliability coefficient, the chosen length measurement (Calc_A1) was split into static, dynamic, male and female groupings (n = 30 for each group).

The ICC analyses suggested high relative reliability, ranging from 0.95 to 0.97 and 95% LOA demonstrated good agreement between the measurements in their groupings although the dynamic prints displayed wider bands than their static equivalents (Table 7.2).

Table 7.2 Reliability analysis of the length measurements from the base of the heel to the apex of the first toe

Variables	Mean Difference (mm)	ICC	95% LOA Upper	95% LOA Lower
Male Static	-0.47	0.96	1.77	-2.71
Female Static	-0.04	0.97	2.93	-2.99
Male Dynamic	0.07	0.95	5.75	-5.61
Female Dynamic	-0.70	0.95	5.54	-6.94

ICC Intraclass correlation coefficient

LOA Limits of agreement

7.4.4 Study 1: Discussion

The descriptive statistics showed that the means for the three dynamic measurements were very close, as were the standard deviations in all groups of measurements. French has argued that observations from descriptive data can be more representative of the reliability of the three measurements than the more complex analyses involving ICCs and LOAs (French, 1988). However, demonstrating the extent of error by way of various statistical tests with supporting reflections of agreement and consistency as procured by the 95% LOA would be more desirable.

High relative reliability, indicated by ICC values greater than 0.9 across all measurements, demonstrated little variation existed between the six repeated footprint dimensions for each person. It did not, however, indicate the amount of disagreement between the measurements and necessitated Bland & Altman LOA graphs be constructed for further analysis.

In the construction of the graphs, paired means and paired differences of two measurements were calculated. By pairing the means, outliers become less prominent. The resulting graphs for the Calc_A1 measurement did actually

present three outliers outside the two standard deviation levels above and below the means, for both static and dynamic footprints. Outliers possibly exist in this study because of the 'people factor' – the natural variability of human footprints. The LOA graph pictorially demonstrated how the scores were tightly clustered around the mean which reflected reliability between the two measurements. Wider interval bands were recognised for the dynamic Calc_A1 measurements. Greater systematic variation occurred within the dynamic measurements possibly as a result of measurement inclusion regarding the toe and heel print flaring effect. This was not only difficult to measure in terms of defining start and end pixel on the footprint image but it was also thought that the amount of flaring may vary between each step taken by a single subject. Both these sources of potential error would account for a larger variance in the dynamic prints compared with the static footprints. When the measurements were split into homogenous groupings, wide intervals were again demonstrated in the dynamic measurements for both male and female footprints. Further investigation showed that the bands for the female dynamic footprint measurements were slightly wider than the equivalent male measurements. It was postulated that determination of the start and end pixel for automatic measurement purposes, would be more difficult in less defined flared heel and toe print images. A heavier person may produce a more defined print than a lighter person, though further research is needed to support this hypothesis. A subsequent t-test determined the male group was on average 14.5kg heavier than the female group, which may support this notion.

7.5 Study 2: Intra-rater reliability

7.5.1 Study 2: Method

A repeated measures study was designed to assess the intra-rater reliability of the measuring method. For this part of the analysis, static and dynamic prints were selected from thirty subjects picked at random. Since all length measurements were highly correlated with one another, ascertained from a previous investigation (section 4.6.1) only one measurement, Calc_A1 was deemed necessary for analysis.

Also included in the reliability analysis was the testing of selected angle measurements (the footprint angle and the 2-5 toe angle of declination) and

width measurements across the heel and forefoot areas of the print. Each print was rescanned and the central axis located on each image. The constructed width, angles and length lines were re-measured according to the Reel method and entered into a new dataset for comparison with the original measurement results.

7.5.2 Study 2: Data analysis

For Study 2, tests of normality were analysed using the K-S statistic and Shapiro-Wilk test. The first part of the intra-rater analysis examined normality plots of each measure across thirty footprints; therefore the Shapiro-Wilk test was included in this analysis as it is deemed suitable for exploring less than fifty data points (D'Agostino, 1971). Homogeneity of variance was examined using Levene's test.

Using the best considered reliability estimation models from the literature, the following statistical tests were chosen to assess the extent of reliability; SEM and 95% LOA for absolute reliability, and ICC calculated from a one-way ANOVA. The SEM and 95% SEM were calculated using the formulae previously discussed in section 5.5.3.

7.5.3 Study 2: Results

Normality tests with supporting histograms and Q-Q plots ascertained a normal distribution of all measurements ($p > 0.05$). Results of Levene's test determined all data assumed homogeneity of variance ($p > 0.05$). Thus parametric analysis and analysis of variance were supported.

Descriptive data for the selected measurements are presented in Table 7.3.

Table 7.3 Descriptive data for repeated length, width and angle measurements from combined static and dynamic footprints (n = 30).

Measurement	Mean (mm)	Min (mm)	Max (mm)	SD
Calc_A1	242.68	216.80	275.20	14.44
Calc_A1 Repeat	242.80	216.90	276.00	14.45
CalcWidth	48.02	38.70	56.20	5.11
CalcWidth Repeat	48.11	38.50	56.60	5.18
MPJWidth	92.64	82.70	108.10	7.18
MPJWidth Repeat	92.46	81.60	108.30	7.21
Measurement	**Mean (°)**	**Min (°)**	**Max (°)**	
2-5 Toe Angle	36.28	22.25	48.27	7.08
2-5 Toe Angle Repeat	36.61	22.49	50.36	7.30
Footprint Angle	46.87	33.60	58.77	6.72
Footprint Angle Repeat	46.42	33.04	58.14	6.51

In the intra-rater reliability analysis, ICCs ranged from 0.98 to 0.99. Results for SEM and LOA also suggested a good agreement between the two tests when original measurements were compared with the scores of repeated measurements (Table 7.4).

Table 7.4 Intra-rater reliability analysis of selected length, width and angle measurements (n = 30)

Measurement	Mean difference	ICC	95% LOA Upper Lower	SEM	95% SEM
Calc_A1	-0.13mm	0.99	0.65 -0.91	0.43mm	0.84mm
CalcWidth	-0.09mm	0.99	1.21 -1.39	0.52mm	1.02mm
MPJWidth	0.18mm	0.99	1.64 -1.28	0.72mm	1.41mm
2-5 Toe Angle	-0.33°	0.98	2.79 -3.45	1.02°	2.00°
Footprint Angle	0.45°	0.98	3.09 -2.19	0.94°	1.84°

ICC Intra-class correlation coefficient

LOA Limits of agreement

SEM Standard error of measurement

7.5.4 Study 2: Discussion

Construction of linear and angle dimensions for measurement across scanned images of inked footprints was a subjective process, despite the advantages of employing automated software to facilitate this method. As previously discussed, determination of appropriate start/end pixels as measurement markers, potentially provided the greatest source of error. In this intra-rater reliability investigation, the extent of this error explored by different statistical approaches was shown to be limited.

The ICCs for each measurement demonstrated near-perfect coefficients (0.98 – 0.99), but the use of this statistic used in isolation can give rise to misleading inferences, as previously demonstrated in the between-print reliability analysis. Results from plots of limits of agreement presented small interval bands and a close position of the mean of the paired differences in comparison with measurement value 0.00mm. Largest bands were seen in the angle

measurements. This was not surprising as creation of the angles on the scanned images requires identification of several key pixels for construction and measurement, thus increasing the amount of potential variation for repeated measurements.

The 95% SEM values provide a reflection of error within a real-world setting. For example, for a heel width footprint measurement of 48mm, 95% of repeated measurements of that same footprint by the same rater would yield scores ranging from 46.98mm to 49.02mm.

7.6 Study 3: Inter-rater reliability

7.6.1 Study 3: Method

The final analysis, inter-rater test, rounds off the comprehensive evaluation of reliability of the new footprint measurement approach. In order to investigate the limitations of inter-rater reliability, two additional operators, both students from the University of York St John, volunteered to be involved in a repeated measures study design after ethical approval had been granted (Appendix D.1). Both students did not have prior experience of footprint measurement and neither was familiar with automatic measurement of scanned images; however both were postgraduate students of physiotherapy and were therefore knowledgeable of measurement concepts. Their inexperience of this particular task was intended to reflect the range of experience in measurement within practitioners actively engaged in forensic footprint evaluation in the field (section 2.1.1) and to fully test the boundaries of measurement reliability of the Reel method. The paucity of forensic podiatry research coupled with recent criticisms regarding a lack of scientifically based experiments in similar forensic identification fields (Saks & Faigman, 2008; National Research Council, 2009), warrant further investigation into footprint measurement. Undergraduate and postgraduate research is currently being undertaken in the area of forensic podiatry (University of Huddersfield, 2012) therefore footprint evaluation undertaken by relatively inexperienced students must be considered. The participants selected for this part of the study were chosen to reflect this range of expertise.

The same group of thirty randomly picked scanned footprint images consisting of fifteen dynamic and fifteen static prints were given to each student in JPEG form. The participants downloaded the scanned images onto their individual PCs. As the footprint measurements were highly associated and displayed similar behaviours (section 4.6), only one measurement (Calc_A1) was chosen for this inter-rater reliability analysis. The volunteers were mentored briefly by the researcher as to the Reel method of line construction and measurement of Calc_A1. In order to further facilitate this part of the study, the students were each given copies of a guide to measurement produced by the researcher (section 3.6). Measurements were constructed and values recorded independently of each other and of the researcher.

In an intention of further exploration of intra-rater reliability, the process of construction and recording measurements was repeated by each volunteer. The recorded values were analysed for estimates of normality and then compared with measurement values from the same rater and between raters, including the same set of measurement results from the researcher (three raters in total). In this instance, the researcher is considered to be practised in the task, therefore the inclusion of the researcher's recorded measurements for the inter-rater reliability analysis serve not only to increase rater numbers but also to challenge the underlying hypothesis regarding the pragmatism of the Reel method. The researcher may have considered the task to be a practical and simple method of measurement, but only by testing the theory using other inexperienced operators and analysing the ensuing data can this be proven.

The literature review of clinical measurement inter-rater studies did not suggest a finite number of raters for adequate reliability testing. For example, in a study investigating the inter-rater reliability of assessing irradiated skin using ultrasound techniques, two operators were chosen who were colleagues and experienced in the field of ultrasound (Huang et al., 2007). Similarly, two experienced physiotherapists were used in an inter-rater reliability study determining the reliability of different measurement tools to assess movement in burns survivors (Edgar et al., 2009). In a study that has parallels with the reliability analysis offered by this thesis, Aubin et al. used data recorded from measurements taken by three operators of differing professions and thus differing experience for an examination of inter-rater reliability of radiographic

software (Aubin et al., 2011). In the field of fingerprint identification, a second fingerprint officer verifies the result of the initial examiner in an inter-rater reliability quest for validation and this number of raters is considered sufficient for the task (Speckels, 2011). The number and the amount of experience of raters employed depend therefore on the test or technique under reliability examination. It was considered that the use of three operators with varied experience would be sufficient to challenge inter- and intra-rater reliability of the Reel method.

7.6.2 Study 3: Data analysis

Data from thirty length measurements were analysed for distribution using Shapiro-Wilks tests, supported by Q-Q plots and histograms. Homogeneity of variance was investigated using Levene's test.

Assuming homogeneity of variance, the data were further explored for intra- and inter-rater reliability within and between raters' measurements using intraclass correlation coefficients based on a one-way ANOVA. In a further exploration and to examine the extent of absolute reliability, 95% SEM and 95% LOA plots were produced for the intra-rater study and 95% SEM for the inter-rater study.

7.6.3 Study 3: Intra-rater results

All measurements displayed scores that were normally distributed. The results of Shapiro-Wilks and Levene's statistical tests were non-significant across the measurements for the three raters, permitting parametric analysis of the data and an assumption of normality throughout ($p > 0.05$). Histograms and Q-Q plots supported the calculation of normality, exemplified by appendices F.6 and F.7.

Descriptive data for sixty measurements recorded by three independent raters are displayed in Table 7.5.

Table 7.5 Descriptive data for thirty repeated Calc_A1 measurements recorded by three raters

Measurement	Mean (mm)	Min (mm)	Max (mm)	SD
Rater 1				
Calc_A1 Original	242.67	216.80	275.20	14.44
Calc_A1 Repeat	242.80	216.90	276.00	14.44
Rater 2				
Calc_A1 Original	242.88	216.70	276.10	14.60
Calc_A1 Repeat	243.10	217.00	276.10	14.57
Rater 3				
Calc_A1 Original	242.57	216.20	275.80	14.49
Calc_A1 Repeat	242.64	216.50	275.40	14.43

The average mean value for the six groups of measurements (242.78mm) was comparable with the average median value (242.73mm) indicating similar traits existed between the measures, despite the large range within this particular footprint length.

The error bar graph (Figure 7.5) illustrates the close agreement between the means across the raters' measurements and the similar amount of variation exhibiting in all measurements.

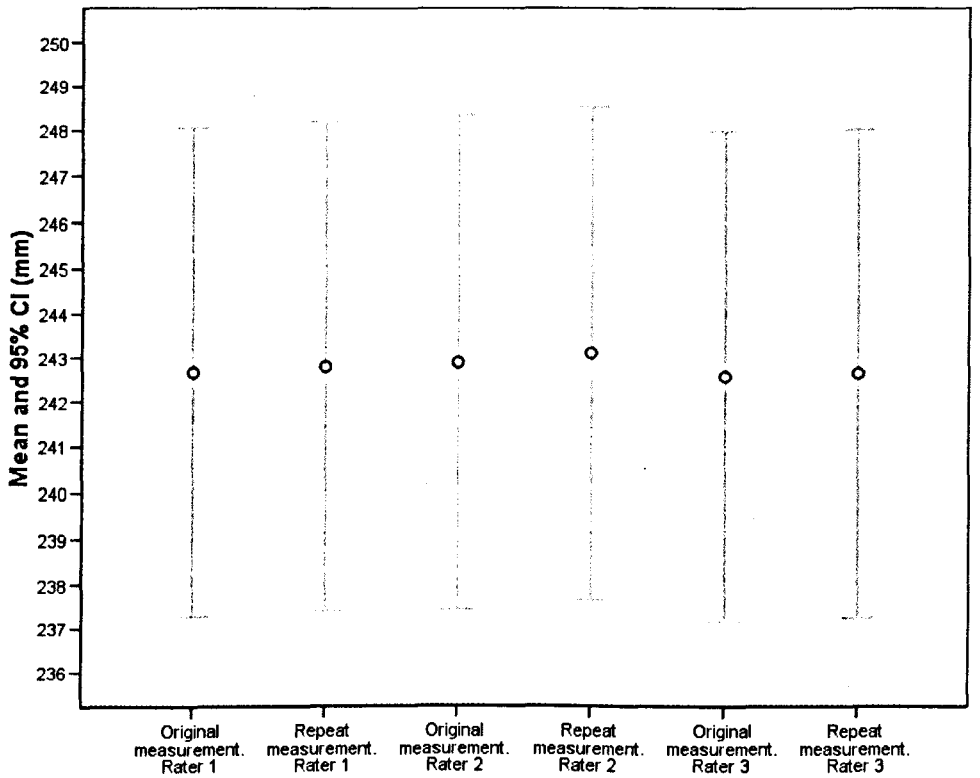


Figure 7.5 Error bar illustrating mean and 95% CI of repeated measurements of Calc_A1 between raters.

In the context of the average length measurement of 242.78 mm, the mean difference between test 1 (original measurements) and test 2 (repeated measurements) for all raters was low; 0.13mm (Rater 1), 0.25mm (Rater 2) and 0.10mm (Rater 3). Reliability statistics as shown in Table 7.6 resulted in high values. ICCs of 1.0 between test 1 and test 2 for all raters demonstrated highest reliability. The 95% SEM values are a reflection of the average standard deviation from the repeated tests combined with the results of the ICC for each rater. Repeated test results gathered by Raters 1 and 3, show greater overall reliability than Rater 2, as indicated by the 95% SEM.

Table 7.6 Intra-rater reliability analysis of Calc_A1 measurements between three raters (n = 30).

Measurement	ICC	95% LOA Upper Lower	SEM (mm)	95% SEM (mm)
Rater 1				
Calc_A1 Original/repeat	1.00	0.65 -0.91	0.43	0.84
Rater 2				
Calc_A1 Original/repeat	1.00	0.32 -0.76	0.44	0.86
Rater 3				
Calc_A1 Original/repeat	1.00	0.55 -0.69	0.43	0.85

ICC Intra-class correlation coefficient

LOA Limits of agreement

SEM Standard error of measurement

The interval widths for all raters illustrated by the Bland & Altman graphs of agreement are perhaps a better reflection of reliability in this instance. The 95% LOA graphs warrant closer attention and are depicted in Figure 7.6. From these visual representations of reliability estimates, it can be seen that although the measurement scores for the tests recorded by Rater 2 appeared to have the closest agreement illustrated by the tight interval bands on the y-axis, the mean paired difference between test 1 and 2 (denoted by the red reference line) is the furthest away from the definitive 0.0 value on the y-axis. In contrast, the widest interval bands in the graph depicting reliability of the scores for Rater 1 indicate the poorest agreement compared with the other raters. However, Rater 3 demonstrated the smallest paired mean differences across all raters.

In an investigation to verify the previous belief that the images from dynamic prints would incur more rater error than their static counterparts, the measurements from the same sample were divided accordingly and further analysed. Descriptive data is displayed in Table 7.7.

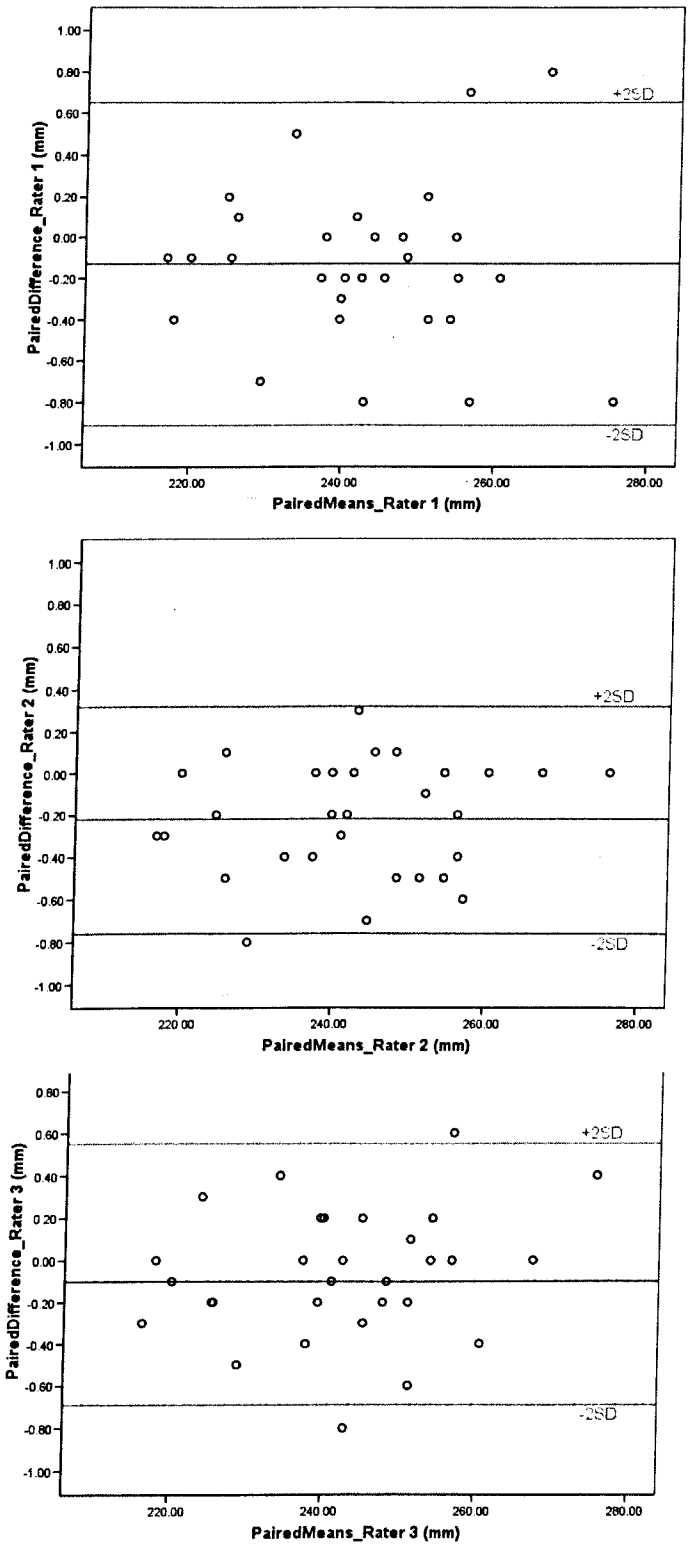


Figure 7.6 LOA graphs for three raters with one repeated measurement (n = 30)

Table 7.7 Descriptive data for repeated Calc_A1 static and dynamic measurements (n = 30) for three raters

Static Measurement	Mean (mm)	Min (mm)	Max (mm)	SD
Rater 1 Calc_A1 Original	237.97	216.80	260.40	14.15
Rater 1 Calc_A1 Repeat	238.00	216.90	260.60	14.14
Rater 2 Calc_A1 Original	238.03	216.70	260.30	14.18
Rater 2 Calc_A1 Repeat	238.29	217.00	260.30	14.17
Rater 3 Calc_A1 Original	238.09	216.20	260.40	14.10
Rater 3 Calc_A1 Repeat	238.19	216.50	260.80	14.10
Dynamic Measurement	**Mean (mm)**	**Min (mm)**	**Max (mm)**	**SD**
Rater 1 Calc_A1 Original	247.37	226.10	275.20	13.56
Rater 1 Calc_A1 Repeat	247.60	226.00	276.00	13.52
Rater 2 Calc_A1 Original	247.73	225.80	276.10	13.80
Rater 2 Calc_A1 Repeat	247.91	225.70	260.30	13.77
Rater 3 Calc_A1 Original	247.05	224.10	275.80	13.91
Rater 3 Calc_A1 Repeat	247.08	223.80	275.40	13.80

All data were found to be normally distributed, allowing for parametric statistical analysis.

ICCs of 0.99 to 1.00 indicated exceedingly good reliability existed across all analyses (Table 7.8). The interval bands in the construction of 95% LOA

remained close, but were wider for the dynamic footprint impression repeated tests for two out of three raters. Rater 2 had wider intervals for the repeated static measurements. 95% SEM, a practical margin of error, resulted in smaller estimates for the evaluation of the dynamic prints rather than the static ones (Table 7.8).

Table 7.8 Reliability analysis of static and dynamic Calc_A1 measurements (n = 30) for 3 raters

Static Measurements	ICC	95% LOA Upper Lower	SEM (mm)	95% SEM (mm)
Rater 1				
Calc_A1 Original/repeat	1.00	0.40 -0.46	0.43	0.83
Rater 2				
Calc_A1 Original/repeat	1.00	0.33 -0.83	0.43	0.83
Rater 3				
Calc_A1 Original/repeat	1.00	0.35 -0.55	0.42	0.83
Dynamic Measurements	**ICC**	**95% LOA Upper Lower**	**SEM (mm)**	**95% SEM (mm)**
Rater 1				
Calc_A1 Original/repeat	0.99	0.76 -1.22	0.41	0.80
Rater 2				
Calc_A1 Original/repeat	1.00	0.33 -0.69	0.41	0.80
Rater 3				
Calc_A1 Original/repeat	1.00	0.74 -0.80	0.42	0.82

ICC Intraclass correlation coefficient

LOA Limits of agreements

SEM Standard error of measurement

7.6.4 Study 3: Inter-rater reliability results

Data from repeated testing of the Calc_A1 static and dynamic measurements between three raters were analysed to explore the extent of error occurring between raters.

95% SEM displayed small variance allowing a ±1.27mm error margin for the Calc_A1 length between all three raters (average SD 14.50, ICC 1.00 (average measures). The high ICC values demonstrate insufficiencies of this particular statistical test, when compared with the 95% SEM calculation.

7.6.5 Study 3: Discussion

Although high ICCs of 1.00 suggested near-perfect reliability between all three raters, further statistical investigations of the intra-rater analysis illustrated different traits amongst the raters (section 7.6.3). This was most evident in the graphs of 95% LOA (Figure 7.6). There are three key features to limits of agreement; the number and positioning of outliers, the position of the mean of the paired differences in comparison with measurement value 0.00mm, and how great or small are the values of the paired difference standard deviations.

Many outliers would imply poor repeatability and therefore poor reliability. Outliers were apparent in all raters' plotted comparisons but none were considered too far away from the 2SD boundary to indicate poor reliability as the upper and lower boundaries fell within 1mm.

The red line in each graph (Figure 7.6) denotes the mean of the paired differences between a rater's scores from the first set of measurements obtained, compared with the scores from the repeated set of measurements. The closer the mean of the paired differences of the scores to 0.00mm, the more reproducible the measurement results are between the original and repeated tests. A mean value disparate to 0.00mm would imply discrepancies when the measurements were repeated.

A small standard deviation of the paired differences would imply that the tester's results were consistent in that little variation occurred for the all the collective repeated measurements and values for all paired differences between the original and repeated test were very small.

The LOA graph representing Rater 1's scores showed measurement recordings that fell within the largest interval band width compared with the other raters. This indicates Rater 1's scores were the least consistent.

The graph representing Rater 2's scores showed small standard deviations yet with mean values of -0.22mm, a comparatively large distance from 0.00mm. This implies that compared with the other raters, Rater 2's reliability results were the most consistent, but the least repeatable.

95% LOA results demonstrating the recorded measurements from Rater 3 displayed paired mean difference values close to 0.00mm. This reflects the greatest repeatability of the comparison scores from Rater 3 compared with the other raters.

Chapter 2 discussed the concept of validity whereby if a test is valid, it measures what it is supposed to measure. This is true if the test is measured against a known gold standard (previously discussed in Chapter 5) which in this case, does not exist. It would be better to say that validity is not necessarily a property of the test, but rather the extent to which an approach yields useful information for a specific purpose (Goodwin & Leech, 2003). The contexts in which the item(s) being measured must be therefore considered before reliability statistics are applied to infer statements about the test's validity. In the 95% LOA example, a paired mean difference value of -0.22mm between two tests would imply high repeatability if the thirty footprint length measurements had a large range e.g. from 180mm to 320mm. A paired mean difference value of -0.22mm would imply poor reliability however, if the footprint lengths ranged from 2.0mm to 2.5mm. For the randomly selected sample in this study, the footprint measurements ranged from 216mm to 276mm; the largest paired mean value across raters of -0.22mm implies the poorest repeatability amongst raters but could not be considered as demonstrating low reliability in the context of the study.

The term 'consistent' is linked with the words 'precision' and 'reliable'. Rater 2's scores were the most consistent out of the three raters, yet the least repeatable. Rater 2 did not practise the construction and measuring of the footprint length line before commencing the exercise; she read the instructions without verbal

guidance and the task was immediately commenced. However, this rater punctuated the task with frequent breaks, after every fifth footprint completed. Rater 2 completed her measurements in the shortest time amongst the three raters. The method used by this rater is analogous to a do-it-yourself enthusiast who measures lengths of wood using their arm instead of a tape measure. The wood finally purchased may appear to be the same length, but not precisely the correct measurement required for the job.

Rater 3 practised several times before performing the task set and was in frequent communication with the researcher in an intention to perfect the measurement technique. This modus operandi is reflected in the smallest mean paired difference value of -0.10 across scores indicating the highest repeatability amongst raters.

Rater 1 displayed the greatest variation for collective repeated measurements indicating the least consistency amongst raters illustrated by 95% LOA graph. This rater, the researcher, was complacent in the task. Instead of taking frequent breaks, Rater 1 spent a few hours at a time on the task before having a rest of several hours. This approach increased the amount of random error in results. However, 95% SEM reliability results of 0.84mm between two tests for Rater 1 indicated a slightly lower error uncertainty than the other raters.

Literature pertaining to inter-rater studies has observed the extent of reliability is often dependent on the experience of the raters. For example, in their study examining the reliability of the manual supination resistance test as a diagnostic aid for prescribing foot orthoses, Noakes & Payne (2003) found differences in raters depending on their experience. They established that ICCs for the two experienced clinicians were good (0.82 and 0.78) but were poor for the two inexperienced raters (0.56 and 0.62). Similarly, in Pomeroy et al.'s investigation of the reliability of an instrumented walkway, GaitMat II, five raters were used. Only one of these raters was inexperienced in both clinical and laboratory-based gait analysis and ensuing ICCs for this rater's measurements were the lowest in comparison with the other raters (Pomeroy et al., 2004).

Despite these accounts of inexperienced rater unreliability, it is apparent that the extent of reliability of the raters performing measurements using the Reel method was reflective of the different practical approaches adopted rather than

experience or expertise. This revelation prompted further practical recommendations when evaluating footprint images using the Reel method; 1) to practice the method several times before the start of data collection, and 2) to rest every fourth or fifth footprint measurement, as supported by the actions of Rater 2. It can be argued that the measurement rigour of the Reel method is so robust it is not influenced by rater inexperience. This holds wider implications in that the method does not need to be confined for use within a particular profession or background.

Split into their homogenous groupings, ICCs were exceedingly high for both static and dynamic footprint measurements inferring high relative reliability. 95% LOA interval bands for both static and dynamic measurement recordings were similar although the bands for the dynamic footprints were slightly wider, echoing the results of the intra-rater study (Study 1, section 7.4). Greater variation within the comparison scores from the dynamic prints were expected over the static prints between raters, due to incursion of error when measuring heel and toe flare. However, calculation of the 95% SEM depicted a different supposition as values were smaller for the dynamic group. Thus absolute and relative reliability displayed good agreement across both static and dynamic groups and previous concerns regarding the unreliability of measuring image flare appear to be unfounded.

In the final inter-rater analysis, the ICC once again displayed exceedingly high relative reliability of scores from all raters, despite the calculation of an actual error margin of ±1.27mm as proposed by the 95% SEM. This can be explained by the calculation of the intraclass coefficient, which determines the smaller the error variance, the closer the coefficient value is to 1.0. However, in this study, the true variance is equivalent to the between-subject differences of the footprint measurements and the measurement error equates to the within-subject differences of the footprint measurements. This means that if the within-subject differences are small, ICC reliability estimates will be high. If the between-subject differences are large, the resultant reliability coefficient will be even greater. This statistical weakness can be likened to a target in which shots fall close to the bull's eye (Figure 7.7):

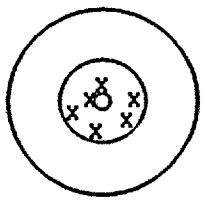


Figure 7.7 Small within-subject difference and high reliability

If the target size is increased yet the shots remain on target as before, the overall effect is that reliability is deemed to be even higher (Figure 7.8).

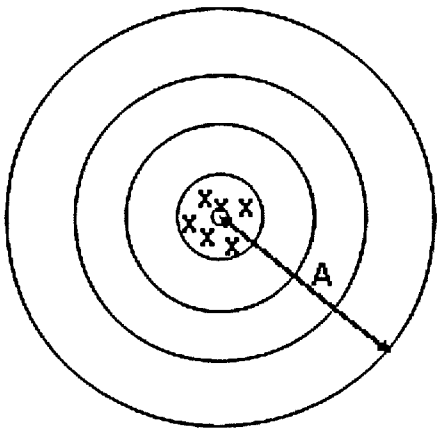


Figure 7.8 Large between-subject difference (denoted by A) increases the reliability coefficient from Figure 7.7 (above), despite values being the same

Therefore the results from the ICC calculation must be taken in the context of the study and highlights the inappropriateness of using this statistical method in isolation to determine differences between groups to estimate reliability of repeated measures.

7.7 Conclusions

Establishing the extent of reliability of a new measurement approach is an essential component of determining its overall validity. Without these fundamental explorations, advancement in the area of footprint evaluation cannot proceed. Despite the Otway appeal case in which experience-based opinion was admitted involving forensic gait analysis evidence (Otway v R.,

2011), footprint identification evidence in a court of law may flounder in the future in the absence of the vital foundations of reliability and validity. This is supported by the forensic science regulator's recent stringent demands relating to demonstrability of acceptable reliability and validity (Rennison, 2011).

A search of the literature did not offer suitable reliability studies pertaining to footprint measurement. Alternative information was sought by reviewing articles pertaining to medicine in line with other areas of this thesis which draws parallels between the ideologies of forensic science and medicine. Thus, the area of reliability assessment of clinical evaluation tools was selected for use as a model for the reliability estimation of the Reel method. In these studies, various statistical approaches had been adopted. Supported by literature in the field of statistical analysis, certain tests were deemed inappropriate for the chosen reliability analysis of the appraised papers. The ICC, 95% SEM and 95% LOA were the preferred choice of statistics for this reliability study.

Despite the arguments put forward in defence of employing this particular trio of statistical tests, flaws were encountered as the separate sections determining the extent of reliability were investigated. For example, the ICC values for the homogenous groupings in Study 1 were much lower than ICC values for the other reliability studies. This may be because the homogeneity of the sample tends to deflate the magnitude of the reliability coefficient. The ICC calculation is dependent upon the extent of the variation between the subjects' scores in relation to the extent of variation within the subjects' scores; if homogeneity is increased, this will cause the former account of variation to decrease in relation to the latter which remains the same, causing an overall decrease in the coefficient.

Conversely, 95% LOA could serve to magnify the variation about the means; limits can be very wide if sample sizes are below fifty and interpretations should be used in context.

95% SEM can provide a numerical, absolute estimate of acceptable error adding to the interpretability of the results when the footprint measurement method is used for evaluative purposes. This should be an essential requirement if point three of the Daubert ruling concerning error accountability is to be interpreted in the correct scientific manner.

An examination of between-print reliability (section 7.4) illustrated intra-variation was greater for the dynamic prints compared with their static counterparts. This may have been due to environmental or individual variation between each volunteer's three given impressions. Alternatively, lack of measurement reliability may have caused larger variation in the dynamic groups. This theory was perceived to be more plausible as the dynamic images incurred greater subjectivity in terms of measurement procedures. However a following intra-rater and inter-rater reliability study did not support this supposition. In the intra-rater study (section 7.5) the interval bands in the construction of LOA were wider for the dynamic measurements, but the 95% SEM and paired mean differences illustrated greater absolute reliability than the static measurements. Inclusion of heel and toe flare measurement did not therefore affect the reliability of the method. However, greater intra-variation in the dynamic state, though slight, was evident and should be taken into consideration in forensic identification investigations by the inclusion of 95% SEM margins.

The inter-rater test (section 7.6) involved two inexperienced and one experienced rater. Initially it was presumed the experienced rater would display reliability estimates allowing for an interpretation of greater repeatability compared with the others. This notion was proved to be incorrect; higher reliability was deemed not to be associated with experience or expertise, rather conscientious and meticulous construction and measurement with frequent break-taking. Therefore, the method can be considered pragmatic, permitting individuals from different professions and backgrounds to undertake two-dimensional footprint measurement.

Chapter 8
Evaluation of the Reel Method

Research in the modern era is expected to be applied and utilised by the wider community. In an exploration of external validity, this chapter presents an approach that aims to translate science into practice by producing a study that tests the utility of the new measurement approach by practitioners and uses the learning from the experience to improve the measurement method and supporting material.

8.1 Literature review

In order to test the effectiveness of the measurement approach, it was decided that a qualitative research approach should be adopted, as this next section of the research was of an exploratory nature. Hendry reiterates this by explaining that qualitative research seeks to reveal an understanding about a little-known area (Hendry, 2003). Kendra & Taplin (2004) describe research as being of a qualitative nature if it is attempting to understand the experiences of a given situation to a group of individuals, in this case, forensic practitioners.

Testing the utility of the new measurement approach can be referred to as translational science; a term often used in health settings in which novel therapeutic strategies are developed through experimentation in a 'bench to bedside' approach (Marincola, 2003, page 1). This type of research requires extrapolation beyond the controlled laboratory environment to the context of life in the real-world and considers both content validity and overall utility (Sirovatka, 2005).

According to Gustafsson et al. (2004) content validity is where the contents of a tool or test are examined to decide if it actually reflects the area of the content the tool is representing. Bowling (2002) suggests that content validity should also involve the judgements of experts as to the degree to which the content of the test examines the area it is intending to assess. Overall utility refers to the usability of a tool in its related setting (Law et al., 2000). The researcher looked to the literature to provide guidance as to an appropriate design, methodology

and analytical procedure that established content validity and utility, and thus could be used as a model for evaluating the approach and supporting material.

The lack of a forensic database or a systematic review centre pertaining to forensic science equivalent to, for example, the Cochrane Collaboration, directed the researcher to search for publications held in clinical databases. Using the databases AMED, CINAHL and MEDLINE, the following search terms were entered; 'tool* OR system* OR guide AND evaluat* OR assessment AND survey AND interview* AND user*', eliciting sixteen results. Of these, nine articles were of particular relevance in that they involved the evaluation of a newly developed tool or test, using a qualitative approach in which the opinions of users were gathered and analysed. These articles were critically appraised using the guidelines developed for qualitative studies by Letts et al. (2007). The levels of evidence based on these relevant articles were also evaluated using the system advocated by the Association of Women's Health, Obstetric and Neonatal Nurses. The quality of evidence ratings ranged from 22.5 – 30 (good), 15 – 22.4 (fair) and less than 15 (poor) (Cesario et al., 2002). A summary of this scoring system is illustrated in Appendix B.3. Articles scoring a total of fifteen or greater using the Association of Women's Health, Obstetric and Neonatal Nurses system (fair to good) elicited four studies which are discussed in more detail next.

Pagliari et al. (2003) developed an online diabetes management tool. The aim of the study was to evaluate its utility at the early stages for refining and further improvement. This was considered a relevant study for the development of an evaluation procedure for the footprint measurement approach, as there were similarities with aims and end users. Thirty eight members of staff from five general practices were sent questionnaires by email. This first part of the process identified nine key respondents consisting of four general practitioners, three nurses and two administrators. This mix of professionals was used in the hope of gaining a wider insight to the evaluation of the online management system. The chosen key respondents were questioned as to their opinions of the prototype web-based resource regarding its usability. This was achieved by using semi-structured, one-to-one interviews at the participant's place of work.

The interviews lasted approximately forty minutes each. The study did not report how the responses were collected for this part of the research, but the responses subsequently underwent a content analysis. The evaluations from the initial questionnaires, the semi-structured interviews from the key informants and on-going online evaluations resulted in the improvement and refinement of the utility and content of the web-based resource. The authors reported the resource has since been rolled out to seventy four practices in the local area.

Part of the aims of 'Developing a treatment manual for attention management in chronic pain' (Morley et al., 2004) was to consider experts' opinions regarding a newly developed manual which advises on protocols for the therapeutic application of controlling chronic pain. Again similarities were apparent between this and the study presented by the researcher. For example, both studies set out to evaluate the effectiveness of the manual and opinions were sought from experts in the field. In Morley et al.'s work, six experts were identified. These were professionals known to the main author and it is acknowledged that the sample was not representative as a result of this type of sample selection. The participants were first asked to read through the manual and provide written suggestions regarding alternative advice, exercises, potential problems and solutions. Later the experts were questioned through semi-structured interviews given over the telephone. This focused on the aims and structure as well as specific aspects of the treatment prescribed by the manual and lasted between forty five minutes to two hours. The responses were transcribed and were then subjected to a thematic analysis in which six main themes were identified. Using this information, the manual was revised and sent back to the experts to ensure their contribution and views had been captured accordingly.

A reliability and validity study of a clinical assessment tool for palliative care providers investigated the usefulness of the tool in a clinical setting in order to establish its validity (Ho et al., 2008). This paper was of relevance as it bore similarities with the evaluative aims of the research presented in this thesis. Doctors and nurses involved in palliative care were contacted by email and invited to participate in the research. Of the fifty three respondents, fifteen experts were identified for further questioning. These were identified through their amount of experience using the system under investigation (two years or more) and interviewed over the telephone. Prior to the interviews, each expert

was sent an email detailing a set of five questions to be discussed. The interviews lasted between twenty to thirty minutes and were audio-recorded as well as notes taken at the time of the telephone conversation. A content analysis was undertaken by all four authors and five themes were identified, presumably indicative of the five key questions asked of the participants. Results suggested that most experts had already incorporated the system into their daily practice and further modifications were not necessary. Difficulties had arisen, however, in initially learning how to use the system and also ascertaining in-between values on the system's scale. Although these points were acknowledged, the authors did not offer alternative suggestions to nullify these problems.

Evaluation of a new tool was also the main focus of Gustafsson et al.'s study (2004) and therefore considered relevant to the evaluation study in this thesis. The assessment tool examined patients' perception of their manual ability. Also the tool was capable of measuring rehabilitation outcomes and provided the basis for treatment planning. Patient participants were initially invited to rate the tool using a computer program which then converted the ratings into a score for further analysis. Twelve participating occupational therapists were invited to trial the tool with five patients each. Both patients and occupational therapists were asked for their opinions relating to the assessment tool in focus-group discussions for an exploration of content validity and clinical utility. The authors explained that focus-groups were used as this method is recommended by others when a tool is in its developmental stage. For example, the researcher in Gustafsson et al.'s study learned phraseology that users may describe to express their own experiences of using the tool. The discussions were audio-taped and later transcribed. Various themes emerged and were categorised. Another author who had not been involved with the focus-group discussions analysed the data independently. Results suggested that the central concepts of the profession were not totally understood by the occupational therapists. This is not a comment concerning the validity of the tool but on the profession itself. Had the study not used a focus-group methodology, this type of information may not have been assimilated. It may not have been directly relevant to the aims of the study, but instead opened a new channel from which further research may be initiated.

In summary, these four studies all evaluated a newly developed tool by asking the opinions of people in the relevant field who would be using the device. They gathered information from these key participants using semi-structured interviews, either face-to-face or on the telephone. The exception to this was the study by Gustafsson et al. which employed a focus-group method. The results using this method were interesting but digressed from the study's aims. All interviews were audio-taped and transcribed, and emergent themes from a content or thematic analysis helped to improve the tool in some studies.

The concept of the new footprint measurement approach was based on theoretical considerations and was developed from the literature; however the perceived pragmatism of the method had not been tested. Following the methods described in the studies from the literature review above, a determination of content validity was proposed in which the new approach would be evaluated by others.

8.2 Method

A proposal regarding this part of the study was presented to the board of ethics committee at York St John University and was approved (Appendix D.2).

8.2.1 Sample

A group of people with experience of forensic footprint comparison and analysis were considered for this next stage of exploring science to utility. The predefined level of experience was set at a minimum of two footprint identification cases in formal forensic investigation. This small amount reflects the relatively few cases that are currently examined by practitioners in this field. However, one case may involve the examination of many footprints, and a practitioner with experience of two cases will have undergone a substantial number of technique practises beforehand to gain proficiency.

Prior to recruitment, it was envisaged that difficulties would arise in gathering interested participants for the study evaluating the newly developed footprint measurement approach due to the low numbers of individuals working within this field. This was a key factor in the design of the study as it determined an in-

depth rather than a broad data-collection, for example by using e-surveys and postal questionnaires (Patton, 1990).

Formative evaluation requires comment from a specific group of people who would be familiar with the constructs of the new measurement approach developed by the researcher (Patton, 1990). People with experience of forensic footprint comparison and analysis were therefore considered cases for analysis for this next stage of exploring science to utility. Actual numbers of experts in this field in the UK are unknown but considered to be small, consisting of members from the professions of forensic podiatry, forensic scientists, anthropologists and specially trained associates of the police force (Robbins, 1978; Laskowski & Kyle, 1988; Kennedy, 1996; Borkowski, 2002). Additional aims of the evaluation were to discover if the measurement approach would be of use to other researchers or students and therefore practitioners in the field with higher education affiliations were particularly sought after.

The researcher knew of only three experts in the field personally. These possible recruits were of different professional backgrounds comprising of a forensic marks examiner employed by West Yorkshire Police, a forensic podiatrist and a principal lecturer in podiatry who had worked independently with the police on several forensic cases. Heterogeneity of this small group potentially offered meaningful insight for the evaluation study, as it was possible that different footprint comparison and analysis methods were being utilised that had not necessarily developed from the literature. For example, the marks examiner was experienced in other areas of forensic identification such as those dealing with fingerprints, shoe wear marks and tyre tread marks. It was possible that due to daily experience in a variety of different types of examination, methods of footprint evaluation may have evolved from influences from these other areas within her work. The three potential recruits were considered adequately varied in professional discipline, background and expertise, to represent this small community of forensic practitioners who deal with casework involving footprints. However, the researcher was keen to recruit other practitioners to permit data saturation in which the answers from further

participants would not offer additional information (Glaser & Strauss, 1967). Therefore another recruitment strategy, snowball sampling, was considered (Patton 1990). Approaching a number of people at forensic-based conferences and meetings across the UK and also in the US the author was able to ask well-situated people for the names of other suitable candidates for this part of the study. In this process three key names were mentioned repeatedly and therefore took on special importance. Two of these named people had already agreed in principle to partake in the evaluation study.

Information gathered from international meetings and conferences concerning the analysis and comparison of crime scene footprints for identification purposes, confirmed to the researcher that the processes involved were similar in many countries. Attaining the thoughts of overseas experts regarding the researcher's measurement approach was thus considered. However, this idea was rejected as the research was limited by time and financial constraints. The purposive sampling combined with snowball sampling finally produced six key names within the UK.

Experts were finally approached by way of telephone and email and were chosen as they were deemed by the researcher to be information-rich due to the diversity of their backgrounds and expertise. Out of the six experts approached, all responded positively; however one respondent failed to answer upon further contact and was not included in the study. The remaining five consisted of a forensic scientist who also lectures forensic science students, a specially trained member of the police force (marks examiner) who is also a part-time degree student of podiatry, and three forensic podiatrists, two of whom are also lecturers at higher education institutes. It was hoped that the diversity of experience in footprint analysis demonstrated by the chosen sample would increase the credibility, dependability and transferability of the findings from this part of the study (Lincoln & Guba, 1985; Patton, 1990; Strauss & Corbin, 1998; Polit & Hungler, 1999).

8.2.2 Study design

Information sheets, consent forms, and a package enclosing the manual, CD and DVD (section 3.6) were sent by post to the respondents. All volunteers sent copies of signed consent forms back to the researcher and dates were organised for individual interviews, allowing sufficient time for viewing the

material beforehand. Copies of the information sheets and consent forms can be seen in appendices D.5 and D.6.

In the formative stages prior to the interview process, a focus-group consisting of the researcher and supervisory team considered various questions relating to the utilisation of the technique which could potentially produce units of interest. Many of the questions produced by the researcher were later deemed irrelevant as they did not pertain to the manual, CD and DVD (the package). In order to surmount this, questions were organised into groups and those not pertaining to the package were excluded. This resulted in eighty-three surviving questions. The questions were collectively condensed further by organising the questions into groups and finally eighteen key questions were agreed upon by the focus-group. The researcher then piloted these questions with three of her work colleagues. Turner recommends that the pilot test be 'conducted with participants that have similar interests as those that will participate in the implemented study,' (Turner, 2010, page 757). However, this was unachievable due to the low number of UK experts in the field available. Instead the author chose allied health professionals who held degrees at master's level and also had an interest and basic knowledge of forensic podiatry. They had previously attended lectures and meetings on the subject and all had expressed an interest in current research practices in the area of forensic podiatry. The three allied health professionals piloted for the study were each given the package to look at and were interviewed accordingly after consent forms had been returned and the study approved by the board of ethics. Further review of the questions using information gleaned at the piloting stage resulted in reducing the number down to just fourteen with associated prompts (Appendix E.1).

Standardised, open-ended questions were employed as the interviewing technique of choice to gain experts' opinions as to the usefulness, or utility of the measurement approach having reviewed the written manual, the DVD demonstrating a method of collecting static and dynamic footprints and the CD displaying real-time moving screen shots of the measurement software in action. In this design, all interviewees are asked the same questions in the same order, but the responses are open-ended and probing questions can be asked by the researcher at appropriate places (Gall et al., 2003). The advantage of this method is the data can be organised succinctly according to

the ordered series of answers. Interviewer bias is minimised as the same question is asked of each participant and because the interview is systematic, the level of researcher judgement during the interview is decreased. The disadvantages are that by standardising the wording of the questions, the significance and spontaneity of the questions and answers may be constrained (Patton, 1990). Mindful of these disadvantages, the researcher utilised probes and prompting strategies during the interviews to elicit further information (Appendix E.1).

The control of researcher bias was an important aspect for this section of the research. Chew-Graham et al. (2002) found that where the respondent regarded the researcher as a professional peer, rich and intuitive responses were elicited from the interview. However, there remained a risk that this could lead to a contextually shared blindness, allowing the researcher's own feelings and opinions about the field to govern the interview and its subsequent interpretations (Hamberg et al., 1994). The researcher was aware that subconscious desires for the expert participants to respond positively to the work that had been produced during the development of the research, posed a threat to internal validity. Therefore, questions were carefully worded and underwent a series of revisions by the research team before a final set of questions were agreed on. The main researcher was both careful not to introduce bias in the questions but keen to extract as much information from the participant as possible, without fear of the expert upsetting or criticising the researcher/interviewer's work. For example, the researcher was inquisitive as to the experts' thoughts regarding a particular aspect of the measurement technique; wanting to explore their thoughts as fully as possible regarding the method but realising also they may not have wanted to imply criticism towards the researcher's work in their answer. To overcome this, the questions sometimes implied fictitious previously accepted criticism, allowing freedom for in-depth critical evaluation. For example, the following was asked; 'In the measurement method, where you have to draw a line skimming the outer pixels, there's some guesswork involved here. Do you think that's too subjective?' This method of objective questioning by the interviewer is described by Patton as 'conveying that important sense of neutrality' (Patton, 1980, page 317). Additionally, part of the researcher's aims of interviewing the experts was to

refine and become aware of any problems regarding the content of the package and hoped that by questioning the interviewees in this manner, they would feel free to think and comment accordingly. Baxter & Jack comment that this type of questioning 'increases the confidence in the findings, as the number of propositions and rival propositions are addressed and accepted or rejected' (Baxter & Jack, 2008).

Prior to each interview, the expert was assured that although their answers would be tape-recorded and transcribed, their anonymity would be protected. Respondent's names were coded by the researcher only, and tape-recordings and transcriptions were kept in a locked cabinet at the University of York St John, to be destroyed after completion of the research. They were instructed that they could withdraw from the study without explanation at any time. The researcher's contact details were made available to all the experts should they wish to discuss any issues further. The interviews were taped using a portable tape-recorder and microphone and later transcribed word-for-word. Each transcript was sent via email to the corresponding expert for verification and all responded that they were satisfied with the content included in their individual transcribed interview with the researcher.

The interview questions constituted a framework for a thematic analysis to be undertaken which allowed the primary patterns in the transcripts to be identified and categorised (Patton, 1990). Using this type of framework allowed for a cross-case analysis which Patton defines as 'grouping together answers from different people to common questions or analysing different perspectives on central issues,' (page 376). By using the interview questions as a framework and identifying patterns within the responses from the experts, it was anticipated that common themes would emerge.

Themes and emergent patterns were identified as answers that cropped up with recurring regularity, sometimes even cutting across questions. These were organised initially by hand using highlighter pens on the printed copies of the transcripts and later by 'cutting and pasting' as suggested by Patton (1990, page 382).

Additionally, two academics read through the completed transcripts and independently identified themes in the manner described above. The emergent

units highlighted by the three autonomous assessors were then examined and all themes appeared to be concurrent. Seeking agreement amongst co-researchers, participants and experts can be construed as a method of achieving credibility and trustworthiness in the interpretation of the study's findings (Adler & Adler, 1988; Woods & Catanzaro, 1988; Polit & Hungler, 1999).

8.3 Findings

The five experts, described in section 8.2.1 had a combined total of fifty-six years of experience in dealing with footprint evidence for forensic purposes and four had undergraduate and/or postgraduate involvement at higher education institutes at the time of interview. Numbers of forensic footprint cases per year were reported as being few with an average of 2.5 cases although all experts thought that this number has the potential to increase with the advancement of footprint research and training.

8.3.1 Approaches utilised in practice

When asked if they used a specific approach to measure two-dimensional footprints, all answered that they used the overlay method in conjunction with a linear measurement method, most notably that of Gunn's (three respondents). The overlay method, although subjective in nature, offers a pragmatic overview of the general morphology of the footprint in question. The method is especially useful when comparing toe arrangement patterns between exemplar and unknown footprints, and to gain information regarding approximate footprint length and breadth evaluations. It is sometimes used exclusively when comparing partial prints due to the lack of footprint detail which is necessary for most linear measurement methods (Vernon, 2007).

Two experts occasionally used the OCM in addition to the Gunn and overlay methods. They explained that the OCM was achieved by overlying a sheet of acetate over the footprint with concentric circles printed onto it from an image obtained from the internet. The concentric circle pattern would be placed over a toe or heel and the optical centre identified. Another respondent had explained he had tried to use the OCM but had found it too subjective. All respondents explained that the overlay method would be the first choice of analysis.

8.3.2 Methods of footprint collection utilised by the experts

When asked about the inkless paper system, all experts were familiar with and regularly used this method of collecting two-dimensional footprints, describing it as "quite a handy kit", "nice and clean to use" and "portable, easy to carry". When asked of other methods they had experienced for collecting comparison footprints four respondents had tried fingerprinting ink but described this as inconvenient. Three respondents used the word "messy" regarding the inked method. Three respondents had used animal blood in order to achieve a "like-for-like" medium but had been unhappy with ethical issues surrounding this method. Instead, two respondents had used a thick solution of poster paint to achieve a similar substrate.

Another respondent described using oil, gel or soap to try and recreate "slip-marks" and then had dusted the prints with black powder for contrast.

When asked if the experts regarded the inkless paper system to be a valid and useful method of collecting footprints, all responded positively adding comments such as "very neat result", "gives a really clear print", "good reliable detail" and "unfussy, quite simple, doesn't need chemicals or reagents, so yes, it's ideal". However, all the experts discussed the disadvantages of using the inkless paper system, especially for the analysis of comparison footprints in criminal cases. All the experts questioned had issues with the method of collecting the dynamic footprints, specifically with the accurate targeting of the foot on both the inkless pad and the paper. They identified that this method appeared time-consuming and four out of the five experts suggested a long roll of paper containing the developing substrate as a solution, rather than individual A4 sized paper for dynamic footprint capture. Two experts had heard of the use of fax paper that can be purchased as a long roll as being capable of developing the inkless footprints. These comments reflect the previous debate outlined in Chapter 3. Justification of the use of the A4 sized inkless paper used in the study is explored in section 3.2.

8.3.3 Thoughts on the evidence underpinning the new measurement approach

The experts were then asked if they thought that the underpinning evidence behind the approach was useful in setting the scene. All responded positively adding that they found it "interesting", and "useful". Expert #2 thought it was a

"good starting point for real-world application" and Expert #3 said it was a "nice overview of the literature". Experts #1, #3 and #4 were surprised at the criticisms directed at the work of Kennedy et al. and Expert #1 had not previously been aware of the problems regarding validity and reliability within the forensic identification sciences, identified by the NAS report and the House of Commons Science and Technology Committee. Three experts admitted that they found the statistical analyses daunting and two of these experts used the phrase "not comfortable with statistics". However, all were in agreement that those chapters dealing with statistics were important in setting the scene.
Expert #1 said in conjunction with these sections of the manual that she found the chapter on reliability and consistency especially useful.
Expert #4 discussed at length reliability issues in measurement as described in Chapter 2 of the manual. He expressed his agreement with the use of multiple statistical methods to ascertain reliability adding that he had personally encountered problems with the use of the ICC on its own, when interpreting measurement data from research studies.
Referring to the evidence supporting the method of footprint collection in the manual, Expert #2 commented that in practice it was advisable to collect more than one footprint, adding that he would collect up to six prints from a suspect, depending on how cooperative the person was at the time. Of Chapter 4, Expert #4 added:

> "I think there has to be a standardised way of how we collect footprints and I think there has to be a standard way of how we're going to measure them. There has to be an accepted method of collecting footprints."

8.3.4 Students

The researcher then asked if the experts thought the approach which included the manual, the DVD and the CD would be useful for students. All responded positively using words such as "...it would be very useful for students" and "...incredibly useful for students". Three experts answered positively in terms of using students' footprints in order to increase databases for future research. Further prompting produced comments regarding forensic science trainees. Expert #2 said:

"Well at the moment, such a guide doesn't actually exist and there is urgent need for protocols and guides at the moment, so it would be invaluable for trainees."

8.3.5 Pragmatism

The researcher then asked if each expert believed that the written guide and guide to collecting footprints DVD would enable them to collect both static and dynamic footprints equally well. All of the interviewed experts responded positively using words to describe the guides as "clear", "helpful", "user-friendly", "nicely set out", and "highly detailed". However, common themes occurred regarding the method of dynamic footprint collection in terms of individuals obtaining a natural gait and also problems encountered in real-life scenarios in terms of lack of space.

Experts #2 and #3 both agreed that the availability of a five-metre walkway in which to practice walking up and down as advised in the guide is unrealistic when collecting footprints for forensic identification purposes. Expert #2 commented:

"...you've got the practical thing that often when you're doing this in practice, rather than having the luxury of a hallway, so you've got a nice 5 metre walkway, you may be stuck to a small backroom in a solicitor's office that's about 12 feet long, so it would be helpful to have something there that if the ideal isn't there, here's what you'd be looking for, as the ideal's rarely there, in my experience."

When asked if the method described for measuring footprints seemed overly-complicated, all experts answered that they did not agree with this statement. All concurred that they approved of the two different learning approaches available in the form of the written guide and the real-time visual CD. Three experts compared their experiences of using Adobe PhotoShop® for digitally measuring footprints compared with what they had viewed and further experienced using GIMP. These experts believed GIMP to be easier software to manipulate than PhotoShop® and therefore quicker to use. When further questioned and asked if they thought it possible that the CD on its own would be enough to instruct someone on how to evaluate footprints (without the written guide), all experts replied positively but added that there was much

value to having both mediums from which to learn from, to support different learning styles. All the experts said given the choice, they would prefer to learn visually from the CD, but also have the written guide available for reference.

The researcher focussed the experts' attention to the detail of the two-dimensional measurement method. They were asked whether they thought the selection of certain pixels of the footprint image as points of reference were too subjective. Experts #1 and #4 agreed that they thought the method was too subjective, the latter wondering also if flattening the image, the use of JPEG rather than TIF, screen resolution and rotating the image might also affect reliability. Experts #2 and #3 said that this part of the measurement process appeared to be subjective, but in practice did not present an issue:

> "What I've found is that as you get experienced at this work the subjectivity tends to go, so I think the subjectivity comes within experience." Expert #2

There appear to be contradictions within this response since subjectivity will remain due to the nature of this part of the method, irrespective of experience. The respondent clarified the point by explaining that he had noted that with greater experience of the method, the differences between repeated measurement results of the same footprint became smaller.

Expert #5 did not agree that the method was too subjective and supported Expert #2's comment regarding that the measurements were obtained in millimetre values rather than pixels. This expert suggested investigating inter-rater reliability of the measurement approach by using thirty volunteers to measure the same footprint length and analysing the differences between results.

All five experts expressed an interest in using the measurement approach in their line of work.

8.3.6 Measurement of ghosting/flaring

'Ghosting' or 'flaring' of the inked footprints was noted to occur around the distal toe areas and the proximal heel print area, especially in the dynamic footprints. The approach instructs the user to include this ghosting in all measurements. However, with most prints of this nature, it is possible to subjectively determine the apparent demarcation line of the actual footprint and the point at which the

foot starts to 'roll' onto (heel) or off (toes) the inkless paper. The experts were asked whether they thought the inclusion of the ghosted area would hinder or help the analysis. The reactions to this question were mixed and related to personal experiences reflecting the way they would approach this situation in practice. Expert #1 commented:

> "From a forensics point of view, if someone's walking and that part of the foot touches the floor when they're making a normal dynamic impression at a crime scene, then you've got to include that bit when you're then taking an impression. Also it then becomes subjective if you discount it, as to where the foot is actually ending because that's just someone's assumption of where the toe is ending, just because of the differences in the ink that's been left."

However, Experts #2 and #3 disagreed:

> "I would always use both measurements but place my reliance on that inner dark measurement because in my experience I've found that that appears to be the true border of the plantigrade foot. The ghosting areas seem to be formed by function. So as you get the ghosting of the heel, that's where the heel sort of moved into its ground contact, but not actually the true picture of that ground contact, and the same with the tips of the toes – they represent where the foot's actually started to move off the normal plantigrade surface during toe-off, and also where the toes appear to have, in effect, scuffed slightly against the ground. Now if those are consistent across the known and unknown, then that's an important factor for comparison, but if it's the true barefoot impression then I would always go for those darker inner lines." (Expert #3)

Expert #1 also suggested taking several dynamic prints and if all of these demonstrated similar amounts of ghosting, then this should be included in the measurements. Experts #2 and #3 said that in practice, they would always record sets of measurement results that both included and discounted ghosting. It was noted that Experts #2 and #3 were the most experienced of the group in the practise of forensic bare footprint analysis and comparison, and therefore this comment held much weight.

Data gathered from the footprint image measurements recorded differences of a maximum value of 9.6mm between the demarcation line of heavier impression in the large toe print and the distal pixel of the flared area. In a further reliability study, intra-rater tests involving the sixty-one participants resulted in high reliability between three measured dynamic footprint lengths collected from each volunteer (Chapter 7, section 7.4). In other words, there appeared to be no statistically significant differences in the amount of flaring produced by a person when the subject created a set of three separate dynamic footprints. This suggests flaring should be accounted for when measuring and comparing dynamic footprints, but measuring the footprint both with and without flaring would probably be best practice, as recommended by Experts #2 and #3. There remains a limited understanding of dynamic print flaring regarding definition, causes and repeatability, and is an area requiring further research.

8.3.7 Partial footprints

When asked if they thought the measurement approach could be used for partial footprints, the experts' answers were largely sceptical. Most argued that partial footprints are usually in the form of a forefoot print. They had identified that the point at the base of the heel print is crucial in determining central alignment of the footprint and also for taking length measurements and therefore the measurement method cannot be used for heelless footprints. Experts #2 and #5 thought measuring from a different 'reference' point such as the medial side of the widest part of the forefoot would allow sufficient measurement reliability. Their suggestions have been reported previously in the literature (Gunn, 1991, Vernon, 2007) and although the measurement approach in the manual can be manipulated to allow for the measurement of partial footprints, reliability and validity of same has not been tested.

8.3.8 Contribution of the new approach to the literature

The final question explored the experts' views regarding the potential for the approach to contribute to the literature. All answered positively, with additional comments such as: "You've got a reliable and valid measurement tool. Everybody in this area should use it"; "It's an evidence-based protocol which is going to be useful"; "We need a sound scientific base – like the NAS report has recommended"; "The current methods haven't been established as reliable or

valid. It's not a question of how many footprints were collected for a study, it's all about how it was statistically proven to be reliable and valid"; and "You need a method that's robust and could be used in day-to-day examinations. It's not just a research tool, it has practical applications also. The more it gets used, recognised, it can only add value to that group of comparisons."

A pertinent question came from Experts #2 and #3, who queried who the end users would be, i.e. the purpose of the manual. For example, a conversation regarding how many prints of a suspect should be taken for further analysis and whether this should be included in the manual, prompted Expert #3 to add:

> "You say what triggered your interest in doing this kind of work but you are doing a PhD study and to me it's just setting the context – is this just the way you do it for your collection of prints and interpretation of prints of your study or would it lead on to this being the way that people start to analyse prints in practice? A users guide in practice would be slightly different from a 'here's how I measure prints for the purpose of my study".

8.4 Discussion

The study was completed as planned and the interviews contributed to the understanding of how to best position the tool for translation into practice.

The answers offered by the experts allowed not only for the exploration of various themes, but also an insight into the academic thinking of the participants' answers. For example, Expert #5 suggested a further inter-rater study involving thirty practitioners. He suggested that in this study the participants would be supplied with the same scanned footprint image and asked to complete one measurement and record the result in millimetres for a reliability analysis using the approach. This expert felt he would have greater confidence in the method if these results produced statistically high reliability and compared testing in this manner with other impression evidence testing carried out such as for finger marks and shoe wear marks. An inter-reliability test of this nature had been considered previously but excluded from the PhD study due to time constraints. This participant's thought processes regarding reliability issues are an example of the high academic calibre noted in all

interviewees and supports the choice of experts in the field used for this validation study.

However, an expert in forensic practice does not solely rely on academic ability. Good practical thinking was demonstrated when the researcher was queried by nearly all experts regarding the collection of dynamic footprints. For example, the experts commented on envisaged practical problems in terms of taking dynamic footprints from a suspect at a police station or solicitor's office, despite the manual not specifying who the prints should be taken from, or where.

This leads to the central discussion point brought up several times by the experts during the interviews as to the intended purpose of the package. Questions asked were often answered with, "It depends who will be using this..." and in these cases the researcher pressed the participants to give answers to include all possible end-users. The hope was that the package could be later adapted or modified to suit all, including under- to post- graduate students, police trainees, and forensic practitioners (podiatrists and marks examiners). In using a variety of experts from all the different fields reflecting these end-users, the researcher anticipated that the interviewees would view the package as a standard method for collecting and measuring footprints and be able to foresee adaptation for different usage. It became clear, however, that the manual was orientated too much around the research with no guidance as to how it could be modified for practical use, or for student use, or for future research. For example, suggestions were made by the experts to include practical advice regarding the collection of footprints from non-compliant suspects. This type of detail was not originally included in the manual, but based on the information offered by the respondents, may be amended for further use.

All the participants agreed that the approach offered a pragmatic alternative to other methods used in the field and were willing to trial the method, if they had not already done so. Another common theme was the suitability of the package and approach for use by students at differing levels of higher education. Explanations of the evidence underpinning the method within the written manual were considered to be helpful by all those interviewed. Finally, the experts agreed that the approach contributed to the literature in this area. The

commonality of themes produced by the participants' answers adds credibility to the validation study, especially as the sample of experts was heterogeneous in their expertise and commented on the questions from different aspects (Graneheim & Lundman, 2004).

An application of the Reel method would be the utilisation of this baseline approach as a foundation for further research examining the variability of a person's footprint. In March 2012, the researcher was asked to present the approach and its findings in a panel interview investigating the potential for the Reel method to be taken up wider. The panel consisting of forensic academics from Staffordshire University are proposing the creation and development of a national footprint database to be piloted in 2013 and to be fully operational by the following year. The Reel approach was duly accepted by the panel, deemed the only approach available that offers enough rigour to be used for this purpose. Collection and measurement will be carried out by various people with limited experience, starting with students. It is envisaged that the pragmatism of the approach identified by the interviewed experts and the various teaching styles offered by the written manual, CD and DVD, will facilitate the learning process.

8.5 Conclusions

The package detailing the researcher's measurement approach examined by the experts was declared to be robust and pragmatic in their opinion. All agreed that in the absence of a standardised footprint measurement method, the new approach was rigorous enough to fill that gap. This supports external validity of the approach.

A potential future development will involve modification of the package to include practical advice that would normally be expected of a field-guide.

Chapter 9
Synthesis of Research Elements

The primary aims of this thesis were to critically review the literature around best practice measurement approaches and to develop a scientifically rigorous footprint measurement approach. In the course of exploring these aims and within the various studies carried out, certain findings took on more meaning when placed into context with the research in its entirety and will be discussed next.

9.1 The parallels, gaps and reconciliation between forensic science and medicine

Throughout the thesis, the researcher has drawn parallels with forensic science and medicine. This is evident in the discussions that create a backdrop for many of the chapters. The literature reviewed for the purpose of informing the research methodology mostly followed the traditional positivist philosophy of exploring science, and the majority of the analyses presented in the thesis are also entrenched in this scientific approach. Initially, the researcher had assumed that the implementation of forensic science can be likened to that of medicine in which the principles of evidence-based practice (EBP) are applied. In this context, methods employed to assess forensic evidence including footprints, aspire to be founded upon the most current, peer-reviewed and scientifically robust research findings. However, as acknowledged in Chapter 2, examples of EBP used in the field of the identification forensic sciences were found to be limited. The best available evidence in some cases was based on practitioner expertise. Searches of forensic science journals using key terms related to specific types of validity such as predictive validity resulted in few articles of relevance. Content-, construct- and criterion-related validities that were addressed in the thesis in order to establish rigour of the Reel method have their concepts rooted in scientific positivism. These types of validity are specifically associated with exploring and defining measurement rigour, an essential component in both disciplines. It can be argued that measurement in medicine is grounded in the social sciences adept in addressing human

variability; however forensic science is not devoid of research involving the measurement of human variation and therefore it is conceivable for the two disciplines to be intertwined.

The difference between medicine-based research and forensic identification research mainly lies in the extent of scientific rigour employed to explore theories and hypotheses. Although efforts to examine measurement rigour in forensic identification science is not apparent in the research literature, a recent review by the UK Government of research and development relevant to forensic science has identified the need for standardised guidelines for forensic practitioners to ensure the quality of the forensic science provided (Silverman, 2011). The author of the review, advocates that these missing standards must be established, underpinned by commissioning 'high quality accessible reviews of the current scientific position of relevant forensic methods' (page 11). This mirrors the EBP principles advocated in medicine, for example the promulgation of the NICE guidelines in clinical practice.

Since the researcher's initial search of the relevant literature in 2006, the gaps first noted in forensic identification science regarding research and application of relevant findings to practice, appear to not only have been formally identified but are perhaps on the brink of being bridged. Adoption of designs and methodologies entrenched in measurement science for use in this thesis is not only justified, but is advocated for similar forensic science research.

9.2 Challenges to forensic footprint interpretation

In the Introduction and Critical Review of the Literature sections of the thesis, wrongful convictions due to unsafe footprint evidence were discussed as exemplified in the cases involving expert witnesses such as Kennedy and Robbins (Hansen, 1993; McRoberts et al., 2004). These miscarriages of justice demonstrated a fundamental lack of scientific rigour of the methods employed from the onset. Evaluative approaches used in these cases did not demonstrate enough scientific rigour to be upheld in a court of law. To make matters worse, inconsistencies seemed to exist in defining the elements of scientific rigour between the areas of law, research and forensic practice. This point was acknowledged by the Silverman review which identified on page 10 'that there is

often a communication gap (and even a philosophical difference) between scientists and forensic providers who develop forensic methods, and judges and juries who deal with evidence based on such methods' (Silverman, 2011). Silverman postulates whether primers to 'set out the science in reasonable text-book style, including a glossary of terms,' would be helpful (page 10). The review therefore not only identifies the confusion that exists regarding the terminologies associated with scientific rigour such as 'validity' and 'reliability', it is actively making plans to address the problem.

Bare footprint evidence is detected and recovered at crime scenes; a prominent example courting much media interest involved the murder of Meredith Kercher in Perugia, Italy in 2007 (Falconi, 2009). The interpretation of footprint evidence in the press coverage of this case and others showed no indication of a scientific evaluation or procedure being undertaken in the process of identification. Instead, the evaluation of evidence had relied on descriptive comparisons observed by experts between exemplar and unknown prints, according to press reports (Gibson, 1986; Rupert, 2004; Falconi, 2009). For clarification, the researcher had looked to published articles to seek evidence of an existing method of high quality. The critical appraisal of the literature uncovered several approaches to the interpretation of footprints in a forensic identification context. The papers were graded in a hierarchy of evidence as to their rigour, particularly in demonstrating reliability and validity, but none presented the extent of scientific rigour expected for this type of evidence. This lack of rigour could not be reconciled in the area of policy either. Despite inconsistencies and confusion over terminologies, the recommendations of the US Daubert ruling (1992), the NAS report (National Research Council, 2009) and the UK's Law Commission report (2011) regarding admissibility of reliable and valid forensic methods in court is pertinent; the current footprint methods of footprint evaluation for identification purposes would not stand up to these recommendations.

9.3 Development and testing of the new footprint measurement method

The researcher therefore proposed a new approach for footprint evaluation. The second aim of the thesis, to develop this method and ensure high levels of

validity and reliability through appropriate testing was thus embarked upon. As the definitions of validity and reliability were the most comprehensive within the area of scientific research (compared with the understandings from the fields of forensics and law), these were selected as templates for the investigation of scientific rigour.

In the process of creating the measurement approach, it was necessary to identify the pertinent operational definitions. For example, the measurements were grounded in the existing literature (the best available evidence) and were developed to enable exploration of the perceived variability of the human footprint, resulting in specific width, length and angle measurements. The method of footprint collection and image storage for the research study also evolved from the appropriate literature and additional experimental studies had to be conducted to justify the chosen method and storage format. The choice of measurement software, the GIMP, had not previously been described in measurement literature pertaining to footprints, or indeed any other form of measurement in forensic identification. However, utilisation of open source software as opposed to closed sourced software such as Adobe Photoshop® or AutoCAD was argued in Chapter 3 to be more appropriate in this instance.

The now newly developed pragmatic approach was employed in the measurement of three hundred and sixty six right footprint images from sixty one volunteers, in a bid to determine and establish the extent of its validity and reliability estimates. Although appropriate for the understanding of footprints at a baseline level, an appropriate type of sample to reflect the nature of the research in a forensic context would have involved collecting footprints from the feet of convicted criminals. This was not achievable, and unlike most research studies, particularly those in the field of medicine, direct inferences to the wider population have yet to be explored. This can only be accomplished as data is gradually collected whilst the method is used in forensic practice. The challenge to this body of work is that the research sets out to legitimise the measurement approach without an appropriate sample. Although a scientifically rigorous method has been implemented, the final test of validation cannot be performed by any one researcher but will require practitioners to implement the approach in the real world. A limitation to using the method in the field is that the concept has not been tested using partial prints and is suitable for complete footprints

only. The challenges to the utilisation of the method in forensic practice were further explored by approaching experts to collate their opinions.

In an evaluative process in which external validity was established, the experts assessed the method of footprint collection and measurement by way of a package detailing the practical aspects of the approach, which also included a written commentary discussing reliability estimates ascertained by the research. These experts were interviewed to allow them to discuss their opinions regarding the approach, which they termed 'the Reel method'. The package consisted of a written manual, a CD explaining footprint measurements using the GIMP software and a DVD demonstrating a method of footprint collection. The conclusions of the experts were favourable in that they all agreed that in the absence of a standardised footprint measurement approach, the Reel method was pragmatic and rigorous enough to fill that gap. The collected opinions of the chosen experts established evidence of external validation of the measurement approach and also the method of footprint collection. Unlike other chapters that deal with the analysis of footprint data, the level of validity in this instance was not quantified using statistical testing, as this part of the research adopted an anti-positivist philosophy. Instead, the methods utilised to obtain relevant information as to the usefulness of the technique, were carried out in a qualitative, transparent and credible manner. For example, independent agreement amongst both co-researchers and the expert participants was sought as to the recognition of the main emergent themes.

Since completing the research studies required by this thesis, the package used as a vehicle to initially establish external validity, has been requested by forensic science practitioners in the US, UK, Canada, China and The Netherlands. It has also been used as an aid to facilitate learning on a forensic podiatry M level module at Huddersfield University in the UK, and also for the collection and measurement of footprints for a national database at Staffordshire University, thus demonstrating a duality of use. In support of the current relevance of the package, the latest publication from the UK's Home Office-appointed forensic science regulator stipulates that 'The forensic practitioner shall have available a library of documents relevant to the authorisation of the new method through validation or verification including any associated supporting material, such as academic papers or technical reports

that were used to support or provide evidence on the applicability of the method' (Rennison, 2011, page 32). It adds, 'Where the method implements a scientific theory/model or an interpretation or evaluation model, the library should include a record of information supporting the use of the theory/model' (Rennison, 2011, page 32). The package produced for the remit of the research appears to be the only validated material in existence to support footprint collection and measurement methods. It is anticipated that the package will now be further requested by practitioners working in the field of footprint identification to fulfil the requirements of the forensic science regulator as part of their validation library. The researcher intends to modify the written manual included in the package to include practical advice that would normally be expected from a field-guide.

The thesis then presents an exploration of different types of validity of the Reel method, and an investigation as to its reliability. The order of these individual analyses was determined not chronologically in terms of the time-span of the research, nor by importance, but by a desire to logically unravel the concepts of validity and reliability beginning with its development through to evaluation by potential users.

Reliability not only estimates error, but also sets the boundaries for decision-making regarding acceptable error. Validity reflects the extent of relationships and prediction. As discussed in the Critical Review of the Literature, if a measurement or test is not reliable, it cannot be valid. Conversely, for a measurement or test to be wholly valid, it must also possess reliability. The relationship between these two basic measurement concepts dictates that they are separate and not interchangeable, yet a presence of one concept without the other negates the definition of scientific rigour. It is therefore difficult to determine which of the two concepts incur most importance and which should gain priority in terms of the structure of the thesis. The development of the method which explored content validity supporting the new approach was the starting point from which to assess scientific rigour. Since this process determined the extent of content validity, it was logical to then explore other relevant types of validity in order to gain continuity. It is for this reason only, that

the chapter detailing reliability concepts of the Reel method is discussed after, as opposed to before, validity. In support of this non-preferential notion, a synthesis of research findings relating to reliability estimates will now be considered first, followed by a discussion of the findings relating to validity concepts.

9.4 Synthesis of the research findings regarding reliability and societal implications

The relevant literature suggested that not only were there different study designs but also various statistical methods available for determining the extent of reliability of a test or measurement (Atkinson & Nevill, 1998; Bruton et al., 2000). Thorough reliability testing should encompass as many of these tests as possible to provide clarity as to the boundaries of acceptable error. In this thesis, the degree of between-print reliability, intra-rater reliability and inter-rater reliability was established using a variety of statistical tests such as ICC, 95% LOA and 95% SEM. The research showed that these decisions based on a single approach are not infallible. Discrepancies and inconsistencies in the results were apparent in all the repeated measures studies and only by further testing using other statistical methods could a comprehensive picture of reliability boundaries be provided. For example, some of the ICC calculations implied perfect correlation with values of 1.0; however 95% LOA and SEM illustrated the existence of variation in all studies. It confirmed that the ICC gives the context for error in respect of group variation (between people variation in the test group) whereas LOA is useful for identifying any bias or outliers between tests, and also visually demonstrates the relationship between the size of the mean and the variance in each sets of measurements. Numerically LOA and SEM provided an estimate of acceptable error when using this footprint measurement method.

It is difficult to define accuracy in the context of this thesis as accuracy usually compares the repeatability of scores of the new method with those from a gold standard method. As reliability scores from a standard method do not exist, accuracy in its true sense cannot be determined for the Reel approach; analysis of data using the method on large populations are required to do this and it is therefore left for future research to establish this outcome. The boundaries to

determine accuracy have yet to be defined. These boundaries may come in the form of the SEM. For example, if repeated test scores using the large population samples fall within 95% SEM values, there may be evidence of accuracy. At present, the research is yet to offer accuracy of the approach, but aspires to this in the future. The determination of accuracy not only involves estimates of reliability, but also of validity. Therefore it could be said that the ultimate definition of rigour can be measured by accuracy.

Although significant differences were detected between static and dynamic length measurements in the repeated measures study, little variation of the measurements between the footprints from each subject was shown when split into their homogenous groupings. LOA graphs illustrated the presence of several outliers, which was attributed to the 'people factor'. Even when split into male/female groupings to improve homogeneity of the sample, wider interval bands were observed for the dynamic footprint measurements compared with their static counterparts. In terms of reliability, this indicated that there was greater variation in the dynamic footprints than the static ones, despite high ICCs for all groupings. The interval bands were noted to be even wider for the female group's dynamic footprints, explained by the fact that the female subjects weighed less on average than the male subjects. The female feet therefore made lighter, less defined impressions, the outlines of which were more difficult to select for measurement.

It was argued that the construction of measurements on the scanned footprint images when toe and heel 'flare' was involved, jeopardised the typically small margins of variation between repeated tests, because the subjective decision as to the start and end pixel from which to execute measurements, incurred more error. The inclusion of measured flare at the heel and toe prints may be regarded as controversial. When interviewed, two of the experts in the field chosen for the evaluative section of the research explained that measurement of the inner darker areas that excluded the flare was preferred. In practice, the researcher found the outer edges of the flaring to be more visually distinct than the inner darker print of the toes and heel; therefore choosing the former as the start/end point for measurement. Anecdotally, this protocol was supported by the opinions of the physiotherapy students who carried out footprint

measurements as part of the inter-rater study. This divergence of opinion may reflect a difference of experiences in footprint evaluation between the students and the researcher, and the practitioners experienced in actual case-work.

The second reliability study established error estimates of the Reel method by way of an intra-rater test. Length, width and angle measurements demonstrated near-perfect ICC values and small interval bands when graphs of 95% LOA were constructed. Means of the paired differences of the measurements resulted in values close to 0.0, indicating high agreement between the repeated tests and 95% SEM intervals were all within ±1.41mm for the linear measurements.

The final inter-rater reliability test illustrated the importance of meticulous and careful measurement-taking. The most experienced rater did not demonstrate a greater repeatability or consistency of measurements between repeated tests when compared with the other two inexperienced raters. When the methods of practice were discussed between the three raters, it became evident that the adopted approaches differed from one to the other. For example, one rater took a minimal time to complete the task in hand, compared to another who took a considerably longer time practicing and perfecting the technique before initialising the recording of measurements. This was identified and influenced the practical advice given in the written manual of the package regarding the optimum approach for undertaking the construction and measurement of two-dimensional foot impressions; to practice the method beforehand and frequent break-taking. Despite these differences between the raters' reliability results, overall error estimates were small. ICC calculations resulted in near-perfect values, 95% LOA interval bands were all within ±1.0 and 95% SEM values fell within ±0.86mm for repeated measurements from all three raters. High reliability estimates for both inexperienced raters illustrates the pragmatism of the approach.

A high degree of intra-rater and inter-rater reliability of the measuring method employed has been established and the extent of error to expect has been determined. Using a combination of all the modern statistical approaches the Reel method is comprehensively supported as being reliable. An article

'Reliability of a two-dimensional footprint measurement approach' (Reel et al., 2010), was published in 'Science and Justice'. The article has been used recently in a pre-trial hearing for a triple murder case in Australia as evidence of the existence of a reliable footprint measurement method (Q v Sica, 2011).

9.5 Synthesis of the research findings regarding validity and societal implications

It has been argued throughout this thesis that validity is an essential requirement to ascertain rigour in respect of the new measurement approach. Despite the conflict and confusion that may lie between the three areas of forensic practice, law and scientific research, all camps agree that without evidence of validity, a test could not be used for forensic measurement. What remains controversial within the three disciplines is the type of validity required, and the methods employed to establish such validity. Much of the published forensic footprint identification literature reported case-studies, demonstrating face validity only (section 2.4.1). The thesis explored other more comprehensive types of validity, which would be suitable for determining the validity of a test, or measurement. These came in the form of content validity, in which the Reel method was developed and later evaluated using the opinions of experts, construct validity which established discriminant and convergent validity and finally criterion-related validity in the forms of predictive and concurrent validity.

The critical appraisal of the literature inferred that differences and similarities between the measurement data would occur when investigating certain variables such as weight, age and height (section 4.2). Predicted differences such as those between static and dynamic and between male and female footprint measurements were unsurprising in that length and width measurements were more informative than the angle measures in this exploration of validity. Statistically significant static and dynamic differences were not displayed for the MPJ width measurement and this was deemed noteworthy. That the width across the forefoot does not vary between the states of standing and walking demonstrates a stable measure which may have further implications for forensic footprint research.

Of non-statistical significance was the impact of high BMI of a subject on footprint dimensions. This was also true for age and ethnicity when these variables were analysed within the footprint measurements. In the future, more specific studies analysing these variables with a more appropriate sample could give rise to a more informative determination of discriminant and convergent validity.

Height was strongly and positively correlated with the paired length and width differences between static and dynamic states and also in the simpler bivariate correlations, as expected. What was not expected was that the highest correlation occurred between height and the Calc_A5 measurement.

This high association between the heel to small toe footprint length and height was further examined through the establishment of predictive validity. In this determination of criterion-related validity, the scores from these two variables were analysed to see if their relationship could be quantified. The Calc_A5 length measurement proved to be the best overall predictor for the stature variable for the analysis of the dynamic prints and displayed an associated coefficient of determination of 0.74. In other words, 74% of the variation of stature is attributable to the Calc_A5 length measurement, supporting the strong relationship between stature and this particular footprint dimension. Split into their homogenous groupings of males, females, static prints and dynamic prints, the highest correlations were strongest for the Calc_A4 and Calc_A5 lengths in the static prints. These correlations were even higher for the dynamic prints and the lateral border of the foot was considered to be a more stable measure compared with the medial border, since the lateral border is devoid of the effects of variation from the medial longitudinal arch. Complimentary regression equations determined that in the absence of all other information, for example sex, age and ethnicity, a regression formula used to calculate height using the Calc_A5 dynamic footprint length measurement could theoretically give rise to the donor's stature within the margins of the length of a credit card (±4.17cm). This has strong practical implications in the field of forensic podiatry as it could aid in the association or disassociation of a person with a footprint. For example, if a mark from a walking foot is left at a crime scene and a suspect happens to have a stature falling within ±4.17cm in accordance with the applied regression formula, there would be a 68% certainty that the donor of the

unknown footprint would be of the suspect's height. These odds could then be factored into a likelihood ratio, along with other factors, to aid in the construction of a case (Redmayne, 2001). The findings from this part of the research have since been disseminated to the forensic community (Reel et al., 2012). It is the only study to date to examine stature prediction from the footprint dimensions of both static and dynamic prints.

Although an existing 'gold standard' approach to footprint measurement was not available to forensic practitioners, the current literature and anecdotal evidence confirmed two popular approaches for quantitative measurement in this field, namely the Optical Center and the Gunn methods. In order to investigate evidence of criterion-related validity, these two methods accepted in the relevant community were compared with the Reel method in terms of validity and reliability. Additionally, the Kennedy approach was included since this method incorporated a linear calculation of the centre of the heel (akin to the Reel method) plus the optical centres of the toe prints. Strong positive PPM correlation coefficients were observed from 0.886 (OCM) to 0.999 (Gunn). This result was unsurprising as all methods were based on similar constructs; however the OCM and the Kennedy methods used different software (AutoCAD) to the Reel method and different methods of line construction. High correlation in this context established criterion-related validity.

A further investigation of reliability was carried out in support of concurrent validity of the approach. The three methods used in the field (Gunn, Kennedy and OCM) were each subjected to an intra-rater repeated measures design study in which one linear measurement from the areas of the heel to large toe were constructed and measured from thirty footprint images picked at random. Results from the repeated tests were collated and analysed along with those of the Reel method. Although all revealed substantial ICC values, interval boundaries for upper and lower limits of agreement were varied, the largest seen in relation to the OCM (-4.756 to +8.678). The inference of the 95% SEM values is that for repeated tests, the measurements would fall within plus or minus *x* mm, 95% of the time. SEM values in this context demonstrated the Reel method had the smallest error variance (±0.10mm). The Kennedy method

displayed 95% SEM values of ±2.23mm. Interestingly, it was Kennedy and his co-workers who suggested an apparently arbitrary 5mm cut-off point (±2.50) in the consideration of the likelihood of a chance match when footprints from their database were analysed (Kennedy et al., 2003). In Kennedy et al.'s study, measurements from two footprints that fell within these limits implied that the footprints belonged to the same person. Reliability estimates from repeated measures of footprints from the same person in the research study using the Kennedy method, determined similar limits of measurement variance.

Strong correlations and high coefficients of determination combined with the highest reliability estimates in a comparison of all methods examined, support the Reel method as an alternative viable and valid measurement approach. Pragmatism of the Reel method over the other community-favoured approaches is considered a further advantage.

9.6 The proclamation of a new, valid and reliable method and subsequent reaction in the field

Having undergone these tests, the Reel method could now be said to possess rigour fulfilling initial requirements stated by the likes of the NAS report (National Research Council, 2009) and the Law Commission Report (2011) regarding the admissibility of a forensic method in a court of law, within the contexts examined in the research. It also complies with the recent requirements of the Codes of Practice and Conduct for Forensic Science Providers and Practitioners (Rennison, 2011).

Complying with the recommendations of the US and UK policies have appeared unpopular with key players in the field of forensic examination, particularly the footwear examiners in the US who appeared to have misunderstood the definitions of validity and reliability and the recommended changes of practice relating to these. An example of this can be seen in the published response to the NAS report from the International Association for Identification (Garrett, 2009). Replying to the challenge regarding the unreliability of friction ridge analysis (fingerprint identification), the response was as follows, 'There is no research to suggest that properly trained and professionally guided examiners cannot reliably identify whole or partial fingerprint impressions to the person from whom they originated' (page 1).

The author of the response appears to believe that reliability is improved by minimising operator bias only. This is in spite of the discussion in the NAS report pertaining to the value of research involving repeated measures tests in order to gain information regarding the reliability of the technique involved in the analysis. The report states that at present there is little information regarding the reliability of fingerprint analysis, relying solely 'on subjective judgments by the examiner' (National Research Council, 2009, page 139).

Barriers to early adoption of the recommendations of the NAS report were also evident at first hand by the researcher. The researcher's article 'Reliability of a two-dimensional measurement approach' (Reel et al., 2010) was reviewed and initially rejected. The reviewer was disparaging of the article's support of the criticisms made by the NAS report and supporting 2009 Law Commission Report commenting thus; 'These references (used in the submitted publication) regard persons who are not practitioners but are from academia and legal circles from those with no practical experience, and often those that have various personal agendas which are for the most part not in the best interest of forensic science' (Margot, 2009, page 1). In response to the reporting of reliability error estimates in this paper the reviewer stated, '[Regarding error levels], although everyone is firmly in favor of protocols and procedures that would contribute to maximizing accuracy and fairness in any forensic comparison, physical comparisons of barefoot impressions deal with evidence that is different from case to case. Variations are normal in both crime scene impressions as well as in exemplar inked impressions. Predicting theoretical error levels in physical match examinations is not possible, nor would the results in one case be applicable to another' (page 1). It is further explained, 'Measurements are usually not made in physical comparisons because of their unreliability in contrast to direct physical comparisons through overlays' (page 1), suggesting that the overlay method is the only reliable method of footprint evaluation. The reviewer appeared unable to comprehend that the article established reliability of the measurement approach as a baseline for footprint measurement and he did not recommend the article to be published, describing it as 'unreliable' (Margot, 2009). Despite this set-back, the paper was accepted for publication in another peer-reviewed journal later that same year and has

received two subsequent citations (DiMaggio & Vernon, 2011; Krishan et al., 2011).

Several presentations were made by the researcher to the International Association of Identification addressing reliability and validity issues in forensic footprint examination. Verbal feedback at these conferences indicated that the practitioners were resistant to the idea that a forensic technique could be stripped down to its core in order to set firm scientific foundations, preferring instead to continue using methods that were scientifically unsubstantiated but deemed acceptable as they had been in use for a considerable amount of time. Misconceptions regarding the practice of bare footprint forensic work were also encountered at first hand from forensic podiatrists, this being one of the contributing factors which gave rise to the development of the published Role and Scope of Practice document for forensic podiatrists (Vernon et al., 2010). This document ascertains the areas of forensic identification in which a forensic podiatrist would be expected to work and identifies fields which would not be in the forensic podiatrist's remit.

However, it was considered that the research conclusions of this thesis were timely and pertinent when the UK Home Office-appointed forensic science regulator published the Codes of Practice and Conduct document, echoing the assertions made by the researcher in relation to the concepts of reliability and validity in forensic practice (Rennison, 2011). For example, section 20.8 of the document discusses validation of measurement-based methods and in particular the performance and functional requirements of the methods as well as the relevant measurement characteristics and parameters. Requirements discussed in this section are that results must be 'consistent, reliable, accurate, robust and with an uncertainty measurement', and that there should be 'a compatibility of results obtained by other analysts using different equipment and different methods' amongst other recommendations regarding validity of a technique (page 28). In terms of reliability, the document recommends the United Kingdom Accreditation Service's UKAS® M3003 publication, 'The expression of uncertainty and confidence in measurement' (UKAS, 2007), which explains the necessity of determining random and systematic error and the calculation of error levels using the SEM. The 2011 codes of practice and conduct for forensic science providers and practitioners in the criminal justice

system (Rennison, 2011) appears to support the scientific development and testing of the Reel method for footprint measurement presented in this thesis. Unlike the NAS and Law Commission report, this document contains a list of requirements rather than recommendations, and therefore must be viewed as essential to practice.

9.7 Future implications of the research

In practical terms, the method has identified previously unknown quantities such as the mean differences between static and dynamic length measurements for homogenous groupings. This type of information is a useful indicator in the assessment of crime scene footprint comparison with exemplar footprints. To further the credibility of forensic footprint examination, consistent measurements may hold importance and are worthy of further investigation. If a measurement remains stable, for example between the states of standing and turning, where others are variable, this could improve confidence as to the likelihood of two footprints belonging to the same person or not. The research demonstrated that between the states of walking and standing, footprint measurement asymmetry existed in all but one measurement; the widest part of the forefoot.

Other questions regarding the behaviour of human bare footprints still remain. Despite Kennedy et al.'s extensive studies investigating the uniqueness of the human footprint, the measurement method used had not been tested for reliability or validity (Kennedy, 1996; Kennedy et al., 2003; Kennedy, 2005; Kennedy et al., 2005). The design of these studies was limited in that they did not follow the scientific Popperian model of falsification. Karl Popper (1959) famously used the example of a black swan to demonstrate the difficulty of proving a belief to be true (that all swans are white), regardless of how many observations appear to support it. In attempting to ascertain whether or not footprint shapes are unique Kennedy and his colleagues chose to seek out similarities within a thoroughly heterogeneous sample. Perhaps a more appropriate approach would have been to assess the probability of footprint individuality by selecting a homogenous sample, for example, an endogamous group of same-sex subjects containing small ranges in height, age and weight values, and with comparable daily activities.

The Reel method has been established as an appropriate tool to be used in furthering this type of research. Its use will also be valuable for other research projects. Variability of the human footprint from the same subject is yet to be fully understood, as is the frequency of certain features displayed by footprints, such as a long second toe (Greek Ideal), within specific larger populations. This type of information would be helpful in building likelihood ratios regarding the probability of an unknown and exemplar footprint belonging to the same person, for presentation purposes in a court of law. Further research of this nature demands the creation of footprint databases of specific populations. The processes of collection, measurement and analyses of footprints for the databases must adhere to standardised protocols in order for the production of meaningful results. In line with the demands of recent law-driven requirements these protocols must be based on scientifically rigorous methods.

At the time of writing, a national footprint database is to be implemented. The researcher has been approached to lead a selected team of academics specifically in the collection and measurement of the footprints using the Reel method, thus authenticating the research.

Chapter 10
Conclusions

The research presented in this thesis grew from an interest in the use of footprints for forensic identification purposes. Ideas generated by this interest focused on the measurement of footprints. The first research objective outlined in Chapter 1 aimed to critically review the literature pertaining to footprint measurement approaches. This was achieved; articles were appraised not only in the field of identification, but also within clinical, biomechanical and shoe design research areas.

The second research objective, evaluation of the extent to which reliability and validity has been utilised in footprint measurement was also met. Exploration in this area crossed over three different disciplines; medicine, forensic science and law. By venturing across these borders the researcher discovered that the interpretation of certain measurement concepts differed between the disciplines, and that the bar for measurement rigour acceptability had been set at varying heights. It became clear that the current methods used in forensic footprint identification had not been sufficiently tested and, as a minimum, a robust measurement approach needed to be established. Developing in tandem and spurring on the research, law-driven policies were actively requesting that new technologies or tests must demonstrate evidence of rigorous scientific foundations before admissibility in a court of law.

The subsequent creation of a new pragmatic method allowed for the researcher to embark upon a feasibility study, contributing towards a scientific underpinning of forensic footprint identification. The new measurement approach complete with an assessment of its utility and limitations, was not a single challenge, but the outcome of tackling many unfolding challenges. Appropriate statistical and evaluative testing ensured confidence in establishing various aspects of the validity and reliability of the concept. In a bid to increase this confidence, the findings have been peer-reviewed and published.

A convenience sample was used to generate footprint data from which the following key findings were made:

- The measurement along the lateral border is highly predictive of stature
- The forefoot width measurement is a stable feature between the static and dynamic states
- Within-subject static and dynamic differences have been quantified

Thus, the third research objective which aimed to develop a new pragmatic approach to footprint impression measurement underpinned by high levels of validity and reliability was duly achieved.

The body of work presented here stands within its parameters. The next stage of the research will be to examine the validity of the approach from data collated and analysed when utilised the field of forensic identification. Until then, implications of the research findings in the field cannot be confidently stated.

Despite this, the Reel method offers a baseline footprint measurement approach to the forensic research community. There remains limited information to help practitioners make better judgements regarding crime scene footprints. An example is in the research area of within-subject variability, where little is known as to the differences in a person's footprint between the states of walking, turning, twisting, running and jumping. Another example is the area of individuality. Forensic footprint identification is dependent on the premise that all footprints are unique. Although Kennedy and his colleagues have made preliminary investigations, the theory of footprint shape individuality is yet to be established. The Reel method offers an approach from which to enter the vast body of research required to prove this theory. There is also a need for a greater gathering of population incidence data. Hopefully the proposed footprint database using the Reel method will provide a starting point for this type of research to be implemented.

Finally, the researcher is of the opinion that more scientifically sound research must be undertaken before footprint evidence can be confidently admissible in a court of law. This will promote the discipline of forensic footprint identification towards that of forensic science, rather than forensic technology.

List of References

Adler, P. A. & Adler, P. (1988) Observational techniques. In Denzin, N. K. & Lincoln, Y. S. (Eds.) *Collecting and interpreting qualitative materials.* Thousand Oaks: London, New Delhi: Sage Publications.

Allen, W.B. (2004) Statistics, science and public policy VIII; Science, ethics and the law. In Herzberg, A.M. & Olford, R.W. (Eds) *Proceedings of the conference on statistics, science and public policy, held at Herstmonceux Castle, Hailsham, UK, April 23-26, 2003.* Kingston, Ont.: Queens University.

Allinson, C. & Hayes, J. (1996) The Cognitive Style Index. *Journal of Management Studies*, 119–135.

Altman, D. G. (1991) Some common problems in medical research. In Altman, D. G. (Ed.) *Practical statistics for medical research.* London: Chapman and Hall.

Anderson, H. (1966) *The influence of hormones on human development,* Philadelphia: W.B. Saunders.

Anderson, M., Blais, M. & Green, W. T. (1956) Growth of the normal foot during childhood and adolescence; length of the foot and interrelations of foot, stature, and lower extremity as seen in serial records of children between 1-18 years of age. *Am J Phys Anthropol,* 14, 287-308.

Arif, M., Ohtaki, Y., Nagatomi, R. & Inooka, H. (2004) Estimation of the effect of cadence on gait stability in young and elderly people using approximate entropy technique. *Measurement Science Review,* 4, 29-40

Ashizawa, K., Kumakura, C., Kusumoto, A. & Narasaki, S. (1997) Relative foot size and shape to general body size in Javanese, Filipinas and Japanese with special reference to habitual footwear types. *Ann Hum Biol,* 24, 117-29.

Ashworth, A. (2006) *Principles of criminal law.* 5th Ed. New York: Oxford University Press Inc.

Atamturk, D. & Duyar, I. (2008) Age-related factors in the relationship between foot measurements and living stature and body weight. *J Forensic Sci,* 53, 1296-300.

Atamturk, D. (2010) Estimation of sex from the dimensions of foot, footprints and shoe. *Anthrop. Anz,* 68, 21-9.

Atkinson, G. & Nevill, A. M. (1998) Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med,* 26, 217-38.

Aubin, C. E., Bellefleur, C., Joncas, J., De Lanauze, D., Kadoury, S., Blanke, K., Parent, S. & Labelle, H. (2011) Reliability and accuracy analysis of a new semiautomatic radiographic measurement software in adult scoliosis. *Spine (Phila Pa 1976),* 36, E780-90.

Barker, S. L. & Scheuer, J. L. (1998) Predictive value of human footprints in a forensic context. *Med Sci Law,* 38, 341-6.

Baumgartner, T. (1989) Norm-referenced measurement: Reliability. In Safrit, M. J. & Wood, T. M. (Eds.) *Measurement concepts in physical education and exercise science.* Champaign, Ill.: Human Kinetics Books.

Baxter, P. & Jack, S. (2008) Qualitative case study methodology: Study design and implementation for novice researchers. *The Qualitative Report.*

BBC News (2009) Key Kercher case evidence criticised. *BBC News Channel* [online] Available at: <http://news.bbc.co.uk/go/pr/fr/-/1/hi/uk/8263676.stm> [Accessed 18th September 2009].

Berger, M. A. & Solan, L. M. (2008) The uneasy relationship between science and law: An essay and introduction. *Brooklyn Law Review,* 73, 847-854.

Birtane, M. & Tuna, H. (2004) The evaluation of plantar pressure distribution in obese and non-obese adults. *Clin Biomech (Bristol, Avon),* 19, 1055-9.

Black, B. (1988) A unified theory of scientific evidence. *Fordham Law Review,* 56, 595-599.

Bland, M. (1987) Clinical measurement. In Bland, M. (Ed.) *An introduction to medical statistics.* Oxford: Oxford University Press.

Bland, J. M. & Altman, D. G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet,* 1, 307-10.

Bland, J. M. & Altman, D. G. (1996) Measurement error. *BMJ,* 313, 744.

Bland, J. M. & Altman, D. G. (2003) Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet Gynecol,* 22, 85-93.

Bloom, M., Fischer, J. & Orme, J. (2009) *Evaluating practice: Guidelines for the accountable professional.* 6th Ed. Boston: Allyn and Bacon.

Bodziak, W., J (2000) *Footwear impression evidence: Detection, recovery, and examination*, Boca Raton: CRC Press.

Bogin, B. & Varela-Silva, M. I. (2010) Leg length, body proportion, and health: a review with a note on beauty. *Int J Environ Res Public Health,* 7, 1047-75.

Bogin, B. (1999) *Patterns of human growth,* Cambridge: Cambridge University Press.

Borkowski, K. (2002) Factors influencing the direct identification of a human being on the basis of footprints. *16th IAFS Conference I.* Montpellier.

Bosch, K., Nagel, A., Weigend, L. & Rosenbaum, D. (2009) From "first" to "last" steps in life–pressure patterns of three generations. Clin Biomech, 24, 676-681.

Bowling, A. (2002) Research methods in health: investigating health and health services. Buckingham: Open University Press.

Brown, J. D. (1999) Standard error vs. Standard error of measurement. *Shiken:JALT Testing & Evaluation SIG Newsletter*, 3, 20-25.

Bruton, A., Conway, J. & Holgate, S. (2000) Reliability: What is it, and how is it measured? *Physiotherapy,* 86, 94-99.

Bryman, A. & Cramer, D. (2005) *Quantitative data analysis with SPSS 12 and 13: a guide for social scientists,* London: Routledge.

Burls, A. (2006) *CASP International Network* [online] Available at: <www.casp-uk.net> [Accessed 14th March 2006]

Burnett, A., Green, J., Netto, K. & Rodrigues, J. (2007) Examination of EMG normalisation methods for the study of the posterior and posterolateral neck muscles in healthy controls. *J Electromyogr Kinesiol,* 17, 635-41.

Byers, S., Akoshima, K. & Curran, B. (1989) Determination of adult stature from metatarsal length. *Am J Phys Anthropol,* 79, 275-9.

Campbell, D. T. & Fiske, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull,* 56, 81-105.

Carey, M. A., Laird, D. E., Murray, K. A. & Stevenson, J. R. (2010) Reliability, validity, and clinical usability of a digital goniometer. *Work,* 36, 55-66.

Carrier, B. (2003) Open source digital forensic tools: The legal argument. *Stake Research Report.*

Cassidy, M. J. (1980) *Footwear identification,* Ottawa, Public relation branch, Royal Canadian Mounted Police.

Cavanagh, P. R. & Rodgers, M. M. (1987) The arch index: a useful measure from footprints. *J Biomech,* 20, 547-51.

Cesario, S., Morin, K. & Santa-Donato, A. (2002) Evaluating the level of evidence of qualitative research. *J Obstet Gynecol Neonatal Nurs,* 31, 708-14.

Chatburn, R. L. (1996) Evaluation of instrument error and method agreement. *AANA J,* 64, 261-8.

Chen, C.-H., Huang, M.-H., Chen, T.-W., Weng, M.-C., Lee, C.-L. & Wang, G.-J. (2006) The correlation between selected measurements from footprint and radiograph of flatfoot. *Arch Phys Med Rehabil,* 87, 235-40.

Chen, H-T., Wei, L-Y., Chang, C-F. (2011). Nonlinear revision control for images. *ACM Transactions on Graphics*, 30, 159-167.

Chew-Graham, C. A., May, C. R. & Perry, M. S. (2002) Qualitative research and the problem of judgement: lessons from interviewing fellow professionals. *Fam Pract,* 19, 285-9.

Chinn, S. (1991) Statistics in respiratory medicine. 2. Repeatability and method comparison. *Thorax,* 46, 454-6.

Chiu, T. T. & Sing, K. L. (2002) Evaluation of cervical range of motion and isometric neck muscle strength: reliability and validity. *Clin Rehabil,* 16, 851-8.

Chockalingam, N. & Ashford, R. L. (2002) Foot length ratios for selected dimensions in a non clinical male sample. *Australasian Journal of Podiatric Medicine,* 36, 45-48.

Chu, W. C., Lee, S. H., Chu, W., Wang, T. J. & Lee, M. C. (1995) The use of arch index to characterize arch height: a digital image processing approach. *IEEE Trans Biomed Eng,* 42, 1088-93.

Claassen, J. (2005) The gold standard: not a golden standard. *BMJ,* 330, 1121.

Clarke, H. H. (1933) An objective method of measuring the height of the longitudinal arch in foot examination. *Research Quarterly,* 4, 99-107.

Cobey, J. C. & Sella, E. (1981) Standardizing methods of measurement of foot shape by including the effects of subtalar rotation. *Foot Ankle,* 2, 30-6.

Coffield, F., Moseley, D., Hall, E. & Ecclestone, K. (2004) Learning styles and pedagogy in post-16 learning. A systematic and critical review. *The Learning and Skills Research Centre.*

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences,* Hillsdale, N.J.: L. Erlbaum Associates.

Cohen, J. (1992) A power primer. *Psychological Bulletin*, 112, 155-59.

Coldwells, A., Atkinson, G. & Reilly, T. (1994) Sources of variation in back and leg dynamometry. *Ergonomics,* 37, 79-86.

Cole, S. A. & Dioso-Villa, R. (2009) Investigating the 'CSI Effect' Effect: Media and litigation crisis in criminal law. *Stanford Law Review,* 61, 1335-1374.

Cole, S. A. (2007) Toward evidence-based evidence: supporting forensic knowledge claims in the post-Daubert era. *Tulsa Law Review,* 43, 263-283.

Cronbach, L. J. & Meehl, P. E. (1955) Construct validity in psychological tests. *Psychol Bull,* 52, 281-302.

Cunningham, D. J. & Romanes, G. J. (1976) *Cunningham's manual of practical anatomy,* London: Open University Press.

D'Agostino, R. B. (1986) Tests for Normal Distribution. In D'agostino, R. B. & Stephens, M. A. (Eds.) *Goodness-of-fit techniques.* New York: Marcel Decker, Inc.

D'Agostino, R. B. (1971) An omnibus test of normality for moderate and large size samples. *Biometrika,* 58, 341-7.

Daubert v Merrell Dow Pharmaceuticals Inc. [1992] 509 US 579

Denegar, C. R. & Ball, D. W. (1993) Assessing reliability and precision of measurement: an introduction to intraclass correlation and standard error of measurement *Journal of Sports Rehabilitation,* 2, 35-42.

DiMaggio, J. (2004) The role of feet and footwear in medicolegal investigations. In Rich, J. & Dean, D. E. (Eds.) *Forensic medicine of the lower extremity: human identification and trauma analysis of the thigh, leg, and foot.* Totowa, N.J., Humana; Oxford: Blackwell.

DiMaggio, J. A. & Vernon, D. W. (2011) *Forensic podiatry: Principles and methods.* Humana Press.

Douglas, K. S. & Webster, C. D. (1999) The HCR-20 violence risk assessment scheme. Concurrent validity in a sample of incarcerated offenders. *Criminal Justice and Behavior* 26, 3 -19.

Douglass, J. (1979) Validation of two subjective rating systems for synchronized swimming educational and psychological measurement *Educational and Psychological Measurement,* 39, 373-80.

Dowling, A. M., Steele, J. R. & Baur, L. A. (2001) Does obesity influence foot structure and plantar pressure patterns in prepubescent children? *Int J Obes Relat Metab Disord,* 25, 845-52.

Dulniak, D. J., Busekist, J. L., Cline, C. J., Jones, M. K. & White, K. M. (1996) Institutional Policy Formation For Fax Documents. *AACRAO Fax Guidelines.*

ECRI Institute (1990) Fading images on thermal paper. *Health Devices,* 19, 374-5.

Edgar, D., Finlay, V., Wu, A. & Wood, F. (2009) Goniometry and linear assessments to monitor movement outcomes: are they reliable tools in burn survivors? *Burns,* 35, 58-62.

Eliasziw, M., Young, S. L., Woodbury, M. G. & Fryday-Field, K. (1994) Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Phys Ther,* 74, 777-88.

ENFSI (2006) European Network of Forensic Science Institutes. *Guidelines for best practice in the forensic examination of digital technology. Version V.* [online] Available at: <http://www.enfsi.eu/sites/default/files/documents/forensic_it_best_practice_guide_0.pdf> [Accessed 8th November 2008]

Endo, M., Ashton-Miller, J. A. & Alexander, N. B. (2002) Effects of age and gender on toe flexor muscle strength. *J Gerontol A Biol Sci Med Sci,* 57, M392-7.

Erdemir, A., Hamel, A. J., Fauth, A. R., Piazza, S. J. & Sharkey, N. A. (2004) Dynamic loading of the plantar aponeurosis in walking. *J Bone Joint Surg Am,* 86-A, 546-52.

Faigman, D. L., Saks, M. J., Sanders, J. & Cheng, E. K. (2005) *Modern scientific evidence. The law and science of expert testimony.*, West Information Publishing Group.

Falconi, M. (2009) Expert: bloody footprint not Italy defendent's. *Associated Press. Newsday.* [online] Available at: <http://www.newsday.com/news/world/expert-bloody-footprint-not-italy-defendant-s-1.1458666> [Accessed 18th September 2009].

Falls, H. B. (1986) Coed football: Hazards, implications and alternatives. *The Physician and Sportsmedicine,* 14, 207-222.

Farrell, M., Hartnett, B. R., Horgan-Černy, R., Hill, V., Mallon, A., O'Keeffe, H., O'Sullivan, C., Mulcahy, C. & Loughman, M. (2010) Records management best practice guidelines. *Office of Corporate and Legal Affairs.* University College Cork.

Fascione, J. M., Crews, R. T. & Wrobel, J. S. (2012) Dynamic footprint measurement collection techniques and intrarater reliability. *J. Am. Podiatric. Med. Ass.* 102, 130-138.

Fawzy, I. A. & Kamal, N. N. (2010) Stature and body weight estimation from various footprint measurements among Egyptian population. *J Forensic Sci,* 55, 884-8.

Field, A. (2005) *Discovering statistics using SPSS: (and sex, drugs and rock'n'roll),* London: SAGE Publications.

Fleiss, J. L. (1986) *The design and analysis of clinical experiments,* New York; Chichester: Wiley.

Fleming, N. D. & Mills, C. (1992) Not Another Inventory, Rather a Catalyst for Reflection. *To Improve the Academy,* 11, 137.

Forensic Science Regulation Unit (2009). *Notes of the ninth meeting, held at 11am on Monday 1 June 2009 at the Home Office, 2 Marsham Street, London SW1P 4DF.* [Online] Available from: *Forensic Science Advisory Council* <http://police.homeoffice.gov.uk/publications/forensic-science-regulator/FSAC-minutes-1-June-092835.pdf?view=Binary> [Accessed 14th November 2009].

Forriol, F. & Pascual, J. (1990) Footprint analysis between three and seventeen years of age. *Foot Ankle,* 11, 101-4.

French, S. (1988) How significant is statistical significance? A critique of the use of statistics in research. *Physiotherapy,* 74, 266-268.

Frye v United States [1923] 54 App. D.C. 46, 293 F.1013

Gall, M. D., Gall, J. P. & Borg, W. R. (2003) *Educational research: An introduction,* Boston, MA, A & B Publications.

Garrett, R. J. (2009) *IAI Letter from President Robert Garrett to its Membership.* [online] Available at: <http://www.theiai.org/current_affairs/nas_memo_20090219.pdf> [Accessed 21st February 2009].

Gauch, H.G., Jr. (2006) Winning the accuracy game. *American Scientist,* 94, 133-141

Gibson, I., Farrelly, P., Harris, E., Hoey, K., Iddon, B., Key, R., McWalter, T., Murrison, A., Smith, G., Spink, B. & Turner, D. (2005) Forensic science on trial. *Science and Technology*. The Stationery Office Limited.

Gibson, R. (1986). Courted expert steps on toes with footprints. *Chicago Tribune [online] 6th April*. Available at <http://www.law-forensic.com/cfr_robbins_7.htm> [Accessed 21st September 2006]

Giles, E. & Klepinger, L. L. (1988) Confidence intervals for estimates based on linear regression in forensic anthropology. *J Forensic Sci,* 33, 1218-22.

Giles, E. & Vallandigham, P. H. (1991) Height estimation from foot and shoeprint length. *Journal of Forensic Sciences,* 36, 1134-1151.

Glaser, B. & Strauss, A. (1967) *The discovery of grounded theory: Strategies for qualitative research.* New York: Aldine Publishing Company.

Goodwin, L. D. & Leech, N. L. (2003) The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Meas Eval Couns Dev,* 36, 181-91.

Gordon, C. C. & Buikstra, J. E. (1992) Linear models for the prediction of stature from foot and boot dimensions. *J Forensic Sci,* 37, 771-82.

Grabiner, M. D., Feuerbach, J. W., Lundin, T. M. & Davis, B. L. (1995) Visual guidance to force plates does not influence ground reaction force variability. *J Biomech,* 28, 1115-7.

Graneheim, U. H. & Lundman, B. (2004) Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Educ Today,* 24, 105-12.

Grčar, M., Šarabon, N., Strel, J., Starc, G. & Labrovič, J. (2006) *Foot study: Automating foot geometry analysis* [online] Available at: <www.footstudyproject.atspace.com> [Accessed 15th October 2007].

Green, S. B. (1991) How many subjects does it take to do a regression analysis? *Multivariate Behavioural Research,* 26, 499-510.

Greenhalgh, T. (2004) *How to read a paper: The basics of evidence based medicine*, BMJ Books.

Greenhalgh, T. & Peacock, R. (2005) Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ,* 331, 1064-5

Grivas, T. B., Mihas, C., Arapaki, A. & Vasiliadis, E. (2008) Correlation of foot length with height and weight in school age children. *J Forensic Leg Med,* 15, 89-95.

Gunn, N. (1991) Old and new methods of evaluating footprint impressions by a forensic podiatrist. *British Journal of Podiatric Medicine and Surgery,* 3, 8-11.

Gustafsson, S., Sunnerhagan, K. S. & Dahlin-Ivanoff, S. (2004) Occupational therapists' and patients' perceptions of ABILHAND, a new assessment tool for measuring manual ability. *Scandinavian Journal of Occupational Therapy,* 11, 107-117.

Hamberg, K., Johansson, E., Lindgren, G. & Westman, G. (1994) Scientific rigour in qualitative research--examples from a study of women's health in family practice. *Fam Pract,* 11, 176-81.

Hamill, J., Bates, B., Knutzen, K. & Kirkpatrick, G. (1989) Relationship between selected static and dynamic lower extremity measures. *Clin Biomech (Bristol, Avon),* 4, 217-225.

Hansen, M. (1993) Believe it or not. *American Bar Association Journal,* 79, 64-70.

Harbour, R. & Miller, J. (2001) A new system for grading recommendations in evidence based guidelines. *BMJ,* 323, 334-6.

Harris, R. I. & Beath, T. (1947) *Army foot survey: an investigation of foot ailments in Canadian soldiers.* Ottawa: National Research Council of Canada.

Hawes, M. R., Heinemeyer, R., Sovak, D. & Tory, B. (1994a) An approach to averaging digitized plantagram curves. *Ergonomics,* 37, 1227-30.

Hawes, M. R., Nachbauer, W., Sovak, D. & Nigg, B. M. (1992) Footprint parameters as a measure of arch height. *Foot Ankle,* 13, 22-6.

Hawes, M. R., Sovak, D., Miyashita, M., Kang, S. J., Yoshihuku, Y. & Tanaka, S. (1994b) Ethnic differences in forefoot shape and the determination of shoe comfort. *Ergonomics,* 37, 187-96.

Hawes, M.,R. & Sovak, D. (1994) Quantitative morphology of the human foot in a North American population. *Ergonomics,* 37, 1213-26.

HCPC (2012) Standards of conduct, performance and ethics. *Health and Care Professions Council publications.* [Online] Available at <www.hpc-uk.org/assets/documents/10003B6EStandardsofconduct,performanceandethics.pdf> [Accessed 24th July 2012]

Hendry, J. R. (2003) Environmental NGOs and business: A grounded theory of assessment, targeting, and influence. *Business & Society,* 42, 267-276.

Herrmann, N. (1996) *The whole brain business book,* New York; London, McGraw-Hill.

Hertzog, K., Garn, S. & Hempy 3rd, H. (1969) Partitioning the effects of secular trend and ageing on adult stature. *Am J Phys Anthropol,* 31, 111-5.

Hicks, C. (2005) *Research methods for clinical therapists: applied project design and analysis,* Edinburgh: Churchill Livingstone.

Hicks, J.H. (1954) The mechanics of the foot. II. The Plantar aponeurosis and the arch. *J.Anat.* 88, 25-30.

Ho, F., Lau, F., Downing, M. G. & Lesperance, M. (2008) A reliability and validity study of the Palliative Performance Scale. *BMC Palliat Care,* 7, 10.

Honey, P. & Mumford, A. (1992) *The manual of learning styles,* Peter Honey.

Houck, M. M. & Siegel, J. A. (2010) *Fundamentals of forensic science,* Academic Press.

Howitt, D. (2006) *Introduction to forensic and criminal psychology,* New York; Harlow, Pearson Longman.

Hsu, S.S. (2012) Convicted defendants left uninformed of forensic flaws found by Justice Dept. *The Washington Post April 16th 2012.* [online] Available at: <http://www.washingtonpost.com/local/crime/convicted-defendants-left-uninformed-of-forensic-flaws-found-by-justice-

dept/2012/04/16/3450dbf8-ea33-11el-87c7-93316b9dfff3_story.html#>
[Accessed 17th April 2012]

Huang, Y. P., Zheng, Y. P., Leung, S. F. & Mak, A. F. (2007) Reliability of measurement of skin ultrasonic properties in vivo: a potential technique for assessing irradiated skin. *Skin Res Technol,* 13, 55-61.

Hughes, J., Clark, P. & Klenerman, L. (1990) The importance of the toes in walking. *J Bone Joint Surg Br,* 72, 245-51.

Hyzer, W. G. & Krauss, T. C. (1988) The bitemark standard reference scale - ABFO No.2. *Journal of Forensic Sciences,* 33, 498-506.

Igbigbi, P. S. & Msamati, B. C. (2002) The footprint ratio as a predictor of pes planus: a study of indigenous Malawians. *J Foot Ankle Surg,* 41, 394-7

Innes, E. & Straker, L. (1999) Validity of work-related assessments. *Work,* 13, 125-152.

Jackson, A. S. (1989) Application of regression analysis to exercise science. In Safrit, M. J. & Wood, T. M. (Eds.) *Measurement concepts in physical education and exercise science.* Champaign, IL: Human Kinetics Books.

Jasuja, O. P. & Manjula (1993) Estimation of stature from footstep length. *Forensic Sci Int,* 61, 1-5.

Jasuja, O. P., Harbhajan, K. & Anupama, K. (1997) Estimation of stature from stride length while walking fast. *Forensic Sci Int,* 86, 181-186.

Jasuja, O. P., Singh, J. & Jain, M. (1991) Estimation of stature from foot and shoe measurements by multiplication factors: a revised attempt. *Forensic Sci Int,* 50, 203-15.

Johnson, D. J. (2008) Ridgeflow of the feet. *IAI 93rd International Education Conference.* Kentucky, USA.

Jung, J.-W., Sato, T. & Bien, Z. (2004) Dynamic footprint-based person recognition method, using hidden markov model and neural network. *International Journal of Intelligent Systems,* 19, 1127-1141.

Kahane, D. & Thornton, J. (1987) Discussion of 'Estimating height and weight from size of footprints'. *Journal of Forensic Sciences,* 32, 9-10.

Kanchan, T., Krishan, K., ShyamSundar, S., Aparna, K.R., & Jaiswal, S. (2012) Analysis of footprint and its parts for stature estimation in Indian population. *The Foot.* doi:10.1016/j.foot.2012.02.010

Kanchan, T., Menezes, R. G., Moudgil, R., Kaur, R., Kotian, M. S. & Garg, R. K. (2008) Stature estimation from foot dimensions. *Forensic Sci Int,* 179, 241 e1-5.

Katzmarzyk, P. T. & Leonard, W. R. (1998) Climatic influences on human body size and proportions: ecological adaptations and secular trends. *Am J Phys Anthropol,* 106, 483-503.

Kendra, K. & Taplin, L. J. (2004) Project success: A cultural framework. *Project Management Journal,* 35, 30-45.

Kennedy, R. B. (1996) Uniqueness of bare feet and its use as a possible means of identification. *Forensic Sci Int,* 82, 81-7.

Kennedy, R. B. (2005) Ongoing research into barefoot impression evidence. In Rich, J., Dean, D. E. & Powers, R. H. (Eds.) *Forensic Medicine of the Lower Extremity: Human Identification and Trauma Analysis of the Thigh, Leg and Foot.* Totowa, NJ: The Humana Press Inc.

Kennedy, R. B., Chen, S., Pressman, I. S., Yamashita, A. B. & Pressman, A. E. (2005) A large-scale statistical analysis of barefoot impressions. *J Forensic Sci,* 50, 1071-80.

Kennedy, R. B., Pressman, I. S., Chen, S., Petersen, P. H. & Pressman, A. E. (2003) Statistical analysis of barefoot impressions. *J Forensic Sci,* 48, 55-63.

Kerr, W. G. (2000) Seen at the scene - plantar dermatoglyphic use in identification and detection. *British Journal of Podiatry,* 3, 57-60.

Kippen, S. C. (1993) A preliminary assessment of recording the physical dimensions of an inked footprint. *Journal of British Podiatric Medicine,* 48, 74-80.

Klackenberg, E. P., Elfving, B., Haglund-Ackerlind, Y. & Carlberg, E. B. (2005) Intra-rater reliability in measuring range of motion in infants with congenital muscular torticollis. *Advances in Physiotherapy,* 7, 84-91.

Kleijn, S. A., Aly, M. F., Terwee, C. B., Van Rossum, A. C. & Kamp, O. (2011) Reliability of left ventricular volumes and function measurements using three-dimensional speckle tracking echocardiography. *Eur J Echocardiogr.* 13, 159-168.

Klementa, J., Komenda, S. & Kratosa, J. (1973) Use of a biometrical method for prediction of body height from the known value of foot length. *Anthropologie,* 11, 31-43.

Knight, B. (2004) The establishment of identity of human remains. In Saukko, P. (Ed.) *Knight's Forensic Pathology.* 3rd ed. London: Arnold Publishers.

Koch, S. & Schneider, G. (2000) Results from Software Engineering Research into Open Source Development Projects Using Public Data. In Hansen, H. R. & Janko, W. H. (Eds.) *Diskussionspapiere zum Tätigkeitsfeld Informationsverarbeitung und Informationswirtschaft.* Wien, Wirtschaftsuniversität Wien.

Kolb, D. (1984) *Experiential learning: Experience as the source of learning and development.* , Englewood Cliffs, NJ: Prentice-Hall.

Krishan, K. & Sharma, A. (2007) Estimation of stature from dimensions of hands and feet in a North Indian population. *J Forensic Leg Med,* 14, 327-32.

Krishan, K. (2007) Individualizing characteristics of footprints in Gujjars of North India--forensic aspects. *Forensic Sci Int,* 169, 137-44.

Krishan, K. (2008a) Estimation of stature from footprint and foot outline dimensions in Gujjars of North India. *Forensic Sci Int,* 175, 93-101.

Krishan, K. (2008b) Determination of stature from foot and its segments in a north Indian population. *Am J Forensic Med Pathol,* 29, 297-303.

Krishan, K. (2008c) Establishing correlation of footprints with body weight--forensic aspects. *Forensic Sci Int,* 179, 63-9.

Krishan, K. (2011) *Footprints in forensics: A study of North Indian population.* [Conference] 96th IAI International Education Conference, Milwaukee, WI. (Personal communication 15th July 2011).

Krishan, K., Kanchan, T. & Sharma, A. (2012) Multiplication factor versus regression analysis in stature estimation from hand and foot dimensions. *Journal of Forensic and Legal Medicine,* doi/10.1016/j.jflm.2011.12.024.

Krishan, K., Kanchan, T., Passi, N. & DiMaggio, J. A. (2011) Heel-Ball (HB) Index: Sexual Dimorphism of a New Index from Foot Dimensions. *J Forensic Sci.* 57, 172-5

Kroke, A., Boeing, H., Rossnagel, K. & Willich, S. N. (2004) History of the concept of 'levels of evidence' and their current status in relation to primary prevention through lifestyle interventions. *Public Health Nutr,* 7, 279-84.

Kulthanan, T., Techakampuch, S. & Bed, N. D. (2004) A study of footprints in athletes and non-athletic people. *J Med Assoc Thai,* 87, 788-93.

Kusumoto, A., Suzuki, T., Kumakura, C. & Ashizawa, K. (1996) A comparative study of foot morphology between Filipino and Japanese women, with reference to the significance of a deformity like hallux valgus as a normal variation. *Ann Hum Biol,* 23, 373-85.

Kwon, O. Y., Tuttle, L. J., Commean, P. K. & Mueller, M. J. (2009) Reliability and validity of measures of hammer toe deformity angle and tibial torsion. *Foot (Edinb),* 19, 149-55.

Lamm, B. M., Paley, D., Kurland, D. B., Matz, A. L. & Herzenberg, J. E. (2006) Multiplier method for predicting adult foot length. *J Pediatr Orthop,* 26, 444-8.

Lamparter, J., Shulze, A., Schuff, A.C., Berres, M., Pfeiffer, N. and Hoffmann, E.M. (2011) Learning curve and fatigue effect of flicker defined form perimetry. *Am J Ophthalmol,* 151, 1057-1064.

Landauer, T. (1997) Behavior research methods in HCI. In Helander, M., Landauer, T. K. & Prabhu, P. V. (Eds.) *Handbook of human-computer interaction.* 2nd ed. Amsterdam; Oxford, Elsevier.

Landorf, K. (2002) Letter to the Editor. Re: Payne C, Oates M, Mitchel A. The response of the foot to prefabricated orthosis of different arch heights. Australasian Journal of Podiatric Medicine 2002; 36 (1): 7-12. *Australasian Journal of Podiatric Medicine,* 36, 49.

Landis, J. R. & Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics,* 33, 159-74.

Laskowski, G. E. & Kyle, V. L. (1988) Barefoot impressions--a preliminary study of identification characteristics and population frequency of their morphological features. *J Forensic Sci,* 33, 378-88.

Laub, J. H. (2011) Translational Criminology. Office of Justice Programs. *National Institute of Justice.* [online] Available at: <http://www.nij.gov/nij/about/speeches/translational-criminology-3-1-2011.htm> [Accessed 4th October 2011]

Law, M. C., Baum, C. M. & Dunn, W. (2000) *Measuring occupational performance: supporting best practice in occupational therapy,* Thorofare, NJ: Slack.

Law, M., Stewart, D., Pollock, N., Letts, L., Bosch, J. & Westmorland, M. (1998) Guidelines for critical review form - quantitative studies. [online] Available at: <http://www.srs-mcmaster.ca/Portals/20/pdf/ebp/quanguidelines.pdf> [Accessed 24th June 2005]

Leard, J. S., Breglio, L., Fraga, L., Ellrod, N., Nadler, L., Yasso, M., Fay, E., Ryan, K. & Pellecchia, G. L. (2004) Reliability and concurrent validity of the figure-of-eight method of measuring hand size in patients with hand pathology. *J Orthop Sports Phys Ther,* 34, 335-40.

Leedy, P. D. (1993) *Practical research: planning and design,* New York: Macmillan; London: Maxwell Macmillan International.

Letts, L., Wilkins, S., Law, M., Stewart, D., Bosch, J. & Westmorland, M. (2007) Guidelines for critical review form: Qualitative studies (version 2.0). Qualitative review form guidelines. *School of Rehabilitation Science, McMaster University* [online] Available at: <http://www.srs-mcmaster.ca/Portals/20/pdf/ebp/qualguidelines_version2.0.pdf> [Accessed 15th May 2009]

Lin, C. H., Chen, J. J., Wu, C. H., Lee, H. Y. & Liu, Y. H. (2004) Image analysis system for acquiring three-dimensional contour of foot arch during balanced standing. *Comput Methods Programs Biomed,* 75, 147-57.

Lincoln, Y. S. & Guba, E. G. (1985) *Naturalistic inquiry,* Beverly Hills, CA: London, Sage.

Luo, G., Houston, V. L., Mussman, M., Garbarini, M., Beattie, A. C. & Thongpop, C. (2009) Comparison of male and female foot shape. *J Am Podiatr Med Assoc,* 99, 383-90.

Maes, R., Dojcinovic, S., Andrianne, Y. & Burny, F. (2006) Study of the plantar arch: correlations between podometrical and radiological parameters. Results of a prospective study of 79 cases. *Rev Med Brux,* 27, 422-9.

Malina, R. M., Little, B. B., Stern, M. P., Gaskill, S. P. & Hazuda, H. P. (1983) Ethnic and social class differences in selected anthropometric characteristics of Mexican American and Anglo adults: the San Antonio Heart Study. *Hum Biol,* 55, 867-83.

Maltais, L. & Yamashita, A. B. (2010) A validation study of barefoot morphology. *Journal of Forensic Identification,* 60, 362-370.

Mann, R. A. & Hagy, J. L. (1979) The function of the toes in walking, jogging and running. *Clin Orthop Relat Res,* 24-9.

Margot, P. (2009) Associate Editor, Forensic Science International. *Ms. No. FSI-D-09-00226.*[online] Available at: <http://ees.elsevier.com/fsi/viewLetter.asp?id=83350&lsid={DEED8E5D-8EF2-4F0F-AF872245602D5471}> [Accessed 8th June 2009].

Marincola, F. M. (2003) Translational Medicine: A two-way road. *J Transl Med,* 1, 1.

Martin, R. (1928) *Lehrbuch der Anthropologie,* Jena, Gustav Fisher.

Mathieson, I., Upton, D. & Birchenough, A. (1999) Comparison of footprint parameters calculated from static and dynamic footprints. *The Foot,* 9, 145-149.

Mathieson, I., Upton, D. & Prior, T. D. (2004) Examining the validity of selected measures of foot type: a preliminary study. *J Am Podiatr Med Assoc,* 94, 275-81.

Matthews, J. N., Altman, D. G., Campbell, M. J. & Royston, P. (1990) Analysis of serial measurements in medical research. *BMJ,* 300, 230-5.

McCaffrey, R.J., Duff, K. and Westervelt, H.J. (2000) *Practitioner's guide to evaluating change with neuropsychological assessment instruments.* New York: Plenum Publishers.

McCrory, J. L., Young, M. J., Boulton, A. J. M. & Cavanagh, P. R. (1997) Arch index as a predictor of arch height. *The Foot,* 7, 79-81.

McRoberts, F., Mills, S. & Possley, M. (2004) Forensics under the microscope: Unproven techniques sway courts, erode justice. *The Print,* 20, 1-7.

MEDLINE Course Materials. *BMA Library Seeking Evidence* (2005). [Online] Available at: <http://library.bma.org.uk/library> [Accessed 9th November 2006].

Menz, H. B. (2004) Two feet, or one person? Problems associated with statistical analysis of paired data in foot and ankle medicine. *The Foot,* 14, 2-5.

Menz, H. B., Zammit, G. V., Munteanu, S. E. & Scott, G. (2006) Plantarflexion strength of the toes: age and gender differences and evaluation of a clinical screening test. *Foot Ankle Int,* 27, 1103-8.

Merriman, L.M. & Tollafield, D.R. (2002) *Assessment of the lower limb,* Edinburgh: Churchill Livingstone.

Messick, S. (1995) Validity of psychological assessment: validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *Am Psychol,* 50, 741-9.

Meyers-Rice, B., Sugars, L., McPoil, T. & Cornwall, M. W. (1994) Comparison of three methods for obtaining plantar pressures in nonpathologic subjects. *J Am Podiatr Med Assoc,* 84, 499-504.

Michels, E. (1985) *Design of research and analysis of data in the clinic: an introductory manual for clinical research,* American Physical Therapy Association, Division of Research and Education.

Miles, J. & Shelvin, M. (2000) *Applying regression and correlation: A guide for students and researchers,* London: Sage.

Milz, S., McNeilly, C., Putz, R., Ralphs, J. R. & Benjamin, M. (1998) Fibrocartilages in the extensor tendons of the interphalangeal joints of human toes. *Anat Rec,* 252, 264-70.

Mochimaru, M. & Kouchi, M. (1997) Automatic calculation of the medial axis of foot outline and its flexion angles. *Ergonomics,* 40, 450-464.

Moenssens, A. A. (1995) *Scientific evidence in civil and criminal cases,* Westbury, N.Y: Foundation.

Moorthy, N., Samsudin, W. & Ismail, M. (2011) A study on footprints of Malaysian athletes and non-athletes for application during forensic comparison. *Malaysian Journal of Forensic Sciences,* 2, 29-35.

Montori, V. M., Wilczynski, N. L., Morgan, D. & Haynes, R. B. (2005) Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. *BMJ,* 330, 68.

Morley, S., Shapiro, D. A. & Biggs, J. (2004) Developing a treatment manual for attention management in chronic pain. *Cogn Behav Ther,* 33, 1-11.

Mosier, C. I. (1947) A critical examination of the concepts of face validity. *Educational and Psychological Measurement,* 7, 191-205.

Myers, J. B., Oyama, S., Wassinger, C. A., Ricci, R. D., Abt, J. P., Conley, K. M. & Lephart, S. M. (2007) Reliability, precision, accuracy, and validity of posterior shoulder tightness assessment in overhead athletes. *Am J Sports Med,* 35, 1922-30.

Meyers-Rice, B., Sugars, L., McPoil, T. & Cornwall, M. W. (1994) Comparison of three methods for obtaining plantar pressures in nonpathologic subjects. *J Am Podiatr Med Assoc,* 84, 499-504.

Morlock, M. & Mittlmeiser, T. (1992) First step method vs. full gait method – results of a comparison. *Eur J Phys Med Rehab*, (suppl 1), 2, 33-4.

Nakajima, K., Mizukami, Y., Tanaka, K. & Tamura, T. (2000) Footprint-based personal recognition. *IEEE Trans Biomed Eng,* 47, 1534-7.

Natarajamoorthy, T., Khairulmazidah, M., Mohamed Hadzri Bin, Y. & Jayaprakash, P. T. (2011) Estimation of stature based on foot length of Malays in Malaysia. *Australian Journal of Forensic Sciences,* 43, 13-26.

Natarajan, N. & Cecil, R. (2005) Computer assisted analysis of footprint geometry. Journal of Forensic Identification, 55, 489-498.

National Institute for Health and Clinical Excellence. 2009. The guidelines manual 2009 - Chapter 6: Reviewing the evidence. [Online] Available at: <http://www.nice.org.uk/media/5F5/22/The_guidelines_manual_2009_-_Chapter_6_Reviewing_the_evidence.pdf> [Accessed 12th July 2010]

National Research Council. (2009) *Strengthening forensic science in the United States - A path forward.* National Academy of Sciences. Washington DC: The National Academies Press.

Nester, C.J. (1997) Rearfoot complex: a review of its interdependent components, axis orientation and functional model. *The Foot.* 7, 86.

Nicholson, D.E., Armstrong, P.F., Macwilliams, B.A., Terry, S., Porter, J. & Miller, M.L. (1998) The effects of velocity, step initiation, and a visible platform on plantar pressures of healthy children. *Gait Post.* 7, 146.

Nikolaidou, M. E. & Boudolos, K. D. (2006) A footprint-based approach for the rational classification of foot types in young schoolchildren. *The Foot,* 16, 82-90.

Noakes, H. & Payne, C. (2003) The reliability of the manual supination resistance test. *J Am Podiatr Med Assoc,* 93, 185-9.

Norkin, C. C. & White, D. J. (2003) *Measurement of joint motion: a guide to goniometry,* Philadelphia: F.A. Davis.

Norton, B. J. & Ellison, J. B. (1993) Reliability and concurrent validity of the Metrecom for length measurements on inanimate objects. *Phys Ther,* 73, 266-74.

Oberoi, D., Kuruvilla, A., Saralaya, K. M., Rajeev, A., Ashok, B., Nagesh, K. & Nageshkumar, R. (2006) Estimation of stature and sex from foot print length using regression formulae and standard foot print length formula respectively. *Journal of Punjab Academy of Forensic Medicine and Toxicology,* 6, 1-9.

Olivier, G. (1965) *Anatomie anthropologique,* Paris, Vigot Freres.

Opila, K. A. (1988) Gender and somatotype differences in postural alignment: Response to high-heeled shoes and simulated weight gain. *Clinical Biomechanics,* 3, 145-152.

Orlin, M. N. & McPoil, T. G. (2000) Plantar pressure assessment. *Phys Ther,* 80, 399-409.

Ostell, C. (2011) Competency assessment. *Interfaces Newsletter. Forensic Science Society,* 65, 5.

Otway v R. [2011] EWCA Crim 3.

Özaslan, A., Iscan, M. Y., Özaslan, I., Tugcu, H. & Koc, S. (2003) Estimation of stature from body parts. *Forensic Sci Int,* 132, 40-5.

Pagliari, C., Clark, D., Hunter, K., Boyle, D., Cunningham, S., Morris, A. & Sullivan, F. (2003) DARTS 2000 online diabetes management system: formative evaluation in clinical practice. *J Eval Clin Pract,* 9, 391-400.

Patton, M. Q. (1990) *Qualitative evaluation and research methods,* Newbury Park, CA: Sage Publications.

Pawar, R. M. & Pawar, M. N. (2012) Foot length – a functional parameter for assessment of height. *The Foot,* 22, 31-4.

Pereira-Maxwell, F. (1998) *A-Z of medical statistics: a companion for critical appraisal,* London: Arnold ; New York: Oxford University Press.

Perens, B. (1999) The open source definition. In Dibona, C. & Ockman, S. (Eds.) *Open sources: voices from the open source revolution.* O'Reilly Media.

Phillips, B., Ball, C., Sackett, D., Badenoch, D., Straus, S., Haynes, B. & Dawes, M. (1998) Oxford centre for evidence-based medicine. In Centre for Evidence Based Medicine, *Evidence Levels.* Oxford, UK.

Polat-Ozsoy, O., Gokcelik, A. & Toygar Memikoglu, T. U. (2009) Differences in cephalometric measurements: a comparison of digital versus hand-tracing methods. *Eur J Orthod,* 31, 254-9.

Polit, D. & Hungler, B. (1999) *Nursing research. Principles and methods,* Philadelphia, New York, Baltimore: J.B. Lippincott Company.

Polski, J. (2007) Message from the Chief Operations Officer. *Journal of Forensic Identification,* 57, 758-763.

Pomeroy, V. M., Chambers, S. H., Giakas, G. & Bland, M. (2004) Reliability of measurement of tempo-spatial parameters of gait after stroke using GaitMat II. *Clin Rehabil,* 18, 222-7.

Popper, K. R. (2002) *The logic of scientific discovery.* 2nd ed. London: Routledge.

Portney, L. G. & Watkins, M. P. (2000) *Foundations of clinical research: applications to practice,* Upper Saddle River, NJ: Prentice Hall; London: Prentice-Hall International.

Pretty, I. A. (2006) The barriers to achieving an evidence base for bitemark analysis. *Forensic Sci Int,* 159 Suppl 1, S110-20.

Q v Sica (2011) Unpublished transcript of court proceedings [07/11/11, Day 4.] Indictment No 68 0f 2011. Brisbane.

Qamra, S. R., Sharma, B. P. & Kaila, P. (1980) Naked foot marks - a preliminary study of identification factors. *Forensic Sci Int,* 16, 145-52.

R v Cannings [2004] EWCA Crim 1.

R v Clark [2003] EWCA Crim 1020.

R v Dallagher [2002] EWCA Crim 1903.

R v Dimitrov [2003] 68 O.R. 3d 641, No. C34922. *Court of Appeal for Ontario.* Sack, Goldblatt, Mitchell LLP.

R v Kempster [2008] EWCA Crim 975.

R v T [2010] EWCA Crim 2439.

Randall, F. E., Munro, E. H. & White, R. M. (1951) Anthropometry of the foot (US Army white male); Report 172. Natick, MA: Environmental Protection Division, Quartermaster Research and Development Centre.

Rani, M., Tyagi, A., Ranga, V., Rani, Y. & Murari, A. (2011) Stature estimates from foot dimensions. *Journal of Punjab Academy of Forensic Medicine and Toxicology,* 11, 26-30.

Rankin, G. & Stokes, M. (1998) Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clin Rehabil,* 12, 187-99.

Redmayne, M. (2001) *Expert evidence and criminal justice,* Oxford: Oxford University Press.

Reel, S., Rouse, S., Vernon, W. & Doherty, P. (2010) Reliability of a two-dimensional footprint measurement approach. *Sci Justice,* 50, 113-8.

Reel, S., Rouse, S., Vernon, W. & Doherty, P. (2012) Estimation of stature from static and dynamic footprints. *Forensic Sci Int,* 219, 283.e1-283.e5.

Reid, A., Birmingham, T. B., Stratford, P. W., Alcock, G. K. & Giffin, J. R. (2007) Hop testing provides a reliable and valid outcome measure during rehabilitation after anterior cruciate ligament reconstruction. *Phys Ther,* 87, 337-49.

Reneman, M. F., Jorritsma, W., Schellekens, J. M. & Goeken, L. N. (2002) Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. *J Occup Rehabil,* 12, 119-29.

Rennison, A. (2011) Codes of practice and conduct for forensic science providers and practitioners in the criminal justice system. *Home Office Forensic Science Regulation Unit.* [Online] <http://www.homeoffice.gov.uk/publications/agencies-public-bodies/fsr/codes-practice-conduct?view=Binary> [Accessed 30th December 2011].

Review of Science in the Home Office, Royal Statistical Society (2003). Royal Statistical Society.

Richer, P. & Hale, R. B. (1973) *Artistic anatomy. Translated and edited by Robert Beverly Hale*, New York: Watson-Guptill; London: Pitman.

Riddiford-Harland, D. L., Steele, J. R. & Storlien, L. H. (2000) Does obesity influence foot structure in prepubescent children? *Int J Obes Relat Metab Disord,* 24, 541-4.

Ridola, C., Palma, A., Cappello, F., Gravante, G., Russo, G., Truglio, G., Pomara, F. & Amato, G. (2001) Symmetry of healthy adult feet: role of orthostatic footprint at computerized baropodometry and of digital formula. *Ital J Anat Embryol,* 106, 99-112.

Riviello, R. (2008) *Manual of forensic emergency medicine*, Jones and Bartlett Publishers, Inc.

Robbins, L. M. (1978) The individuality of human footprints. *J Forensic Sci,* 23, 778-85.

Robbins, L. M. (1985) *Footprints: collection, analysis, and interpretation,* Springfield, Ill: C. C. Thomas.

Robbins, L. M. (1986) Estimating height and weight from size of footprints. *J Forensic Sci,* 31, 143-52.

Roberts, P. (2009) Forensic science evidence - a consumer perspective. [Online] Available at: *Scottish Institute for Policing Research.* <http://www.sipr.ac.uk/downloads/Forensics2/Paul_Roberts.pdf> [Accessed 26th March 2010].

Robson, C. (2002) *Real world research: a resource for social scientists and practitioner-researchers,* Madden, MA: Oxford, Blackwell Publishers.

Roig, M. (2010) Special issue: Responsible writing in science. *Biochemia Medica,* 20, 295-300.

Roman, J., Adams, W., Reid, S. & Reid, J. (2008) *Can DNA solve property crimes? Results from a randomized control trial.* Atlanta, GA: American Society of Criminology.

Romanos, M. T., Raspovic, A. & Perrin, B. M. (2011) The reliability of toe systolic pressure and the toe brachial index in patients with diabetes. *J Foot Ankle Res,* 3, 31.

Rose, G. K., Welton, E. A. & Marshall, T. (1985) The diagnosis of flat foot in the child. *Journal of Bone and Joint Surgery,* 67B, 71-78.

Rosnow, R. L. & Rosenthal, R. (2005) *Beginning behavioral research: a conceptual primer,* Upper Saddle River, N.J: Pearson/Prentice Hall.

Rossi, W. A. (1992) Podometrics: An new methodology for foot typing. *Journal of Testing and Evaluation,* 20, 301-311.

Rowntree, D. (1981) *Statistics without tears: a primer for non-mathematicians,* Harmondsworth, Penguin.

Rupert, J. 2004. Courts trample on life's work. *The Ottawa Citizen,* [online] 22nd June. Available at: <http://edelsonandassociates.com/news/dimitriov/dimitriov-23-03-04.htm> [Accessed 17th March 2006]

Rumsey, S. (2008) *How to find information: a guide for researchers,* Maidenhead, McGraw-Hill/Open University Press.

Sacco, I. D. C., Noguera, G., Bacarin, T., Casarotto, R. & Tozzi, F. (2009) Medial longitudinal arch change in diabetic peripheral neuropathy. *Acta Ortopedica Brasileira,* 17, 13-16.

Sackett, D. L., Rosenberg, W., Gray, J. A. M., Haynes, R. M. & Richardson, W. S. (1996) Evidence based medicine: what it is and what it isn't. *British Medical Journal,* 312, 71-72.

Safrit, M. J. & Wood, T. M. (1989) *Measurement concepts in physical education and exercise science,* Champaign, Ill: Human Kinetics Books.

Safrit, M. J. (1981) *Evaluation in physical therapy,* Englewood Cliffs, NJ, Prentice Hall.

Safrit, M. J. (1989) An overview of measurement. In Safrit, M. J. & Wood, T. M. (Eds.) *Measurement concepts in physical education and exercise science.* Champaign, Ill: Human kinetics books.

Salthouse, T.A. & Tucker-Drob, E.M. (2008) Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology,* 22, 800-811.

Saks, M. J. & Faigman, D. L. (2008) Failed forensics: How forensic science lost its way and how it might yet find it. *Annu. Rev. Law Soc. Sci.,* 4, 149-171.

Saks, M. J. & Koehler, J. J. (2008) The individualization fallacy in forensic science evidence. *Vanderbilt Law School,* 61, 199-220.

Saltzman, C. L. & Nawoczenski, D. A. (1995) Complexities of foot architecture as a base of support. *J Orthop Sports Phys Ther,* 21, 354-60.

Sanderson, D. J., Franks, I. M. & Elliott, D. (1993) The effects of targeting on the ground reaction forces during level walking. *Human Movement Science,* 12, 327-337.

Sanli, S. G., Kizilkanat, E. D., Boyan, N., Ozsahin, E. T., Bozkir, M. G., Soames, R., Erol, H. & Oguz, O. (2005) Stature estimation based on hand length and foot length. *Clin Anat,* 18, 589-96.

Science Review of the Home Office and Ministry for Justice (2003) *Government Office for Science.* [Online] Available at: <http://www.bis.gov.uk/assets/goscience/docs/science-review-ho-moj/ho-moj-review> [Accessed 3rd April 2007]

Sechrest, L. (2005) Validity of measures is no simple matter. *Health Services Research,* 40, 1584-1604.

Selfe, J. (1998) Validity and reliability of measurements taken by the Peak 5 motion analysis system. *J Med Eng Technol,* 22, 220-5.

Selfe, J., Harper, L., Pederson, I., Breen-Turner, J. & Waring, J. (2001) Four outcome measures for patellofemoral joint problems. *Physiotherapy,* 87, 516-522.

Sen, J. & Ghosh, S. (2008) Estimation of stature from foot length and foot breadth among the Rajbanshi: An indigenous population of North Bengal. *Forensic Sci Int,* 181, 1-6.

Sforza, C., Michielon, G., Fragnito, N. & Ferrario, V., F. (1998) Foot asymmetry in healthy adults: elliptic fourier analysis of standardized footprints. *Journal of Orthopaedic Research,* 16, 758-65.

Sharkey, N.A., Donahue, S.W., Ferris, L. (1999) Biomechanical consequences of plantar fascial release or rupture during gait. Part II: alterations in forefoot loading. *Foot Ankle Int.* 20, 86-96.

Sharma, B. R. (1970) Foot and foot wear evidence. *Journal of the Indian Academy of Forensic Sciences.,* 9, 9-13.

Shiang, T. Y., Lee, S. H., Lee, S. J. & Chu, W. C. (1998) Evaluating different footprint parameters as a predictor of arch height. *IEEE Eng Med Biol Mag,* 17, 62-6.

Shrout, P. E. (1998) Measurement reliability and agreement in psychiatry. *Stat Methods Med Res,* 7, 301-17.

Silverman, B. (2011) Home Office Chief Scientific Advisor. *Research and Development in forensic science: a review.* [online] Available at: <http://www.forensicdentalservices.co.uk/wp/wp-content/uploads/2011/07/forensic-science-review-report.pdf> [Accessed 29th August 2011].

Sim, J. (2001) Unreliable reliability analysis. *Physiotherapy,* 87.

Sirovatka, P. (2005) NIMH series transforms insights into clinical strategies. *Psychiatric News* 40, 17.

Smerecki, C. J. & Lovejoy, C. O. (1985) Identification via pedal morphology. *International Criminal Police Review,* 40, 186-90.

Smith, A. F. M. (1996) Mad cows and ecstasy: chance and choice in an evidence-based society *Journal of the Royal Statistical Society,* 159, 367-83.

Souza, R. B. & Powers, C. M. (2009) Concurrent criterion-related validity and reliability of a clinical test to measure femoral anteversion. *J Orthop Sports Phys Ther,* 39, 586-92.

Speckels, C. (2011) Can ACE-V be validated? *Journal of Forensic Identification,* 61, 201-109.

Staheli, L. T., Chew, D. E. & Corbett, M. (1987) The longitudinal arch. A survey of eight hundred and eighty-two feet in normal children and adults. *J Bone Joint Surg Am,* 69, 426-8.

State v Berry [2001] No. COAOO-263 546 S.E. 2d 145

State v Jones [2001] 343 S.C. 562, 541 S.E.2d 813

State v Jones [2009] South Carolina State Court of Appeal. 383 S.C. 26699

Stavlas, P., Grivas, T. B., Michas, C., Vasiliadis, E. & Polyzois, V. (2005) The evolution of foot morphology in children between 6 and 17 years of age:

a cross-sectional study based on footprints in a Mediterranean population. *The Journal of Foot and Ankle Surgery,* 44, 424-8.

Sternberg, R. J. (1997) *Thinking styles,* Cambridge: Cambridge University Press.

Stevens, P. J. M. (1993) *Understanding research: a scientific approach for health care professionals*, Campion.

Stewart, T. D. (1952) *Hrdlicka's practical anthropometry,* Philadelphia.

Stone, C. A., Nolan, B., Lawlor, P. G. & Kenny, R. A. (2011) Hand-held dynamometry: tester strength is paramount, even in frail populations. *J Rehabil Med,* 43, 808-11.

Strand, S., Belfrage, H., Fransson, G. & Levander, S. (1999) Clinical and risk management factors in risk prediction of mentally disordered offenders-more important than historical data?: A retrospective study of 40 mentally disordered offenders assessed with the HCR-20 violence risk assessment scheme. *Legal and Criminological Psychology* 4, 67-76.

Strauss, A. & Corbin, J. (1998) *Basics of qualitative research,* Thousand Oaks, CA: Sage publications.

Streiner, D. L. & Norman, G. R. (2006) "Precision" and "accuracy": two terms that are neither. *J Clin Epidemiol,* 59, 327-30.

Stripp, D. (2008) Another earprint conviction reversed. *Identification Evidence.* [online]. Available at: <http://www.forensic-evidence.com/site/ID/dallangher_UK.html> [Accessed 17th June 2008].

SWGIT (2001) Draft recommendations and guidelines for the use of digital imaging processing in the criminal justice system. Version 1.1. *Scientific Working Group on Imaging Technologies* [online] Available at: <www.fdiai.org/images/SWGITguidelines.pdf> [Accessed 8th November 2011]

Teh, E., Teng, L. F., U, R. A., Ha, T. P., Goh, E. & Min, L. C. (2006) Static and frequency domain analysis of plantar pressure distribution in obese and non-obese subjects. *Journal of Bodywork and Movement Therapies,* 10, 127-133.

The AGREE Collaboration (2001). Appraisal of Guidelines for Research & Evaluation (AGREE) Instrument [online]. Available at: <www.agreecollaboration.org > [Accessed 6th May 2006].

The Law Commission (2009) The admissibility of expert evidence in criminal proceedings in England and Wales. A new approach to the determination of evidentiary reliability. *Consultation Paper No 190.* [online]. Available at: <http://lawcommission.justice.gov.uk/docs/cp190_Expert_Evidence_Consultation.pdf> [Accessed 7th August 2009].

The Law Commission (2011) Expert evidence in criminal proceedings in England and Wales. *The House of Commons.* London: The Stationery Office.

Thomas, J. R., Nelson, J. K. & Silverman, S. J. (2005) *Research methods in physical activity,* Champaign, Ill: Leeds, Human Kinetics.

Thompson, A. L. T. & Zipfel, B. (2005) The unshod child into womanhood - forefoot morphology in two populations. *The Foot,* 15, 22-28.

Thompson, T. & Black, S. M. (2007) *Forensic human identification,* Boca Raton: Taylor & Francis.

Thordarson, D.B., Kumar, P.J., Hedman, T.P., Ebramzadeh, E. (1997) Effect of partial versus complete plantar fasciotomy on the windlass mechanism. *Foot Ankle Int.* 18, 16-20.

Thurzo, A., Javorka, V., Stanko, P., Lysy, J., Suchancova, B., Lehotska, V., Valkovic, L. & Makovnik, M. (2010) Digital and manual cephalometric analysis. *Bratisl Lek Listy,* 111, 97-100.

Topinard, P. (1890) *Anthropology,* [S.l.], Chapman & Hall.

Tortora, G. J. & Grabowski, S. R. (2003) *Principles of anatomy and physiology,* New York; [Great Britain]: Wiley.

Trotter, M. & Gleser, G. (1952) Estimation of stature from long bones of American Whites and Negroes. *Am J Phys Anthropol,* 10, 463-514.

Tsung, B., Fan, Y. B., Zhang, M. & Boone, D. A. (2003) Quantitative comparison of plantar foot shapes under weight-bearing conditions. *Journal of Rehabilitation Research and Development,* 40, 517-526.

Turlik, M. A. & Kushner, D. (2000) Levels of evidence of articles in podiatric medical journals. *J Am Podiatr Med Assoc,* 90, 300-2.

Turner, D. W. (2010) Qualitative Interview Design: A Practical Guide for Novice Investigators. *The Qualitative Report.*

Tuttle, R. H. (1986) Review: L.M. Robbins' Footprints: Collection, analysis and interpretation. *American Anthropologist,* 88, 1000-1002.

Urwin, C. (2012) Forensic podiatry inquiry [letter] (Personal communication, 10th August 2012)

US v Allen. (2002) 207 F. Supp. 2d 856.

US v Trala. (2001) No. CR.A.00-23-GMS.

UKAS (2007) M3003: The Expression of Uncertainty and Confidence in Measurement JH 2nd Ed. [online] Available at: <http://www.ukas.com> [Accessed 4th January 2009].

University of Huddersfield (2012) *Course Finder.* [online] Available at: <http://www.hud.ac.uk/courses/course/index.php?ipp=00006658> [Accessed 28th February 2012].

Urry, S. R. & Wearing, S. C. (2001) A comparison of footprint indexes calculated from ink and electronic footprints. *J Am Podiatr Med Assoc,* 91, 203-9.

Urry, S., R & Wearing, S., C. (2005) Arch indexes from ink footprints and pressure platforms are different. *The Foot,* 15, 68-73.

Van De Graff, K. M. (1988) *Human anatomy,* Dubuque, Iowa, Wm C Brown.

Van Schie, C. H., Abbott, C. A., Vileikyte, L., Shaw, J. E., Hollis, S. & Boulton, A. J. (1999) A comparative study of the Podotrack, a simple semiquantitative plantar pressure measuring device, and the optical pedobarograph in the assessment of pressures under the diabetic foot. *Diabet Med,* 16, 154-9.

Vernon, D. W. & McCourt, F. J. (1999) Forensic podiatry - a review and definition. *British Journal of Podiatry,* 2, 45-48.

Vernon, W. (2006) Personal communication.

Vernon, W. (2007) The Foot. In Thompson, T. & Black, S. (Eds.) *Forensic Human Identification: An Introduction.* Boca Raton FL: CRC Press.

Vernon, W. (2009) Forensic podiatry: A review. *Axis: The Online Journal of Centre for Anatomy and Human Identification,* 1, 60 - 70.

Vernon, W., Brodie, B., DiMaggio, J., Gunn, N., Kelly, H., Nirenberg, M., Reel, S. & Walker, J. (2010) Forensic podiatry: Role and scope of practice (in the context of forensic human identification). *IDentification News,* 40, 22-24.

Vernon, W., Parry, A. & Potter, M. (1998) Preliminary findings in a delphi study of shoe wear marks. *Journal of Forensic Identification,* 48, 22-38.

Vidya, C. S., Shamsundar, N. M., Saraswathi, G. & Nanjaiah, C. M. (2011) Estimation of stature using footprint measurements. *Anatomica Karnataka,* 5, 37-9.

Villarroya, M. A., Esquivel, J. M., Tomas, C., Moreno, L. A., Buenafe, A. & Bueno, G. (2008) Assessment of the medial longitudinal arch in children and adolescents with obesity: footprints and radiographic study. *Eur J Pediatr,* 168, 559-67.

Wearing, S. C., Hills, A. P., Byrne, N. M., Hennig, E. M. & McDonald, M. (2004) The arch index: a measure of flat or fat feet? *Foot Ankle Int,* 25, 575-81.

Wearing, S. C., Urry, S. R. & Smeathers, J. E. (2000) The effect of visual targeting on ground reaction force and temporospatial parameters of gait. *Clin Biomech (Bristol, Avon),* 15, 583-91.

Wearing, S. C., Urry, S. R., Smeathers, J. E. & Battistutta, D. (1999) A comparison of gait initiation and termination methods for obtaining plantar foot pressures. *Gait and Posture,* 10, 255-263.

Weijers, R. E., Walenkamp, G. H. I. M., Van Mameren, H. & Van Den Hout, J. A. A. M. (2003) Changes of the soft tissue of the forefoot during loading: a volumetric study. *The Foot,* 13, 70-75.

Weiner, J. S. & Lourie, J. A. (1969) *Human biology: a guide to field methods,* Oxford; Edinburgh, Blackwell Scientific Publications.

Welton, E. A. (1992) The Harris and Beath footprint: interpretation and clinical value. *Foot Ankle,* 13, 462-8.

Whittle, M. (2003) *Gait analysis: an introduction,* Oxford, Butterworth-Heinemann.

Winkelmann, W. (1987) [Use of footprints, especially forefoot prints, from the forensic viewpoint]. *Z Rechtsmed,* 99, 121-8.

Wood, T. M. (1989) The changing nature of norm-referenced validity. In Wood, T. M. & Safrit, M. J. (Eds.) *Measurement concepts in physical education and exercise science.* Champaign, Ill: Human kinetic books.

Woods, N. F. & Catanzaro, M. (1988) *Nursing research. Theory and practice.,* St. Louis, Washington DC, Toronto: The C.V. Mosby Company.

Wunderlich, R. E. & Cavanagh, P. R. (2001) Gender differences in adult foot shape: implications for shoe design. *Med Sci Sports Exerc,* 33, 605-11.

Xiong, S., Goonetilleke, R. S., Witana, C. P., Weerasinghe, T. W. & Au, E. Y. (2010) Foot arch characterization: a review, a new metric, and a comparison. *J Am Podiatr Med Assoc,* 100, 14-24.

Yamamoto, L. G. & Wiebe, R. A. (1989) Improving medical communication with facsimile (fax) transmission. *The American Journal of Emergency Medicine,* 7, 203-208.

Yamashita, A. B. (2009) Personal communication.

Zammit, G. V., Menz, H. B. & Munteanu, S. E. (2011) Reliability of the TekScan MatScan(R) system for the measurement of plantar forces and pressures during barefoot level walking in healthy adults. *J Foot Ankle Res,* 3, 11.

Zeybek, G., Ergur, I. & Demiroglu, Z. (2008) Stature and gender estimation using foot measurements. *Forensic Sci Int,* 181, 54 e1-5.

List of Abbreviations

AI Arch Index

ANOVA Analysis of Variance

CI Confidence Interval

CV Coefficient of Variation

df Degrees of freedom

EBP Evidence-Based Practice

GIMP GNU Image Manipulation Program

ICC Intraclass Correlation Coefficient

JPEG Joint Photographic Expert Group

K-S Kolmogorov-Smirnov

LOA Limits of Agreement

MLA Medial Longitudinal Arch

MPJ Metatarsophalangeal Joint

NAS National Academy of Sciences

NICE National Institute for Health and Clinical Excellence

OCM Optical Center Method

OCPM Ohio College of Podiatric Medicine

OLE Oxford Levels of Evidence

PPM Pearson Product-Moment correlation coefficient

Q-Q Quantile-Quantile

SE Standard Error

SEE Standard Error of Estimate

SEM Standard Error of Measurement

SPSS Statistical Package for Social Scientists

Glossary

Degrees of Freedom (df) refers to the number of items, for example footprint measurements, that are free to vary when estimating a statistical parameter of a test.

Digitisation refers to the process of transforming data into a digital form so that it can be processed by a computer.

Dynamic relates to physical force or activity; in this thesis it specifically refers to the action of walking.

Footprint refers to the mark made onto a surface transferred from the plantar surface of a human naked foot.

Forensic relates to the use of technology and science to investigate and establish facts in criminal or civil courts of law.

Gold standard in the context of measurement testing refers to a valid diagnostic tool which is also reliable and accurate. In practice, gold standards are rarely 100% accurate, but are the best method of testing according to the current dogma.

Homogeneous samples include a narrow range or single value of a particular variable or variables, for example, static footprints from a group of male subjects only.

Heterogeneous samples involve the selection of subjects varying widely on the characteristic of interest, for example a group containing male subjects, female subjects, static and dynamic footprints.

Identification refers to the use of evidence to establish the identity of a single person from a larger population.

Impression in a context of forensic evidence involves a donor and a recipient. The donor contains some three-dimensional markings and the recipient in this thesis refers to a material that can form and hold a two-dimensional negative image of the donor markings.

P-value in the context of the statistical significance of a test represents the probability that any particular outcome would have arisen by chance. In this

thesis, p-values are set at < 0.05 if results are to determine significance, and at the < 0.01 level for more robust significance testing, depending upon the statistical method utilised for the analysis.

r refers to Pearson product-moment correlation coefficient values and is a measure of the strength of the linear relationship between two variables.

R^2 refers to the coefficient of determination and states how much of the value of one of the variables can be attributed solely to the other value of the other variable(s).

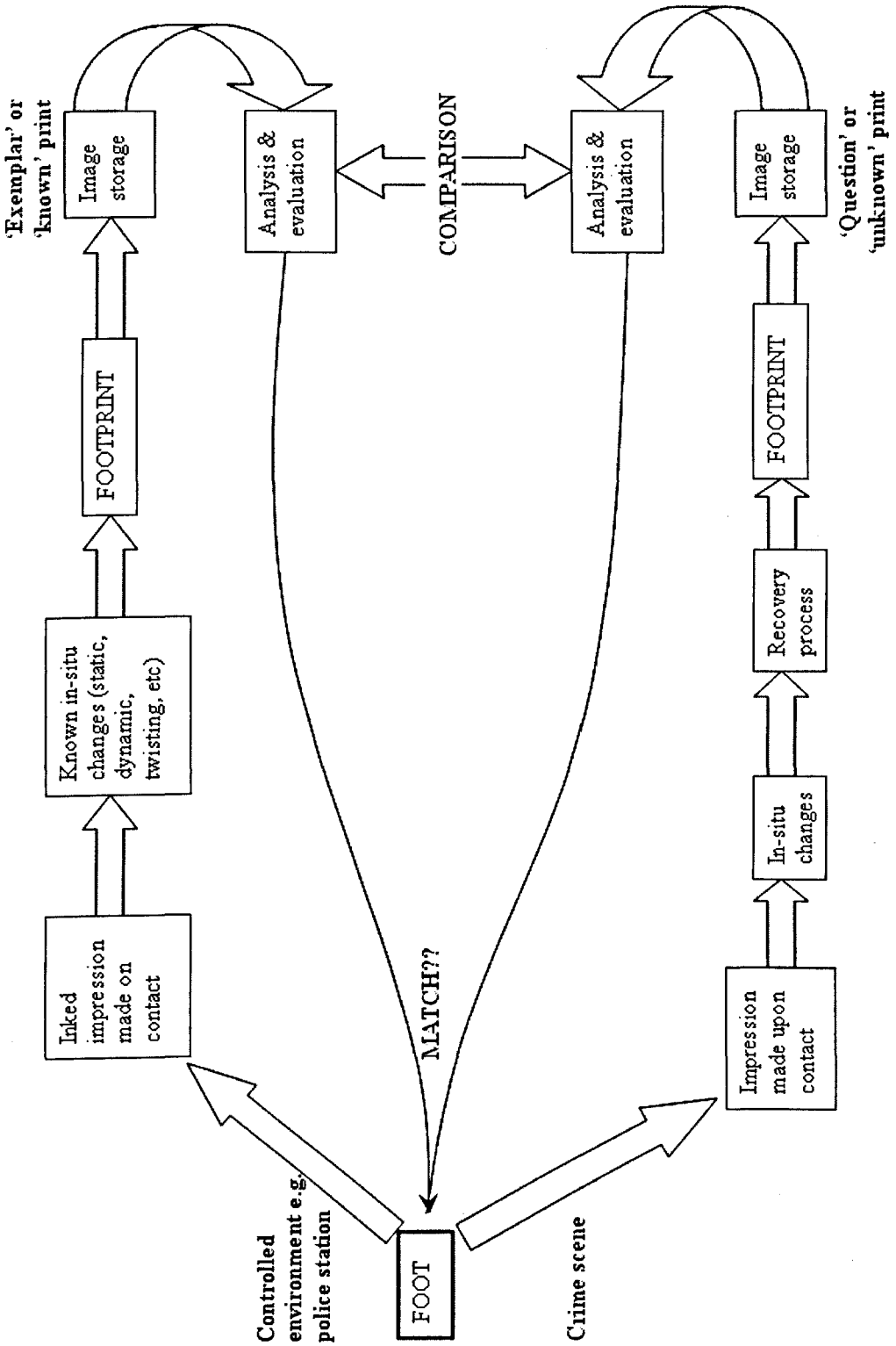
Static refers to the state of inactivity or stationary; in this thesis it specifically refers to the action of standing. Static footprint impressions are formed when the subject stands onto the inkless paper from the ink pad.

Two-dimensional describes a shape devoid of range or depth.

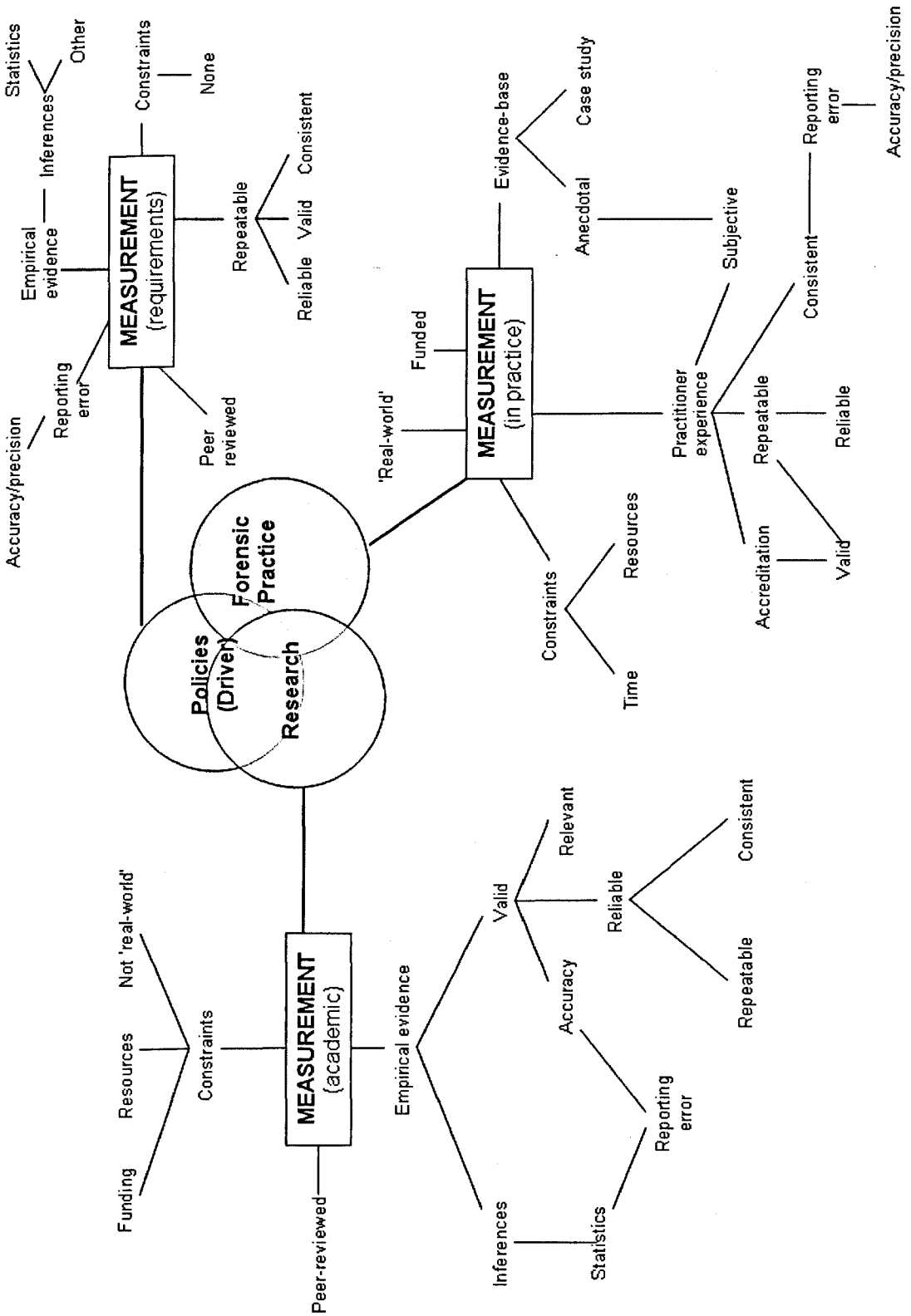
Unique refers to a footprint shape without an equal.

Appendices

A.1 Footprint Identification Process



A.2 Measurement Concepts in Footprint Identification



B.1 Critical Appraisal Tool

1. Is the title a clear and succinct statement of the research study?

2. Does the abstract provide a clear statement of the aims, methods, results and conclusions/implications of the study?

3. After reading the abstract, are you clear in your mind about the nature of the study?

4. Is there an adequate description of the general context for the study?

5. Is the literature review thorough, relevant, recent and properly used to provide a structured argument leading to the reason for conducting the reported piece of research?

6. Is the hypothesis (if appropriate) clearly stated, and the predicted relationship between the variables apparent?

7. If the research does not test the hypothesis, are the aims of the study clear?

8. Are the aims or hypothesis useful to my research?

9. Is the project likely to be of value to my research?

!0. Has the design of the study been properly described?

11. Has the researcher made it clear why this design was chosen?

12. Is the design appropriate for the aims/hypothesis stated in the introduction?

13. Are the sources of error acknowledged and controlled?

14. Is the sample suitable? Of an appropriate size? Fully described? Properly selected?

15. Were any sources of bias or error evident in the sample and/or in the process by which they were chosen?

16. Would this impact on the study's outcome?

17. Was any mechanical apparatus used in the study and, if so, was it properly described? Was it suitable for the project?

18. Were any other materials used, such as questionnaires, score sheets, attitude scales, etc?

19. Were these described fully and/or included in the appendix, if appropriate?

20. Were any questionnaires or scales which were used properly constructed and adequately tested before using them in the study? Were they suitable for their purpose?

21. Is the description of what was done absolutely clear?

22. Does it state the order in which things were done?

23. Does it provide a verbatim report of any instructions given to the subjects? Were the instructions clear?

24. Were the sources of error dealt with appropriately?

25. Was the method of data collection clearly described and appropriate?

26. Were the data a suitable measure of the dependent variable (if the study tested a hypothesis) or of the information required by the survey's aims?

27. Were the subjects treated well, their rights and confidentially protected?

28. Was the study ethical?

29. Could you repeat this study to the letter if it was considered necessary?

30. Are the graphs (if provided) clear, self-explanatory and useful?

31. Are the tables (if used) clearly labelled and constructed and with an obvious relevance to the study?

32. Are the statistical tests used the correct ones for the project's design?

33. Is the selected level of significance appropriate for the topic area?

34. Is the *p* value clearly stated and correct for the hypothesis as stated (i.e. one- or two-tailed)?

35. Are the results and conclusions clearly stated?

36. Are they related to other studies in the area, thereby putting them into a broader research framework?

37. Is a cogent theoretical explanation for the findings provided?

38. Are the results interpreted fully and correctly, or selectively and/or extravagantly?

39. Are any flaws in the study's design highlighted, together with recommendations for improvement?

40. Are the results interpreted with these limitations in mind?

41. Are any practical ramifications of the results discussed?

42. Do any ideas for future projects emerge?

43. Is every article, study, research report and book quoted in the reference section?

44. Do these references give all the required information?

45. Was the project a worthwhile one, contributing to the knowledge base of your field?

46. Was it clearly written, so that the content was easily accessible to the reader?

47. Is the report scientific and objective both in the way in which it was conducted as well as the way in which it was analysed and written up?

48. Is the article devoid of jargon?

49. Has the research project advanced this field in any way?

B.2 Grading of the Relevant Literature

Oxford Centre for Evidence-based Medicine Levels of Evidence

Level	Type of article	Definition/feature
1	RCT/Meta-analysis	RCT controls for selection of subjects, ttmt bias, and analysis of defined end points. Meta-analysis is the process of combining results from several different RCTs.
2	Cohort study	Prospective study of two uncontrolled groups
3	Case-control study	Retrospective study of two uncontrolled groups.
4	Case report/series	A study of a single subject or group of subjects receiving some uncontrolled ttmt. Best for generating, not testing a hypothesis.
5	Expert opinion	Based on years of clinical experience and intuition.

Adapted from Phillips et al., (2009).

Ohio College of Podiatric Medicine tool for grading the validity of an article

- Was the instrument used to collect the data validated?
- Were there an adequate number of patients in the study?
- Were the subjects suitable for the type of study/comparable at baseline?
- Were the data collected compatible with the statistical tests utilised, explanation of rationale for uncommon statistical tests

- Were appropriate descriptive and inferential statistics presented to allow for analysis?
- Were 95% CI/error rates presented about point estimates?

SCORING:	YES		UNCLEAR/POSSIBLY		NO
	4	3	2	1	0

Adapted from Turlick and Kushner (2000).

B.3 Instructions for Evaluating Qualitative Literature

Read the qualitative study and score each of the categories listed using the quality rating scale of 0 through 3 described below.

1. DV = Descriptive Vividness

2. MC = Methodological Congruence a. RD = Rigor in Documentation b. PR = Procedural Rigor c. ER = Ethical Rigour d. C = Confirmability

3. AP = Analytical Preciseness

4. TC = Theoretical Connectedness

5. HR = Heuristic Relevance a. IR = Intuitive Recognition b. RBK = Relationship to Existing Body of Knowledge c. A = Applicability

SCORING SCALE

3 = Good = 75%–100% criteria met

2 = Fair = 50%–74% criteria met

1 = Poor = 25%–49% criteria met

0 = No evidence that criteria met = < 25% criteria met

FINAL QUALITY OF EVIDENCE RATING

The quality of evidence rating was based on the total scores for each of the five categories described above. A quality of evidence rating for each qualitative study was assigned using the legend below:

QI: Total score of 22.5–30 indicates that 75% to 100% of the total criteria were met. (Good)

QII: Total score of 15–22.4 indicates that 50% to 74% of the total criteria were met. (Fair)

QIII: Total score of less than 15 indicates that less than 50% of the total criteria were met. (Poor)

(Cesario et al, 2001, page 711)

C.1 Searching strategy example

	Concept 1 Search Terms	Concept 2 Search Terms	Concept 3 Search Terms	Concept 4 Search Terms	a	b	c	d
		AND	AND	AND				
	Footprint*	Height	Human	Adult	5	1	3	3
OR		Stature						
OR		Age			4	2	6	0
OR		Weight			4	2	3	3
OR		Sex			2	6	7	1
OR		Gender						
OR		Body Mass Index			0	4	4	0
OR		BMI						
OR		Ethnic*			3	2	5	0
OR		Race						
OR		Racial						
OR		Differen*			4	0	0	4
OR		Dynamic						

a Articles retrieved

b Serendipitous searching

c Excluded e.g. pertaining to electronic footprints, carbon footprints, DNA footprints, shoeprints and animal footprints

d Final number of articles appraise

D.1 Ethical Approval 2007 Study

Sarah Reel

York St John University

Dr Simon Rouse
Chair of Research Ethics,
Direct Line 876901
e-mail s.rouse@yorksj.ac.uk

20 April 2007

Dear Sarah

RE: Validity of dynamic footprint measurement and an evaluation of effectiveness in forensic and clinical examination.

I can confirm that you have been granted research ethical approval for your research proposal submitted on the 12/3/07.

Yours sincerely

Cc Professor Patrick Doherty

Lord Mayor's Walk
York YO31 7EX
T. 01904 624 624
F. 01904 612 512
www.yorksj.ac.uk

D.2 Ethical Approval 2010 study

Sarah Reel
PhD Student
Faculty of Health & Life Sciences

York St John University

Dr Simon Rouse
Chair of Research Ethics
Direct Line 876901
e-mail: s.rouse@yorksj.ac.uk

20 January 2010

Dear Sarah

RE: Validity of a two-dimensional footprint measurement approach and an evaluation of its utility in forensic examination.

REF: UC/20/1/10/SR

I can confirm that your ethics proposals has been reviewed and approved

Yours sincerely

P.P. S

D.3 Information Sheet

Dear potential volunteer,

I am undertaking a research project as part of my PhD, on the individuality of human footprints in order to further the field of forensic identification. The PhD is registered with York St John University and the University of Leeds and is being supervised by Professor Patrick Doherty, Professor Wesley Vernon and Dr Simon Rouse. The study has been approved by the University ethics board.

The study will take place on the University campus during the month of June 2007 and it is estimated that it will require approximately 30 minutes of your time. If you decide to take part, you will be asked to remove your shoes and socks and walk along a short walkway where images of your footprint will be captured and used at a later date for measurement analysis. Your identity will remain anonymous and all information from participating in this study will be confidential. The data from the footprints will be kept on a password protected computer and only the researchers will have access to your personal details. You can withdraw at any time without explanation.

Previous research has shown that extremes in height and weight may alter the footprint's outline. In order to make the study more uniform it will therefore be necessary for your height and weight to be measured on the day. The weighing will be done behind a screen in private and all details will be confidential.

Studies have also suggested that foot shape can differ between ethnic groups. It is for this reason that you will be asked to volunteer your perceived racial background. Again, this information will be confidential.

'Inkless System' Mat. You will be asked to step bare foot on to a chemically impregnated mat and then onto a sheet of paper which will instantly capture your footprint. There will be no messy ink to clear up from your foot and therefore it should be a simple process. I have been assured that there has to date been no known allergies or issues regarding cross-infection/hygiene concerning the multiple use of this mat and its chemical substrate. There have been several studies using this inkless mat system involving thousands of volunteers, and no adverse reactions have been reported.

Clinical wipes will be available, should you feel at any time the necessity to clean your feet.

Exclusion Criteria

Unfortunately, you will not be able to take part in the study if:

- you are under 20 years of age
- you are unable to feel your feet
- you are unable to walk independently
- you have a known foot pathology e.g. arthritic conditions, recent surgery to the foot/feet, recent trauma including partial loss of foot tissue.
- you have a foot infection or open wounds on the foot, e.g. open heel cracks, ulcerations, fungal infections, verrucas.

Any Questions?

If you have any questions about the study, please do not hesitate to contact me at sarah.reel@yorksj.ac.uk.

D.4 Consent Form

- Name of Researcher: SARAH REEL.
  Contact details:
- Title of study: VALIDITY OF DYNAMIC AND STATIC FOOTPRINT MEASUREMENT AND AN EVALUATION OF EFFECTIVENESS IN FORENSIC EXAMINATION

Please read and complete this form carefully. If you are willing to participate in this study, ring the appropriate responses and sign and date the declaration at the end. If you do not understand anything and would like more information, please ask.

- I have had the research satisfactorily explained to me in verbal and / or written form by the researcher.

  YES / NO

- I understand that the research will involve walking barefoot along a short walkway. I will be asked to step on to a chemically impregnated mat containing a colourless, odourless, quick-drying ink on the walkway and then onto a corresponding square piece of paper which will immediately develop my footprint impression. This procedure will be repeated until 3 clear prints from my right foot are obtained. Additionally, 3 more static prints of my right foot will be taken by side-stepping onto the paper. It is estimated that this will take approximately 15-30 minutes of my time. I will also be required to have my height and weight measured in order to calculate my body mass index.

  YES / NO

- I understand that I may withdraw from this study at any time without having to give an explanation.

  YES / NO

- I understand that all information about me will be treated in strict confidence and that I will not be named in any written work arising from this study.

  YES / NO

- I understand that any of my footprints will be used solely for research purposes and will be destroyed on completion of your research.

  YES / NO

- I understand that you will be discussing the progress of your research with others at York St John University

  YES / NO

I freely give my consent to participate in this research study and have been given a copy of this form for my own information.

Signature: .....

Date: .....

D.5 Information sheet for experts

Dear

I am undertaking a research project as part of my PhD*, examining the validity of a method of measuring two-dimensional inked footprints. This is a continuation from a previous study which showed that the measuring method I used to measure walking and standing inked footprints was reliable. I am now keen to find out if the approach has practical uses in the 'real' world and if it could be at all utilised in your area of expertise.

One of my supervisors, Professor Wesley Vernon, OBE, suggested that you may be able to help me in this next part of my study, as you are an expert in the field of forensic footprint examination.

If you think you may be able to help me further in this project, the process will require you to watch a DVD illustrating a method of collecting two-dimensional footprints and also a CD detailing a method for measuring a scanned footprint image. You will then be required to read through a manual detailing the measurement approach which could be used for comparison and analysis of footprints for forensic purposes. The manual and supporting CD and DVD will be sent to you by post for you to peruse in your own time. I would then arrange to meet you at a time and place convenient to yourself, during May or June, in order to ask you several questions as to your thoughts on the information sent to you. The conversation will be recorded by me on tape and later transcribed for research purposes. The whole process involving watching the CD and DVD, reading the manual and completing the interview, will take approximately 3 - 4 hours of your time. All information from your participation in the study will be confidential.

Thank you for taking the time to read this email. I would be most grateful if you could contact me at sarah.reel@yorksj.ac.uk if you can help me. Alternatively, if you have a colleague who you think may be interested in this, perhaps you could pass on their contact details so that I could get in touch with them directly. I would be grateful if you could respond to me by 30th April 2010.

Best wishes

Sarah Reel

* The PhD is registered with York St John University and the University of Leeds and is being supervised by Professor Patrick Doherty, Professor Wesley Vernon and Dr Simon Rouse. The study has been approved by York St John University Research Ethics Committee.

D.6 Consent form for experts

Name of Researcher: SARAH REEL.

Contact details: **sarah.reel@yorksj.ac.uk**

Title of study: VALIDITY OF A TWO-DIMENSIONAL FOOTPRINT MEASUREMENT APPROACH AND AN EVALUATION OF ITS UTILITY IN FORENSIC EXAMINATION.

Please read and complete this form carefully. If you are willing to participate in this study, ring the appropriate responses and sign and date the declaration at the end. If you do not understand anything and would like more information, please ask.

- I have had the research satisfactorily explained to me in written form by the researcher.

  YES / NO

- I understand that I will be asked to watch a CD/DVD and read a manual detailing a method of collecting and measuring two-dimensional footprints. I will then be contacted directly by the researcher at a later date where I will be asked of my expert opinion regarding these methods.

  YES / NO

- I understand that the whole task will take approximately 3 - 4 hours of my time.

  YES / NO

- I understand that I may withdraw from this study at any time without having to give an explanation. **YES / NO**

- I understand that all information about me will be treated in strict confidence and that I will not be named in any written work arising from this study.

  YES / NO

- I understand that the taped conversation I complete will be destroyed at the end of the study. However, the transcribed conversation I had with you may appear in your thesis, but my identity will remain anonymous.

  YES / NO

- I understand that you will be discussing the progress of your research with others at York St John University.

  YES / NO

I freely give my consent to participate in this research study and have been given a copy of this form for my own information.

Signature: .....

Name (capital letters).....

Date: .....

Contact details:.....

E.1 Interview Questions

1. How many years have you been involved with dealing with footprint evidence and analysis?

   - What does your job involve?
   - How many footprint cases do you deal with per year?

2. Do you use a specific approach to measure your footprints?

3. In the guide I talk about the 'inkless system' for collecting footprints. Are you familiar with the inkless system?

   - Have you tried any other methods for collecting footprints?
   - Do you think that the inkless paper system is a valid and useful way of collecting footprints in your opinion?

4. Did you get a chance to read through the first few chapters of the guide – the underpinning evidence behind the approach?

   - Did you find this aspect useful in setting the scene?
   - Do you think there are bits of it that are superfluous to the guide?

5. Did you manage to have a look at this bit referring to the different methods of collecting footprints (Chapter 4)?

6. Do you think the approach which includes the manual, the DVD and the CD, would be useful for students?

   - What type of students?
   - How about forensic science trainees?

7. Do you think that the guide and the footprint collection DVD could enable you to collect both static and dynamic footprints equally well?

8. On the whole, do you think that the CD and the section in the guide describing the measuring method are overly-complicated?

   - Do you think the CD on its own would be enough to instruct someone on how to evaluate footprints – without the guide, just the CD on its own?

9. In the measurement method, where you have to draw a line skimming the outer pixels, there's some guesswork involved here. Do you think that's too subjective?

10. Would you be interested in trying this method in your line of work?

11. When doing the linear measurements, most methods discount the ghosting or flaring that occurs especially in the dynamic footprints, and measure to where the expert perceives the 'end' of the toe/heel print to be. In my approach, I include all the ghosted part of the print. Do you think this could hinder or help the analysis in any way?

12. Do you think that the method could be adapted for use with partial prints?

13. The literature underpinning footprint comparison and analysis is relatively weak, but the call for scientifically researched evidence behind the forensic methods is on the increase. As a practitioner do you think that this approach has the potential to contribute to the literature?

14. Are there any points you'd like to raise that haven't already been discussed?

F.1 Tests of normality (all measurements)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Weight (kg)	.099	61	.200*	.955	61	.025
Height (cm)	.091	61	.200*	.987	61	.761
heel to apex of 1st toe(D)	.106	61	.083	.980	61	.429
heel to apex of 2nd toe(D)	.111	61	.057	.970	61	.146
heel to apex of 3rd toe(D)	.100	61	.200*	.978	61	.336
heel to apex of 4th toe(D)	.096	61	.200*	.973	61	.195
heel to apex of 5th toe(D)	.121	61	.027	.966	61	.085
width of calcaneum(D)	.078	61	.200*	.978	61	.336
width of ball of foot(D)	.070	61	.200*	.968	61	.108
footprint angle(D)	.097	61	.200*	.940	61	.005
2-5 toe angle(D)	.058	61	.200*	.978	61	.355
1-5 toe angle(D)	.074	61	.200*	.978	61	.336
dist. met. angle(D)	.066	61	.200*	.988	61	.812
2-4 base toe angle(D)	.082	61	.200*	.981	61	.461
heel to apex of 1st toe(S)	.088	61	.200*	.977	61	.308
heel to apex of 2nd toe(S)	.098	61	.200*	.982	61	.523
heel to apex of 3rd toe(S)	.076	61	.200*	.987	61	.768
heel to apex of 4th toe(S)	.067	61	.200*	.983	61	.562
heel to apex of 5th toe(S)	.072	61	.200*	.973	61	.185
width of calcaneum(S)	.080	61	.200*	.977	61	.301
width of ball of foot(S)	.063	61	.200*	.971	61	.156
footprint angle(S)	.087	61	.200*	.935	61	.003
2-5 toe angle(S)	.106	61	.088	.955	61	.025
1-5 toe angle(S)	.072	61	.200*	.990	61	.916
dist. met. angle(S)	.066	61	.200*	.982	61	.497
2-4 base angle(S)	.049	61	.200*	.985	61	.677

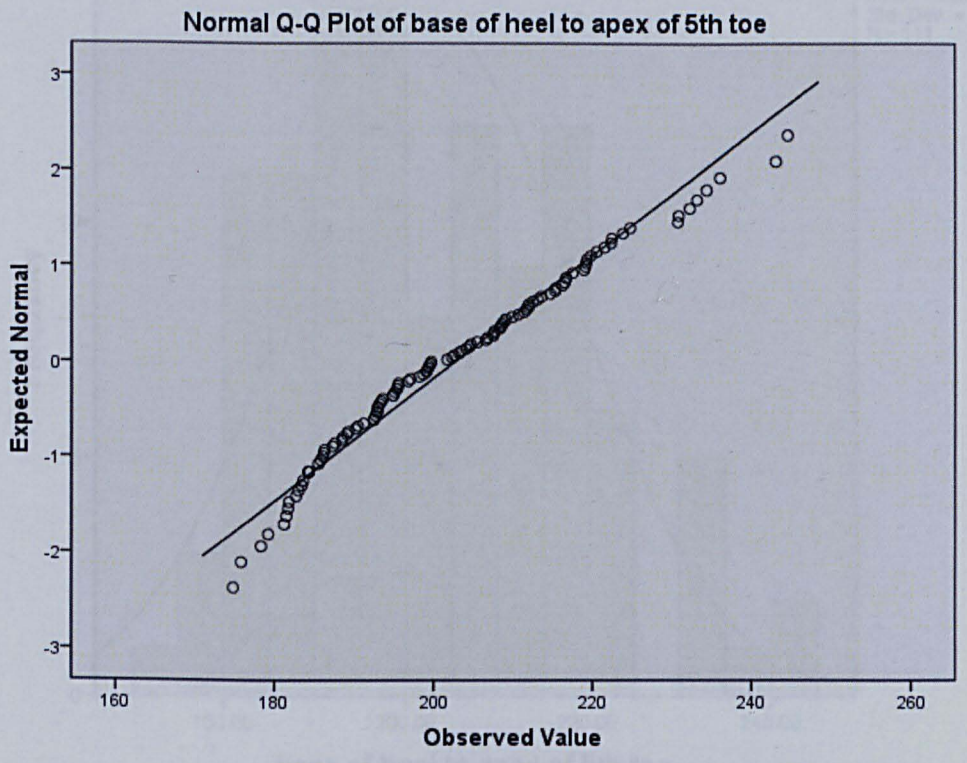
a. Lilliefors Significance Correction

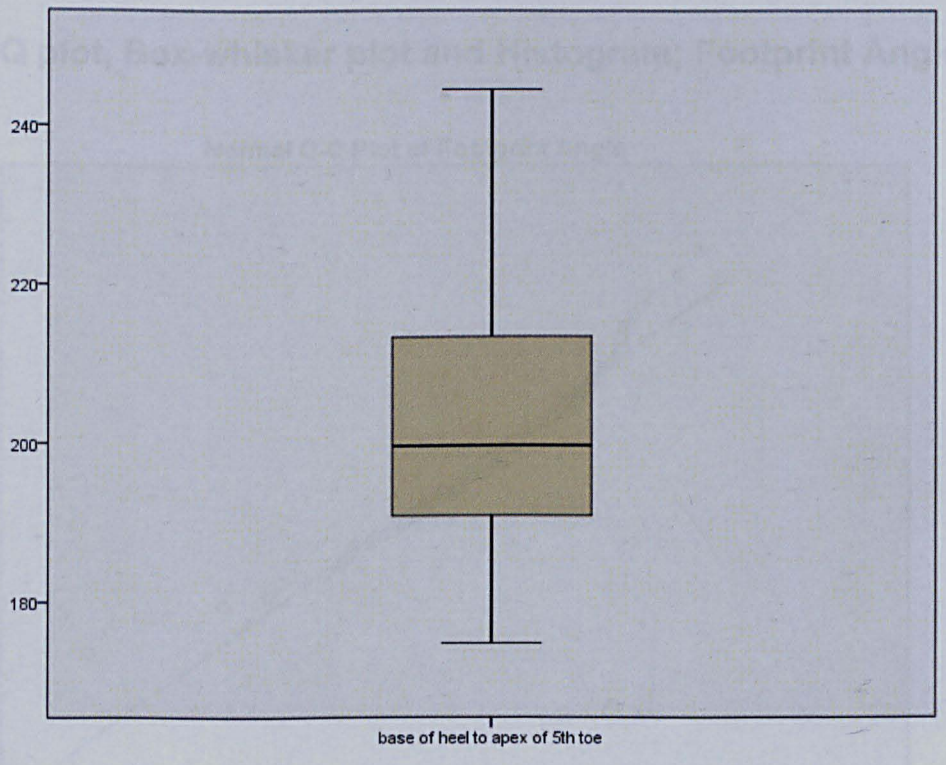
*. This is a lower bound of the true significance.

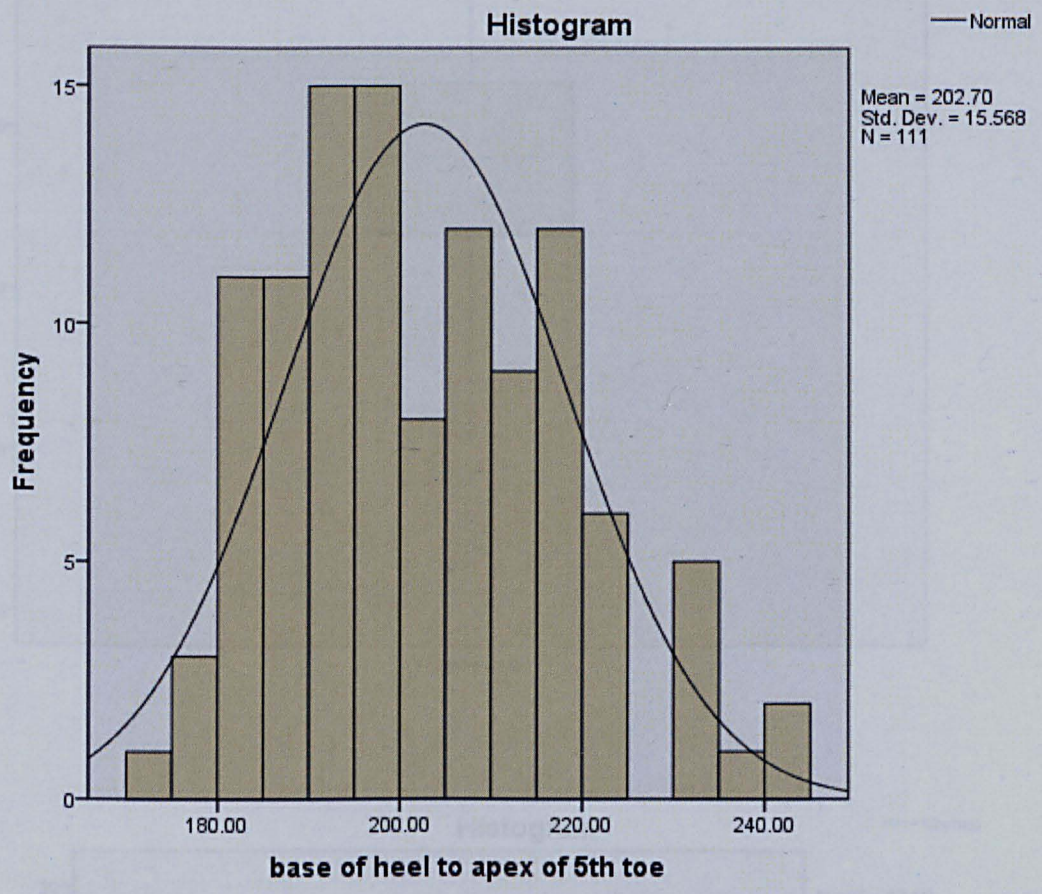
(D) Dynamic

(S) Static

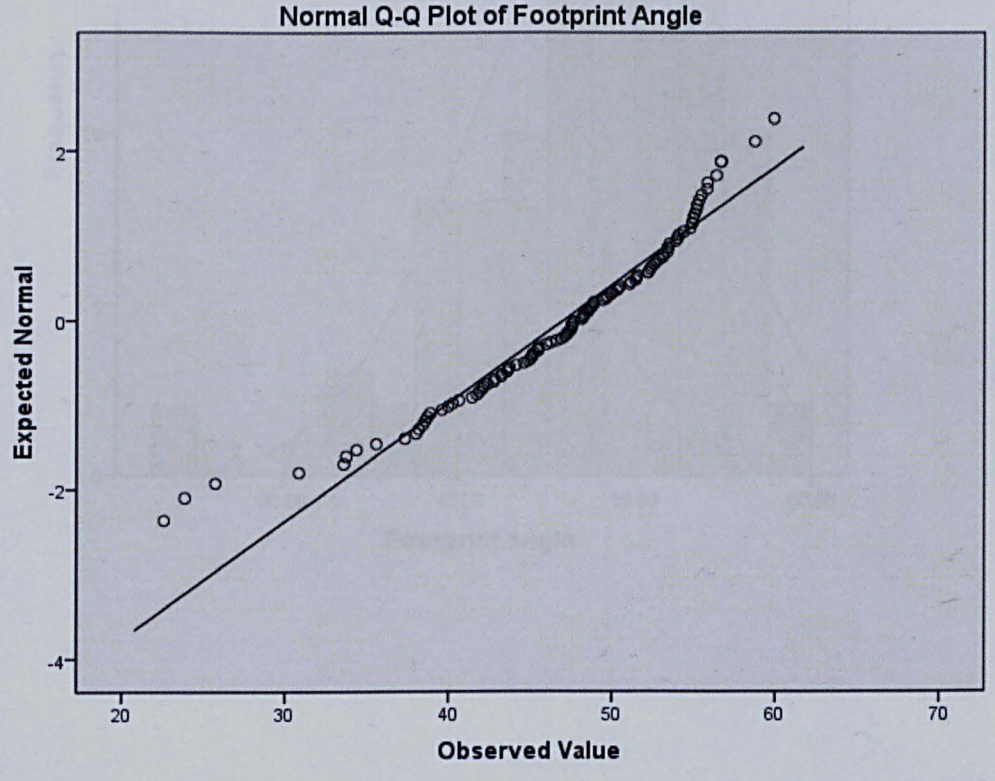
F.2 Q-Q plot, Box-whisker plot and Histogram; Calc_A5

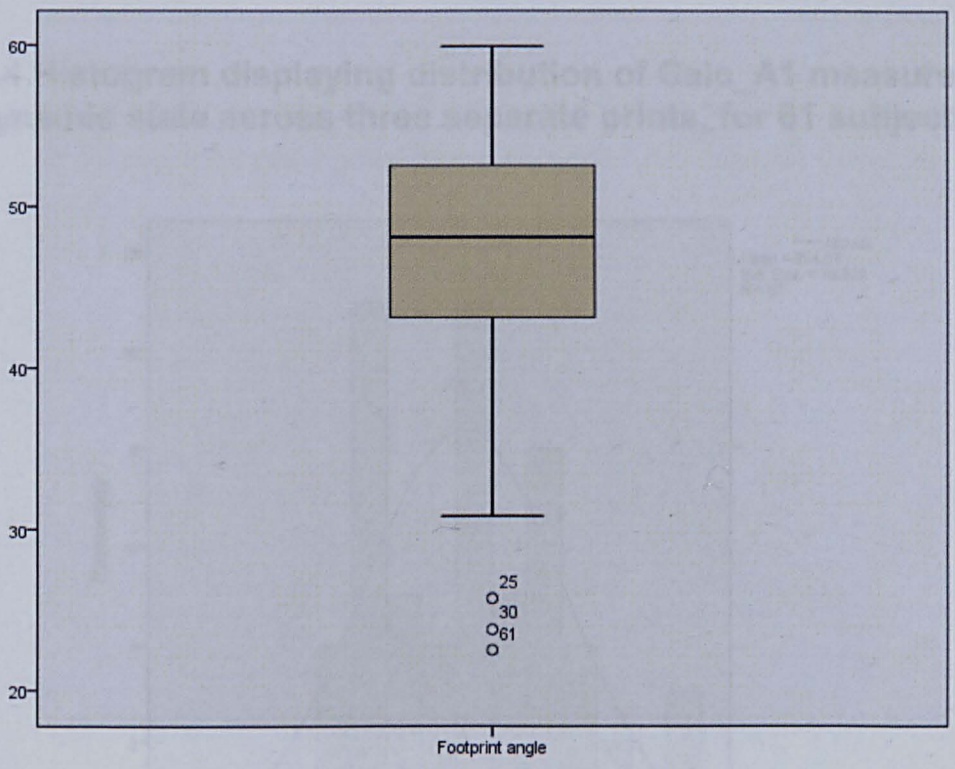


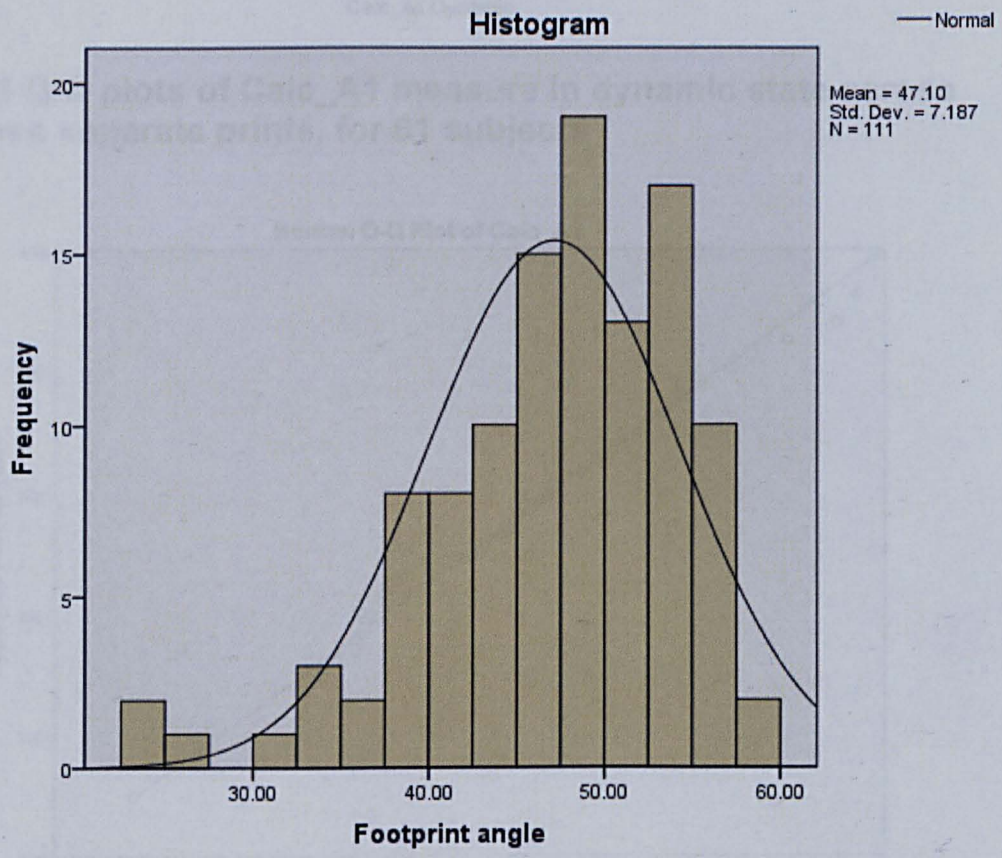




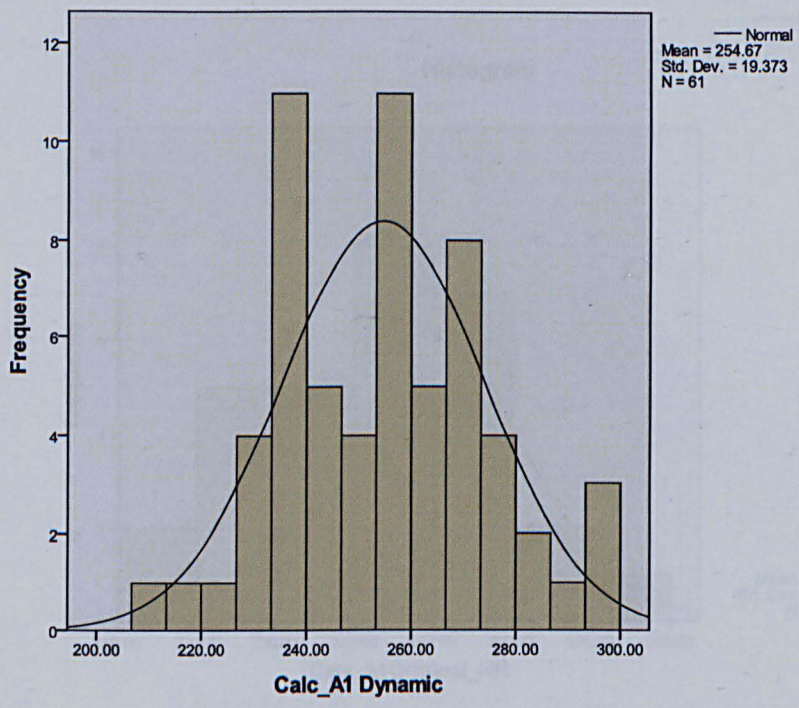
F.3 Q-Q plot, Box-whisker plot and Histogram; Footprint Angle



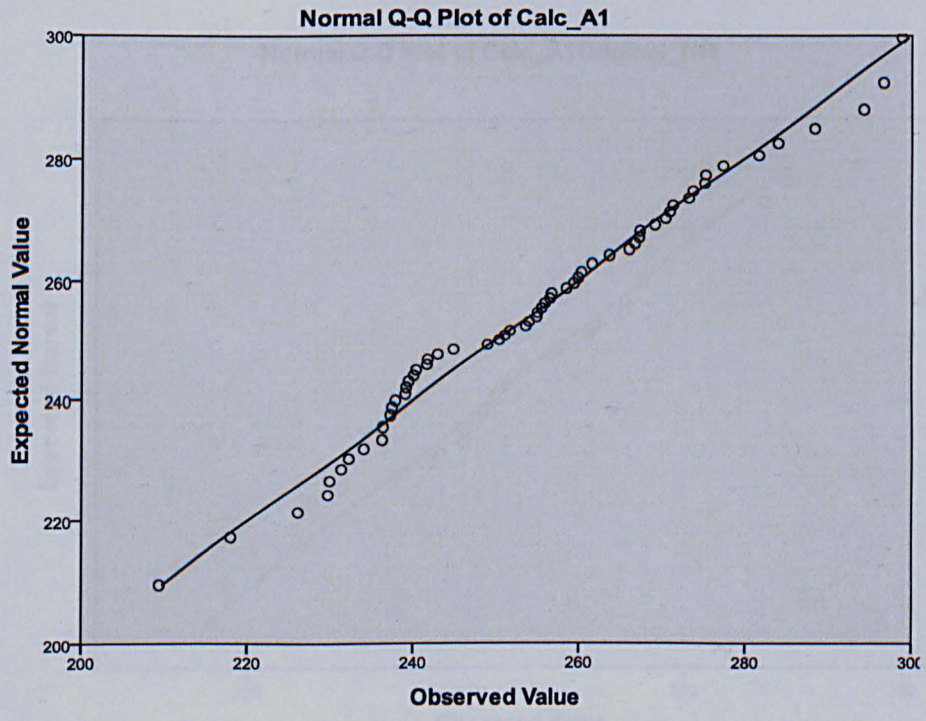

60
50
40
30
20
25
30
61
Footprint angle


Histogram
Normal
Mean = 47.10
Std. Dev. = 7.187
N = 111
Frequency
20
15
10
5
0
30.00
40.00
50.00
60.00
Footprint angle

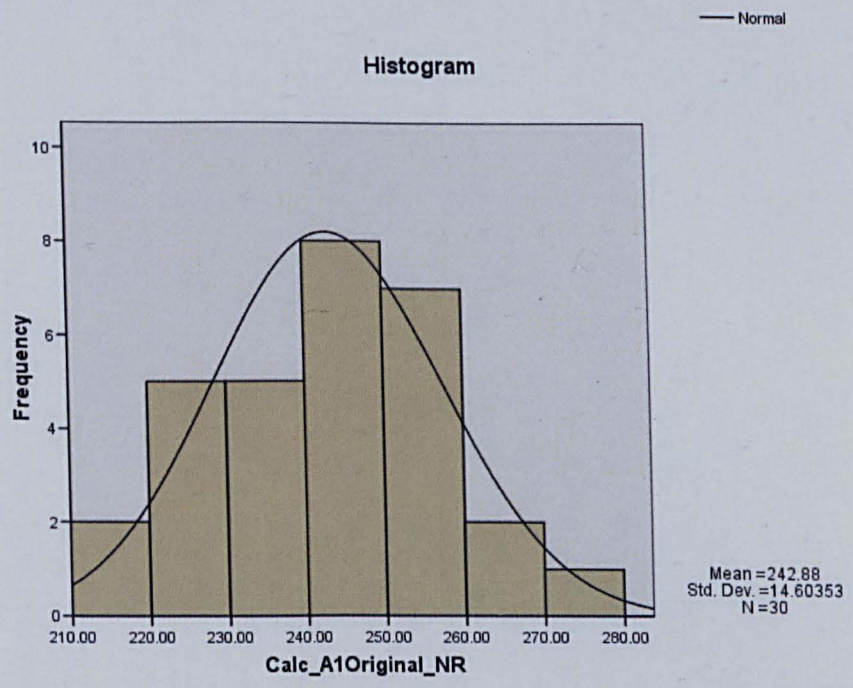
F.4 Histogram displaying distribution of Calc_A1 measure in dynamic state across three separate prints, for 61 subjects



F.5 Q-Q plots of Calc_A1 measure in dynamic state across three separate prints, for 61 subjects



F.6 Histogram displaying distribution of Calc_A1 measurement for 30 footprints recorded by volunteer as part of the inter- rater study



F.7 Q-Q plots displaying distribution of Calc_A1 measurement for 30 footprints recorded by volunteer as part of the inter- rater study