What's still wrong with psychology, anyway? Twenty slow years, three old issues, and one new methodology for improving psychological research.

A thesis submitted in partial fulfilment

of the requirements for the

Degree of Master of Arts in Psychology

at the University of Canterbury

by

Bradley Woods

University of Canterbury

2011

Acknowledgements

I am grateful to a great many people who have contributed in different ways to the production of this thesis.

- Foremost among these is my supervisor, Professor Brian Haig, whose erudition, humility, candour, integrity, and passion is truly an inspiration.
- Professor James Grice, a true pioneer and inventor of OOM, for providing "more light" and a much needed way forward.
- Associate Professor Neville Blampied whose insightful commentary provided the requisite coat of polish.
- Professor Paul Barrett for the presentation on the much needed way forward.
- To both the Psychology and Philosophy faculty at the University of Canterbury, for their collective knowledge, dedication, and zeal that help make "UC" such a fabulous institution for higher learning.
- To Tanya and Stephen for making their house, especially the deck, such a philosophical paradise.
- To Diana and Alistair for a dictionary at just the right time.
- To Alex for the discipline.
- Finally, to my family for all the love, respect, and support a "professional student" could need. My love and gratitude.

Table of Contents

Table of Figures

Abstract

Recent retrospectives of classic psychology articles by Meehl (1978) and Wachtel (1980), concerning problems with psychology's research paradigm, have been viewed by commentators, on the whole, as germane as when first published. However, no similar examination of Lykken's (1991) classic criticisms of psychology's dominant research tradition has been undertaken. Twenty years on, this thesis investigates whether Lykken's criticisms and conclusions are still valid via an exposition of three contentious issues in psychological science: the measurement problem, null hypothesis significance testing, and the granularity of research methods. Though finding that little progress has been made, Observation Oriented Modelling is advanced as a promising methodological solution for improving psychological research.

Chapter 1: Introduction

Classic Articles

Retrospectives

In recent years, two retrospectives of classic psychology articles have appeared in *Applied and Preventative Psychology*. The first, published in 2004, reprinted and commented on Paul Meehl's (1978) seminal article from the *Journal of Consulting and Clinical Psychology*: "Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology". The second, published in 2007 and reprinted with comments, was Paul Wachtel's (1980) *American Psychologist* essay, "Investigation and its discontents: Some constraints on progress in psychological research". Although elucidating, analysing, and appraising different issues, both authors lamented the state of psychological research.

Meehl

Twenty Difficulties with Psychology

Meehl's article began with a discussion of twenty problems in "scientizing the human mind" (Meehl, 1978, p.808). Everything from units of measurement, ethical constraints, and context dependent stochastologicals[1] to idiographic methodologies and the import of cultural factors were documented. In so doing, Meehl offered "one of the most incisive reviews of the problems inherent in studying complex psychological phenomena in existence" (Hinshaw, 2004, p.39). However, this was not the article's only contribution.

The Feeble Practice

The central postulate "particularly hammered home" (Miller, 2004, p.62), and "the lasting memory most readers have" (Hinshaw, 2004, p.40), concerned the critique of significance testing. Highlighting the trivial, but statistically significant, differences between means of large samples, Meehl demonstrated that the null hypothesis is quasi-

[1] Meehl uses this term to describe the probabilistic laws that undergird psychological conceptualisation, in contrast with the strict law-like inference given by the term *nomological* when applied in the natural sciences.

always false: "Putting it crudely, if you have enough cases and your measures are not totally unreliable, the null hypothesis will always be falsified, *regardless of the truth of the substantive theory*" (Meehl, 1978, p.822; italics in original). As a result, psychological hypotheses are subjected to only "feeble danger" (p.821) of refutation; psychological theories become like old generals; "they never die, they just slowly fade away" (p.807), and psychological science "shows a disturbing absence of that *cumulative* character that is so impressive in disciplines like astronomy, molecular biology, and genetics." (p.807; italics in original).

Wachtel

The Predilection with Experimentation and Theoretical Laxity

Psychological theory also came in for criticism in Wachtel's (1980) article, though for entirely different reasons. The lack of recognition, encouragement, and opportunity for theoretical psychologists, and an over-reliance on experimental work, to the detriment of observational and conceptual undertaking, were cited as the most prominent problems. Wachtel contended that in combination they constituted "a kind of intellectual Gresham's law" (Wachtel, 1980, p.401), wherein bad practice had driven out good theorizing. Bemoaning the resulting "misleading findings and one-sided conceptualizations" (p. 407), Wachtel echoed sentiments famously expressed by Wittgenstein: "in psychology there are experimental methods and conceptual confusion" (Wittgenstein, 1958, p.132).

The Perils of Productivity

Going further, Wachtel identified pecuniary incentives and researcher output as additional issues exacerbating the problem of poor psychological theorizing. In warning of the distorting role of grants that drive research programs, along with problems inherent in an overemphasis on publication quantity, rather than publication quality, Wachtel "accurately foretold the future of our field" (Ceci & Williams, 2007, p.14). When the focus on publication quantity and incentive issues are amalgamated with the preceding concerns with theoretical laxity, Wachtel concluded an "industrial model of intellectual endeavour" (p .403) had been created that encouraged "*activity* at the expense of *thought*" (p.399; italics in original), a conclusion noted as "remarkably contemporary" (Cuthbert, 2007, p.15), and striking in its "prescience and timeliness" (Lilienfeld, 2007, p.1).

Retrospective Summary

It should be noted that the articles of both Meehl and Wachtel are not without criticism. In laying the blame for the deficiencies of significance testing solely at the feet of Ronald Fisher, Meehl overlooked the contributions of Jerzy Neyman and Ergon Pearson to the currently used, hybridised, methodology (Gigerenzer, 2004). Similarly, Wachtel's argument, concerning the dangers of valuing research quantity over quality, is somewhat blunted by research showing the most prolific authors have the highest, as well as the lowest, quality of research output (Kazdin, 2007; Simonton, 2004).

Notwithstanding these criticisms, it is a telling and somewhat distressing indictment of the state of psychological science that both articles have been viewed by commentators, on the whole, as being as germane now as when first published (Hinshaw, 2004; Lilienfeld, 2007; Wakefield, 2007; Walker & Mittal, 2007; Waller, 2004; Wampold, 2004). Though overlapping to a small degree, the issues outlined in both articles can be seen as largely complementary. Whereas Meehl exposed the collective failure of psychologists to subject theories to rigorous tests, Wachtel emphasised "in one way or another…the reward structure of our field" (Wachtel, 1980, p.399). Taken together, Meehl and Wachtel's arguments are cogent and conclusive. The unsatisfactory progress of psychology is a direct result of both the dominant research tradition, and the institutional structures that buttress it.

Lykken

Bad Tradition

The view that problems in psychology can be attributed to the research tradition and accompanying support structure is also advanced by David Lykken. In his now classic 1991 essay, "What's wrong with psychology anyway?", he continued where Meehl and Wachtel had left off. Citing depressing figures concerning the number of successful grant applications, journal rejection rates, and the contribution to cumulative knowledge of published articles along with article author and readership rates, Lykken concluded that something is wrong with the established research tradition in psychology. So much so that "It is hard to avoid the conclusion that psychology is a kind of shambling, poor relation of the natural sciences" (Lykken, 1991, p.14).

Castles in the Sand

Like Meehl, in whose honour the essay was written, Lykken characterised the lack of cumulative development of theory in psychology as "just this year's ant hill, most of which will be abandoned and washed away in another season" (p.7), with null hypothesis significance testing, "the favourite ritual" (p.30). Using three historical examples, he also demonstrated that much psychological research fails to replicate before he asserted that the substantial work in psychology over the last hundred years was more neuroscience than psychology. Additionally, in echoing Feynman's (1986) sentiment, whereby psychologists copy the form, but not the substance, of the natural sciences, Lykken contended psychology exists as a "Cargo-cult science"[2] (p.13).

Not Quite the Ailing Lady

Despite his "idiosyncratic and sometimes overstated" critique (Lykken, 1991, p.37), Lykken's message was not all negative. Finding little evidence that psychologists are cognitively inferior to their natural science cousins, acknowledging that experimental control is very difficult, and that the human mind encapsulates the most complicated mechanism known to man, Lykken asserted that psychology is "more difficult, more intractable, than other disciplines" (p.15). Although contending that psychologists have "an abundance of bad habits" (p.29), the good news was that every bad habit jettisoned would improve psychological research. To this end, Lykken nominated the overuse of scientific jargon and over-reliance on significance testing as but two, prominent bad habits (Lykken, 1991).

Although offering methods for improvement, and expressing hope for future research, Lykken's general theme is not oversold. In examining and extending the work of earlier contributors, Lykken showed psychology, and more particularly, its predominant research tradition to be anathema to scientific advancement. Though both Meehl and Wachtel's articles have stood the test of time, one might wonder whether twenty years on, given improvements in both technology and methodology, Lykken's suppositions still hold true?

[2] This is an analogy referring to a tribe that observes cargo arriving during wartime and who try to duplicate the results after war's end by building an airport, control towers, and antennae out of bamboo. Although everything looks the same, the planes fail to land.

Thesis Overview

Outline and Aims

This thesis answers the question of whether Lykken's critique is still valid by undertaking a theoretical exposition of three critical issues addressed respectively by Meehl, Wachtel, and Lykken. Specifically, the thesis updates, synthesises, and evaluates the theoretical and methodological issues surrounding: the measurement of psychological constructs, the role of null hypothesis significance testing (NHST) in theory evaluation, and the tension between the search for general laws and the understanding of individual cases in psychology. In addition, the new statistical methodology of Observation Oriented Modeling (OOM) is evaluated as a means of alleviating, and/or avoiding many of the problems associated with each issue.

Although the recent retrospective commentaries on the articles by Meehl and Wachtel are valuable contributions, in targeting one article each, only a restricted evaluation of some of the critical issues in psychology was articulated. Taking advantage of more contemporary work and possessing fewer word and time constraints than journal articles, this thesis undertakes a deeper, more detailed, and integrated analysis of each of the issues. Furthermore, in documenting and evaluating OOM as a prospective methodological nostrum, the thesis makes a new and unique contribution to the psychology literature. Further details, including the chapter outlines, are summarised below.

The Measurement of Psychological Constructs

The majority of contemporary psychometricians stand by the definition of measurement formulated by Stevens in 1946, namely, that measurement is the assignment of numerals to objects or events according to rule (Michell, 1997). However, few psychometricians are aware that this definition emerged in response to the findings of the Ferguson Committee, which opined that psychophysical methods did not constitute scientific measurement (Ferguson et al., 1940). Despite its now widespread acceptance, Stevens' definition of measurement is not without criticism, and Chapter 2 investigates the

conceptual and theoretical problems with it, and with psychological measurement in general.

The chapter begins with an historical overview documenting early contributions on the issue from such luminaries as James McKeen Cattell, Edward Lee Thorndike, and Gustav Fechner. This exposition provides the backdrop for the analysis of the Ferguson Committee's findings and Stevens' subsequent response. Drawing from the more recent work of Michell (1999; 1997), the problems with Stevens' definition are delineated, particularly its failure to satisfy the quantitative axioms formulated by Holder (1901) and the philosophical issues surrounding operationism.

From there, the chapter explores two alternatives to the traditional approach to measurement, those of Rasch modelling and the theory of conjoint measurement. An analysis of the empirical task in quantification is elucidated next via an appraisal of the recent argument presented by Trendler (2009). Consistent with the contentions of both Michell (1999, 1997) and Trendler (2009), I conclude that psychology, in all but a few instances, has failed to prove that its measurement regime is quantitative and is unlikely to be so in future. Evaluating three options, I suggest that without more interest in conjoint measurement theory, non-quantitative statistical analysis tools such as OOM represent the best conceptual value moving forward.

Null Hypothesis Significance Testing

Though Sir Ronald Fisher is commonly credited with popularising significance testing, little is recognised of the influence William Sealy Gosset had on the early development of statistics used in psychology (Ziliak & McCloskey, 2008). Beginning with Gosset, a historical exposition traces the development of null hypothesis significance testing that includes the contributions of Sir Ronald Fisher, Jerzy Neyman, and Egon Pearson. The hybridised methodology employed by psychologists today is shown to be an incompatible marriage of convenience of competing statistical philosophies.

Examining the assumptions underpinning the modern hybridised NHST methodology, the conceptual and logical inadequacies of such a statistical amalgamation are delineated. Meehl's conjecture, sequence effects, and Lindley's paradox are but several prominent

examples. The contention that NHST can and should be employed as a method for assessing sampling uncertainty is also evaluated and rejected.

Subsequently, contemporary findings on the growth of NHST, attempts at statistical reform, and the level of understanding of significance testing within the psychological fraternity are tabled and evaluated. In conclusion, I contend that despite being the most popular tool for statistical inference, significance testing is largely misunderstood, and misapplied within the domain of psychology; with only rare exceptions should it be employed as a research methodology.

Interindividual versus Intraindividual Research

The research tradition that overwhelmingly characterises contemporary psychological science is one that focuses on the analysis of groups of individuals via statistical aggregates. This interindividual (IEV) research paradigm contrasts with intraindividual (IAV) approaches which emphasise that idiosyncratic patterns of variability within a person are the fundamental level of analysis from which to build generalised psychological knowledge. Though commonly characterised as an idiographic-nomothetic debate, the chapter begins with a definitional assay illustrating how, in drifting from the philosopher Windelband's original conception, a spurious distinction between the terms *idiographic* and *nomothetic* has permeated psychology's research tradition.

The historical development of the rise of IEV at the expense of IAV research is subsequently detailed. It is shown that "the triumph of the aggregate" arose from the conquest of the Galtonian over the Wundtian research philosophies, with causal factors including developments in statistics along with educational, military, and social demand for practical applications of psychological science. Statistics demonstrating the change in fortunes of each research approach are also presented that serve to highlight the serious imbalance of research perspectives that exists within contemporary psychology.

A critical exposition of the strengths and weaknesses of IEV and IAV methods is presented next. Whereas, many of the common criticisms of IAV research are shown to be erroneous and ill-founded, the many touted benefits of IEV research are shown to be premised on dubious conceptual assumptions. Ergodic mathematical theory and empirical research prove that what is common to an aggregate cannot describe that which is

common to an individual. I conclude that redress of the fundamental research imbalance between IEV and IAV research pervading psychology is the only remedy that will allow for the construction of substantive, cumulative, and generalised psychological knowledge.

Observation Oriented Modeling

Observation Oriented Modeling is a novel statistical methodology created by James Grice and founded on the moderate realist philosophy of Thomas Aquinas and Aristotle (Grice, 2011). Emphasising the primacy of real, repeatable, observable, non-aggregated events, OOM seeks to create integrated models that accurately and consistently explain scientific phenomena. Using binary transformations of ordered data elements, termed *deep structures*, sets of observations can be mapped onto one another to infer causal linkages via *Binary Procrustes Rotation*. Predicated on the rejection of traditional positivist methodologies such as NHST, OOM represents an alternative methodology for analysing data and evaluating hypotheses (Grice, 2011).

The chapter begins with an articulation of the rationale undergirding OOM. The technical and statistical elements unique to OOM analysis are elucidated subsequently and include topics concerning matrix rotation, randomised resampling, and generalisation. A hypothetical example follows that illustrates the matrix algebra operations of OOM analysis, and presents statistical outputs from the OOM software, before benefits and criticisms of OOM are considered. In focusing on observations rather than generalities, without relying on assumptions of continuous quantitative measurement, or those assumptions particular to NHST, I contend that OOM is uniquely suited to addressing several problems resulting from psychology's dominant research paradigm.

Chapter 2: The Measurement Problem

Early History of Psychological Measurement

Early Antecedents

Although some form of test or competition has probably been invoked since time immemorial to compare and distinguish between the abilities of humans, the earliest evidence of psychological testing and measurement originates in China (Kaplan & Saccuzzo, 2005). Historical records indicate that oral examinations were given in China every third year in order to help determine work evaluation and promotion decisions, as early as 4000 years ago. The Han Dynasty[3], used test batteries in the fields of civil law, agriculture, geography and military affairs, and it is likely the western world learned about testing mental abilities from the Chinese (DuBois, 1966, 1970). Letters from British missionaries in 1832, for example, encouraged the English East India Company to copy the Chinese system of selecting employees for overseas assignment. Indeed, it was the success of the testing system for the English East India Company that led to the British government adopting a similar regime for its civil service in 1855 (Kaplan & Saccuzzo, 2005).

The inception of modern psychology is commonly dated toward the end of 1879, when Wilhelm Wundt designated space at the University of Leipzig for psychological experimentation, however, it is often overlooked that Wundt was the most significant disciple of Gustav Fechner (Michell, 1999). There had been earlier attempts to found a quantitative psychology, most notably by Johann Friedrich Herbart, who tried to develop mathematical models of the mind, and E. H. Weber, who attempted to demonstrate the existence of a psychological threshold. However, it was Fechner who was successful, in that the science of psychology grew from his undertakings, and the publication of *Elemente der Psychophysik* in 1860, can be seen as the foundation stone of modern psychology, both quantitative and experimental (Michell, 1999).

[3] 206 B.C.E to 220 C.E.

Fechner

Psychophysics

The goal of psychophysical measurement, as conceived by Fechner, was to quantify the intensity of sensations (Fechner, 1860). It was driven by a rejection of Descartes' dualist contention that the mental is not material. Instead, Fechner believed that mind and body were composed of only one substance that could be cognitively related to in two ways, via sensory observation or introspection. It was via the former method, "the common subordination of both the mental and physical realms to the principal of mathematical determination" (Fechner, 1887, p.213), that Fechner hoped to found an exact theory of body and mind relation.

Fechner's first supposition was that a physical stimulus acted upon the nervous system in a manner proportional to magnitude. Presuming that the intensity of sensation was a logarithmic function of the strength of the neural effect, Fechner concluded that stimulus and sensation were related logarithmically (Michell, 1999). However, proving this required a method for establishing the equality of sensation differences. Not believing that subjects could directly judge the quantitative structure of their perceptions, Fechner came up with an experimental method to indirectly measure sensations.

Presenting elements from stimulus sets two at a time, Fechner asked subjects to report which of the pair was greater in some magnitude, for example, which of two weights was heavier. Repeating this procedure many times under controlled conditions, and with varying elements, allowed Fechner to determine the value of a *just noticeable difference* for a given stimulus pair. Invoking an additional premise, that the *just noticeable differences* for different stimulus magnitudes corresponded to equal sensation differences, Fechner was able to show that the intensity of any sensation in a series was measured by the number of *just noticeable differences* between the stimulus producing it and the minimum stimulus threshold (Michell, 1997).

Criticism

Criticism of Fechner's approach was extensive and varied. Bergson rejected the idea that sensations could even be ordered let alone measured. German psychologists such as Delboeuf, Ebbinghaus and Muller contended that Fechner's methods measured only the magnitude of sensation distances and not intensity of sensations. However, it was von

Kries who was one of the most vociferous dissenters (Michell, 1999). Arguing that it was impossible to determine that one pain was exactly10 times stronger than another, von Kries rejected as meaningless the claim that distinct sensation differences were equal.

Whereas von Kries believed intensive physical quantities such as velocity, force, and pressure could be measured via fixing them relative to extensive physical quantities like length, time, and mass, he thought Fechner's fixing of just noticeable differences to equal sensation differences for intensive psychological quantities was arbitrary, with no possibility of being proven correct or incorrect (Michell, 1997). Von Kries' objection faltered, however, as he could do little to prove the difference between the physical and psychological cases, a fact Fechner, famously, was only too happy to acknowledge: "all philosophical counter-demonstrations are, I think, mere writing in the sand' (Fechner, 1887, p.215).

Conclusion

In the end, Fechner's approach became the established psychological *modus operandi* (Michell, 1997). Despite the great philosophical nervousness of his critics, Fechner's methods proved to be too valuable practically for those wishing to adhere to a quantitative imperative in investigating psychophysical phenomena. As will be illustrated further below, the theme of researcher pragmatism overwhelming philosophical reservations continues to this day.

Cattell

Mental Testing

James McKeen Cattell was one of the earliest designated professors of psychology in the United States and had a great influence on the development of American psychology (Boring, 1957). Studying under Galton, a cousin of Darwin, Cattell's doctoral dissertation of 1886 was based on Galton's work on individual differences in reaction times. The first to coin the term *mental test*, Cattell wrote a number of pivotal papers (Cattell, 1890, 1893) that widened the scope of psychology from psychophysics toward the study of intellectual abilities (Michell, 1999).

Like many of his contemporaries, Cattell held a monistic[4] conviction that mental capacities were intricately linked with physical aspects. So much so that the first mental test he developed included tasks such as the pressure exerted by a hand squeeze on a dynamometer and the time taken to move the right arm 50 centimetres (Cattell, 1890). He provided justification for this conviction in his later writings, opining that no difference could be made between mental and physical energy (Cattell, 1902).

The Quantitative Imperative

Like Fechner before him, Cattell believed quantification to be of the utmost importance. Defining measurement as the determination of a magnitude in a standard unit, he saw the ratio as the basis for all measurement (Cattell, 1902). Further, he believed that such measurement was the only way for psychology to advance and attain the certainty and precision of the physical sciences (Cattell, 1890).

Ironically, it was one of Cattell's own students, Clark Wissler who dealt the final blow to the psychophysical approach for measuring intelligence. Wissler calculated the correlations between different psychophysical measures of intelligence and found them too low to be practically useful (Beins, 2010). In another parallel with Fechner, Cattell used pragmatism to justify the conceptual deficiencies of his methods: "It may be at present pseudo-science, in that sense that we have drawn conclusions without adequate knowledge, but it is none the less the best we can do in the way of the application of systematised knowledge to the control of human nature" (Cattell, 1904, p.186).

Conclusion

Though Cattell's approach to psychophysical methods for measuring intelligence ultimately proved futile, he stimulated and perpetuated forces that ultimately led to modern psychometrics (Kaplan & Saccuzzo, 2005). A notable example was his supervision of another student's dissertation at Columbia University in 1898, one Edward Lee Thorndike.

[4] Monism is the conviction that mental and physical phenomena are fundamentally composed of the same singular essence (Feser, 2005).

Thorndike

Credo

Thorndike's first major contribution, after completing his doctorate under Cattell in 1898, was the 1904 publication of *An Introduction to the Theory of Mental and Social Measurement*. This book would go on to become a classic in the field and is still seen as a remarkably sophisticated text (Kaplan & Saccuzzo, 2005). Building on the work, and following the quantitative and practicalist imperatives of his forebears, Thorndike famously declared: "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality…We have faith also that the objective products produced, rather than the inner condition of the person whence they spring, are the proper point of attack for the measurer…the nature of educational measurements is the same as that of all scientific measurements." (Thorndike, 1918, p.16-17). With this statement, Thorndike ushered in a new way of conceptualising psychological measurement (Michell, 1999).

Objective Products

Though admitting psychological measurement instruments had improved since Fechner's time, Thorndike highlighted three fundamental defects that remained, namely, those of ambiguity in content, arbitrariness of units, and ambiguity in significance (Thorndike, 1924, p.219). That is, Thorndike thought that because items in a test might differ in level of difficulty, an observed score was potentially a sum of different magnitudes and thus, not quantitative. Believing a change in focus, from inner sensations of the person to objective outputs produced, was required, Thorndike proposed measurement by relative position as the best way forward. In Thorndike's opinion, it was misleading to judge abilities from casual observations and measures of only a few individuals. Instead, Thorndike believed it was only from a comparison to the total distribution of all individuals that some semblance of value was obtained (Thorndike, 1918).

Practicalism and the Protégé

Having already previously stated that measurement by relative position gave as true a measurement as by a direct unit of amount (Thorndike, 1904), Thorndike drew heavy criticism - most notably, Boring (1920), who claimed Thorndike's method reduced

psychology to operations of mere rank ordering. Once again, practicalism[5] was latched onto as a defence by pro measurement psychologists. The view expressed by Truman Lee Kelley, Thorndike's protégé and for some years America's leading psychologist-statistician (Michell, 1999), is a clear expression of practicalism: "Our mental tests measure something, we may or may not care what, but it is something which it is to our advantage to measure, for it augments our knowledge of what people can be counted upon to do in future. The measuring device as a measure of something that is desirable to measure comes first, and what it is a measure of comes second" (Kelley & Shen, 1929, p.86).

Conclusion

Several themes are readily apparent in the development of psychological measurement from Fechner through Cattell to Thorndike and Kelley.

Monism

Firstly, monism, the belief that the mental and the physical were of the same singular essence, was the default metaphysical stance for all of the early developers. The adoption of such a position inherently implied and allowed for the measurement of the mental, via the physical. Fechner's *just noticeable differences* and the physical tasks in Cattell's mental tests are two germane examples.

Quantification

Secondly, the emphasis on quantification, and the desire to adhere to principles of empirical investigation similar to those used in the natural sciences, was the driving force behind the methodological developments of each contributor. The effects of this were two fold. Attention shifted from metaphysical concerns to instrumental ones and issues with measurement gradually became assimilated to statistical issues. Thorndike's measurement by relative position, and Kelley's assertion that psychometric theory was a branch of statistics (Michell, 1999), provide solid illustration of this shift.

[5] *Practicalism* is the term Michell (1997) uses to describe a view that science should pursue practical ends.

Practicalism

Finally, when confronted at each turn with either metaphysical or empirical opposition to their work, practicalism was the de rigeur defence of measurement methods. That is, the value of information extracted about human abilities via measurement was thought to outweigh any methodological or conceptual objection. Indeed, as seen later, this thought remains a common fall back position to this day.

The 1940s Watershed

The Ferguson Committee

An Extensive Foundation

Through the 1920s, debate continued to rage around the practices of psychological measurement as ever more measurement instruments were designed, implemented, and analysed. Though American psychology, particularly its measurement regime, was under attack, psychology in Britain was even more vulnerable. As late as 1939, there were only six chairs of psychology in all of England and psychology did not gain a substantial foothold in English universities prior to the Second World War (Michell, 1999). It was perhaps no surprise then that the measurement issue came to a head in Britain with the convening of the Ferguson Committee.

Established at York, in 1932, the British Association for the Advancement of Science appointed 19 psychologists and scientists to debate the merits of the possibility of quantitative estimates of sensory events (Ferguson et al., 1940). Propelled by the views of prominent philosophers of physics, particularly N.R. Campbell (1920, 1928) and Percy Bridgman (1927), it was commonly held by the scientific community of the time that measurement was necessarily extensive in nature. That is, physical measures were predicated on the *A-magnitudes* of mass, length and time. All other measures, such as the *B-magnitudes* of momentum, density, and temperature, were seen as derived, being mere manipulations of products of the fundamental *A* units (Narens & Luce, 1986). As psychophysics was commonly held to have no fundamental units, numerical operations such as concatenation could not be performed. Thus, it was believed by detractors that measurement in psychophysics was impossible (Borsboom, 2005; Michell, 1999).

The Nature of Measurement

The scientific camp, led by Campbell, easily won the debate but the case against the possibility of measurement in psychology was not conclusive (Borsboom, 2005; Michell, 1999). The opposing positions were perhaps best exemplified by Guild (1938) and Craik (1940). The former, a physicist, thought using the term *measurement* to apply to psychological practices destroyed all meaning of the word. Craik, on the other hand, emphasised the overly restrictive definition in the physical case: "It is important not to base the definition of measurement only on the most stringent instances, such as length; for 'measurement' is applied to scales of temperature, density, time etc., which fail to fulfil one or other of the conditions which are fulfilled by length." (1940, p.343).

What the Ferguson report (Ferguson et al., 1940) really served to highlight was the importance of the definition of measurement. As part of the report, Stevens' *sone* scale of loudness had been debated as an example of putative psychophysical measurement. Stevens, therefore, had an intense interest in its critique and it spurred him to propose a new definition of measurement. It was a definition that would radically alter the common conceptions of measurement in psychology (Michell, 1997).

Stevens' Definition of Measurement

Subjective Quantification

Like earlier psychologists, Stevens endorsed the importance of quantification opining that the history of science was nothing more than man's search for procedures to measure and quantify the world around him (Stevens, 1967). However, unlike Fechner, Stevens believed subjects were capable of quantifying the nature of their sensations. The secret, he believed, relied on satisfying two conditions. Firstly, adopting a scale of true numerical magnitude and secondly, showing that the scale bore a reasonable relation to the experience of the observer (Stevens, 1936b).

With the first condition, Stevens believed that if the numbers of a scale were amenable to arithmetic, the result should cohere with a set of physical conditions. As an example, he contrasted the cases of length, in which two lots of 3 centimetres combined equalled 6 centimetres, with that of electrical induction in which doubling the length of a wire did not double the inductance (Stevens, 1936b). With respect to the later condition, Stevens

believed that so long as the magnitude of a stimulus designated N by a subject was half as great as the magnitude designated 2N, the scale was satisfactory (Stevens, 1936b). Should both conditions be met, as he believed true of his sone scale of loudness, then measurement was the result: "Provided a consistent rule is followed, some form of measurement is achieved." (Stevens, 1959, p.19).

Representationalism and the Four Scales

The form of measurement was one of Stevens' key themes. He believed measurement was possible only because some degree of isomorphism existed between empirical relations among properties of objects, and the properties of the number system. As such, Stevens proposed that measurement was predicated on four different scales with differing permissible mathematical operations. The result was the construction of the now famous nominal, ordinal, interval and ratio scales, and their respective operations of identity, order, difference, and equality (Stevens, 1946).

As noted by Michell (1997), this reconstruction of measurement effectively disarmed Campbell and other dissenters on the Ferguson committee of their staunchest weapon, i.e., the claim that relevant additive relations between sensory intensities were not demonstrable. It also wedded Stevens' scales to the popular statistical methodology of the Pearson-Fisherian tradition, virtually guaranteeing widespread adoption (Grice, 2011). However, if Stevens' representationalism[6] was correct, then, given a realist view[7] of empirical structures, Stevens had another hurdle to surmount, i.e., the logically prior issue of whether relations of the required kind held in a given empirical domain (Michell, 1999). If Stevens' initial reconstruction was masterful, his second could only be described as extremely intrepid.

Operationism

Invoking a definition of operationism espoused by Bridgman (1927), one of psychophysics' staunchest critics, Stevens constructed an operational interpretation of representational number theory. Agreeing with Bridgman that the meaning of a concept was synonymous with the operations used to identify it, Stevens concluded that the

[6] Representationalism is the view that measurement involves the numerical representation and assignment to empirical relational structures.

[7] Realism mandates that there is an independently existing natural world which humans are able to successfully cognize via observational methods, at least sometimes.

empirical relations represented numerically in measurement must likewise be defined by the operations used to identify them. In doing so, Stevens replaced the prevailing natural science attitude of realism with a form of relativistic subjectivism (Michell, 1997).

Stevens' justification came primarily from two sources. The first was Einstein's repudiation of the classical physics' concepts of absolute time and absolute space (Stevens, 1935b). The second was the domain of mathematics and the rules and symbols used to represent discriminable aspects of nature. Combined, they formed the basis for Stevens' famous dictum: "with the ultimate decoupling of the formal, arbitrary, empty, game like aspects of mathematics from the empirical pursuits of the "concrete" disciplines it became clear that *the province of measurement extends to wherever our ingenuity can contrive systematic rules for pinning numbers on things*. The number system is merely a model, to be used in whatever way we please. It is a rich model, to be sure, and one or another aspect of its syntax can often be made to portray one or another property of objects or events. *It is a useful convention, therefore, to define as measurement the assigning of numbers to objects or events in accordance with a systematic rule*." (Stevens, 1959, p.609, italics added for emphasis).

Conclusion

Inversion via Operationism

In adopting a thoroughgoing representationalism, and giving it an operationist interpretation, Stevens turned the classical definition of measurement on its head. The classical concept asserted that numerical measurements supervene on quantitative attributes. That is, if an attribute had the right kind of structure, then numerical measures were intrinsic to it. According to Stevens, however, this was incorrect; measurable attributes supervene on numerical assessment. If a consistent process could be created for assigning numbers to objects then measurement was achieved.

Overriding Acceptance

By appealing to conceptions of measurement held by his harshest critics, such as Bridgman, and cheekily reinterpreting definitions of others, such as Campbell (1920), Stevens deflected the criticisms of the Ferguson Committee (Michell, 1999). By also incorporating the dominant statistical tradition into the conception of measurement,

Stevens provided psychologists with a justification for their measurement practices, one which they were only too happy to accept. Indeed, the contrast of definitions prior to 1951 with those coming after could not be more striking, with the classical conception of measurement almost unseen after 1951. Instead, a variation of Stevens' form prevails, namely, that measurement is the assignment of numerals to observations according to rule (Michell, 1997).

To this day, the standard definition of measurement in psychology, the one that undergirds traditional approaches to quantifying human abilities, is that of Stevens. As noted by Michell (1999), Stevens' definition is so entrenched that Stevens' dictum is often paraphrased without any acknowledgement that the idea originated with him. That this should be so, coupled with the fact that Stevens' definition of measurement has remained more or less unchanged for more than 50 years, is a testament to its durability. Stevens' definition is, however, uniquely at odds with the common conception of measurement in other sciences, and one might wonder whether this schism is justified?

The Problem with the Modern Definition of Measurement

The Classical Conception of Measurement

Definition and Premises

The classical concept of measurement asserts that all measurable attributes are quantitative and rests on two premises: first, that measurement depends on ratios; and second, that quantitative attributes are the only attributes that sustain ratios (Michell, 1999). Whereas, the second premise is predicated on the work of Holder (1901), the first is derived from Book V of Euclid's *Elements* (Heath, 1908).

Euclid

Building on the Aristotelian definition of quantity[8], Euclid showed that a numerical characterisation of magnitudes was possible via the concept of *identity of ratio*. Defining a ratio of magnitudes as a form of relationship between magnitudes of the same aspect

[8] The Aristotelian definition comprises something that, when divisible into constituent parts, contains "a one and a this" (Michell, 1999, p. 26).

with respect to size (Elements, Bk. V, Dfn. 3; in Heath, 1908, p. 113), Euclid created the conditions necessary to identify two ratios. That is, two ratios of magnitudes are identical when, and only when, they are both less than, greater than, or equal to exactly the same numerical ratios. Or, in modern terms, Euclid located the ratio of magnitudes relative to the series of rational numbers (Michell, 1999).

Euclid's achievement cannot be overstated, providing as it did for a conceptual basis of measurement. Notwithstanding, it was incomplete as a full theory of measurement and required extension in three ways. First, the concept of magnitude had to be defined. Second, the relationship between ratios and all numbers, not just rational ones, had to be explicated. Finally, an account of the contexts in which numbers are properly applied, empirically, was required. It was not until 1901, with the publication of Holder's "*The Axioms of Quantity and the Theory of Measurement"*, that these breakthroughs occurred.

Holder

The critical theorem of Holder's (1901) paper showed that if an attribute is quantitative then it is, in principle, measurable. Using the example of a continuous series of points on a straight line, and invoking ten axioms (see appendix 1), Holder proved that a relation of addition amongst a series of three points must implicitly exist. That is, the magnitude of point *A* may always be expressed relative to point *B* by a positive real number, *R,* where $A=R \times B$ and thus, the ratio of *A* to *B* is the measure of *A* in units of *B* (Michell, 1997).

Holder's contribution can rightly be seen as one of the great intellectual achievements in modern mathematics. In proving that the system of ratios of magnitudes in an unbounded continuous quantity is isomorphic to the system of positive real numbers, Holder filled an important gap in the understanding of measurement (Michell, 1999). That is, the possession of quantitative structure is the precise reason why some attributes are measurable and others not. Hence, "scientific measurement is properly defined as *the estimation or discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute"* (Michell, 1997, p.358; italics in original).

It should be noted that there is no logical necessity for any attributes to possess a quantitative structure, even an ordered series of attributes. As highlighted by Michell (2008), Holder's conditions of magnitude (axioms 4-10) entail the conditions of order (axioms 1-3) and not vice versa; see Appendix 1. Thus, the contention that any attribute,

including a series of ordered attributes, possesses quantitative structure is one that cannot be assumed outright and must be subject to validation. With that in mind, attention is now turned back to an analysis of Stevens', and by default psychology's, conception of measurement.

Critique of Traditional Psychological Measurement

Stevens' Sleight of Hand

In presenting an operational rationale for representational measurement, and by introducing the new terminology of nominal, ordinal, interval, and ratio scales, Stevens effectively wiped the concept of quantification from psychology. It also allowed him to conclude that his ratio scales were on a par with measurement sales used in the natural sciences (Michell, 1997). However, Stevens' position has two critical, conceptual deficiencies. The first concerns the nature of quantification and the second, the veracity of operationism.

Quantification

As demonstrated by Holder's analysis above, the practice of measurement requires capturing, either directly or indirectly, the additive structure of an attribute so that ratios between magnitudes of the attribute may be discovered or estimated. In proposing that measurable attributes supervene on numerical assessments, Stevens totally ignored the question of whether attributes are quantitative. This represents a case of accepting a hypothesis without adequate evidence. Without a demonstration, either logical or experimental, of the hypothesised additive structure of the attributes in question, no inference as to their quantitative nature can be drawn, nor critically, whether the procedures used actually measure them. Barrett (2003) summarises this point succinctly: "The problem is not one of 'permissible statistics' or that one cannot produce numerical results from such techniques, but, the status of any conclusions drawn remains in doubt while the quantitative structure of the variables so manipulated remains untested." (p.427).

Operationism

According to Kerlinger (1979), an operational definition of a construct or variable assigns meaning by specifying the activities or operations necessary to measure it. Though this

might result in a plurality of operational definitions for a construct, Kerlinger's advice was not to let such a consideration trouble a researcher because multiple definitions only demonstrated the flexibility and strength of psychology's measurement regime (Kerlinger, 1979). Unfortunately for Kerlinger, and by extension Stevens, this advice presents serious metaphysical problems.

As noted by Suppe (1977) and others (Luce, Steingrimsson, & Narens, 2010; Trendler, 2009), whenever measurement is defined solely in terms of a measurement process, the inevitable result is a multiplication of theoretical terms. For instance, if intelligence is equated with the operations of the Stanford-Binet test, it immediately follows that the WAIS-R and the Ravens Progressive Matrices cannot also measure intelligence. The implicit conclusion, that no instruments can measure exactly the same attribute, is a rather implausible, indeed absurd result (Borsboom, 2005).

Practicalism

Broadly construed, practicalism is the view that science should serve practical ends (Michell, 1997). In this regard there is no doubt that modern psychometric evaluation can be said to serve practical purposes. Modern psychometric instruments are used in a diverse range of fields, and intelligence tests, to take just one example, are accurate predictors of academic achievement and occupational success, among other things (Kaplan & Saccuzzo, 2005; Kline, 1998). One might conclude then that so long as an instrument serves a practically useful purpose, concerns about what is measured and how the measurement is taken can be ignored. Unfortunately, this view is mistaken. The discovery that an instrument *A* predicts behaviour *B* raises a scientific issue, it does not solve one.

The critical issue that remains is why does *A* predict *B*? Of course it's perfectly plausible to hypothesize that *A* predicts *B* because *A* measures construct *X* which in turns causes *B*. However, this is but one of many possible explanations. Until substantiated by empirical investigation, the claim that *A* measures *X* remains completely speculative.

As highlighted by Barrett (2003), the proper role of the scientist is to find explanations for phenomena, not to create statistical indices of some immediate practical value. Foreshadowing these sentiments, Michell (1997) contends that practicalism, if divorced from robust philosophical underpinning, corrupts the investigative process: "If the

methods of science are not sanctioned philosophically then the claim that science is intellectually superior to opinion, superstition and mythology is not sustained." (Michell, 1997, p.356).

Conclusion

It is hard not to be impressed with Stevens' response to the claims of the Ferguson committee that psychophysical measurement was impossible. In changing the definition of measurement, introducing new terminology, incorporating the dominant statistical tradition, and turning the critics own ideas and words against them, Stevens neutralised the Committee's attack and constructed a paradigm that has lasted to this day. However, it is one built on soft metaphysical sand. Measurement cannot be the process of pinning numbers to things.

Having unfolded the logic of quantification, as founded by the work of Aristotle, Euclid, and Holder, it is clear that measurement is the discovery or estimation of numerical relations, viz ratios, between magnitudes of a quantity, and a unit of that quantity. Though this might be loosely interpreted as the assignment of numerals to objects according to rule as Stevens espoused, Stevens' definition is one that is delusional, ill founded, non rigorous, and conceptually bankrupt (Barrett, 2003; Luce, Steingrimsson, & Narens, 2010; Michell, 2000; Trendler, 2009). Ignoring in totality the logically prior task of proving the existence of quantitative structure, Stevens' definition endorses measurement procedures so lax that virtually anything can be conceived as measurement. Attempting to justify such procedures on the grounds of practicalism serves only to reinforce the view of psychology as a pathological science[9], deflecting psychologists further from the metaphysical commitments of scientific measurement.

Despite the problems with the dominant definition of measurement used in psychology today, it would be hasty to automatically conclude that the Ferguson Committee was correct, that fundamental measurement is, in principle, impossible in psychology. The methods of conjoint measurement and Rasch modelling offer alternatives to the traditional measurement approach. The germane question is are they any better?

[9] Pathological science is characterised by Michell (2000) as a situation in which a hypothesis is accepted as true without serious attempt to test it, and without any recognition of its occurring.

Conjoint Measurement and Rasch Modelling

Conjoint Measurement

Definition Expansion

The genesis of the theory of conjoint measurement is commonly attributed to Duncan Luce and John Tukey and their 1964 publication in the *Journal of Mathematical Psychology*: "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement" (Michell, 1999; Narens & Luce, 1986). Luce and Tukey's key insight was to shift the focus away from trying to prove additive relations within attributes to trying to discover additive relations between sets of attributes. Building on the work of Suppes and Zinnes (1963), and employing insights from the mathematics of indifferences curves in economics, Luce and Tukey took the axioms set forth by Holder (1901) and generalised them to contexts with combinations of attributes (see Appendix 2). The subsequent mathematical refinement by Krantz (1964) showed how the traditional psychological definition of measurement could be expanded without sacrificing the classical concepts of measurement (Michell, 1999).

Derived Measurement

In its simplest form, a conjoint structure exists as an ordered structure that can be factored into two (or more) ordered substructures (Narens & Luce, 1986). If it can be demonstrated that some construct *Z* only increases with increases in *X* and *Y*, and if increasing *Y* has the same effect on *Z* as increasing X does, and if both *X* and *Y* are quantitative, then, it can be shown that *Z* is necessarily quantitative as well (Krantz et al., 1971). Essentially, this is a form of derived measurement as used in physics and other natural sciences. In Luce and Tukey's (1964) case, they used attributes of intensity and frequency to derive a quantitative measure of loudness.

It should be noted that conjoint measurement theory has had implications for the natural sciences as well as psychology. Prior to the conception of conjoint measurement, it was unclear how derived measurement worked. Physicists, for instance, knew that the density of an object was the ratio of its mass to its volume, but exactly which kinds of observations sustained the relationship was uncertain. Conjoint measurement theory

changed this by illustrating how density and volume could be seen to trade off against each other, relative to mass (Michell, 1999).

The Power and Problem of Conjoint Measurement

The invention of conjoint measurement theory is one of the most important theoretical developments in psychology. In showing that fundamental measurement does not require a concatenation operation within an attribute, it provides a justification for psychological measurement on a par with that of physics (Borsboom, 2005). One might assume then that this would have ushered in a new wave of psychometric development, and instrument reappraisal and construction, however, this assumption would be incorrect.

Conjoint measurement theory takes psychology only so far along the path to quantification. The problem with many constructs in psychology is finding conjoined constructs that have already been shown to possess quantitative structure, from which to derive measurement. Though conjoint measurement has been demonstrated with psychophysical attributes such as loudness and brightness (Luce, Steingrimsson, & Narens, 2010), and represents a significant achievement, one can readily see the difficulty in trying to do the same for a construct such as extraversion.

The Unexploited Resource

The lack of attention paid to conjoint measurement by the broader psychological community is so striking that it has been described as a revolution that never happened (Cliff, 1992). As an example, a PSYCInfo database search by the author for full text, peer reviewed articles with psychometrics or psychological assessment as subject items for the period 2000-2010 returned 20,198 articles. Adding conjoint analysis or conjoint measurement as a subject item to the search reduced that number to 45. However, even this is overstated as only 3 of those articles directly dealt with quantifying psychological constructs, with the rest referring to consumer research, education or healthcare concerns where conjoint measurement is used to infer participant preferences.

With no research program dedicated to conjoint measurement, and few researchers working to develop its theory, conjoint measurement theory represents an unexploited psychological resource (Michell, 1999). Borsboom (2005) highlights the problem arising from such a lack of attention: "The logical foundation for psychological measurement has

thus become available, only to be neglected by its presumed audience – and psychologists have continued to use the term measurement for everything else." (p.87).

Rasch Modelling

Scale Comparison with Expected Orders

Rasch analysis involves the formal testing of a measurement scale against a mathematical measurement model originally devised by the Danish mathematician Georg Rasch (Rasch, 1960; Tennant & Conaghan, 2007). Specifically, the pattern of results from a set of items on a test is compared against a theoretical pattern of results that are purported to meet the fundamental axioms of measurement. In doing so, Rasch analysis relates the probability of a person getting an item correct to the ability of the person, and to the difficulty of the item, such that the probability increases with the difference between them in a log interval fashion (Kyngdon, 2008). In this sense, like conjoint measurement, Rasch modelling seeks to simultaneously measure two variables via a third (Michell, 2008).

For example, it can be expected that the results on a spelling test of a person with high spelling ability will be quite ordered. That is, the subject will likely spell words correctly up to the point that the words become very difficult. If the pattern of results derived from the spelling test is ordered in such a way that fit the Rasch measurement model, then it is inferred that spelling ability is necessarily quantitative.

Problems with Rasch Modelling

For over 30 years the idea has been advanced by psychometricians that Rasch modelling is conjoint measurement or a probabilistic variant thereof (Kyngdon, 2008). Indeed, some have called it no less than the empirical benchmark for the quantitative structure of latent variables (Barrett, 2003). Furthermore, in terms of practical application, Rasch modelling has been used successfully to improve traditional psychometric instruments and related outcomes (Pallant & Tennant, 2007). However, Rasch modelling is not without several conceptual problems, including the claim that its modelling is a form of conjoint measurement.

Latent Variables

An initial problem concerns the nature of latent variables purportedly measured via Rasch modelling. As highlighted by Barrett (2005), even if the outcomes of a scale conform to the Rasch model, that is no guarantee it bears any relation to a substantive, meaningful psychological variable. This was illustrated clearly by Wood (1978) who fitted a series of random coin tosses to the Rasch model, demonstrating the latent variable of coin tossing ability. To assume that a well-fitting Rasch model implies accurate measurement commits the psychometric fallacy noted above, of presuming a quantitative structure from a merely ordinal one (Michell, 2008).

A Quantitative Assumption

Although Rasch modelling follows in the conceptual footsteps of conjoint measurement, by attempting to measure two variables by means of a relation to a third, it shares an assumption reminiscent of Stevens' definition of measurement. Specifically, it is taken as given that test performance is a conjoint structure comprising only person ability and item difficulty. Luce (1987) suggested one way to test this assumption is by subjecting item difficulty and person ability to the axioms of Holder (1901). However, to date, no such investigation has been undertaken (Kyngdon, 2008).

The Rasch Paradox

A further troubling issue with Rasch modelling concerns the nature of error in measurement and is known as the *Rasch Paradox* (Michell, 2008). Mathematically, Rasch modelling is equivalent to a transformation of ordinal Guttman scaling, with the only difference being the addition of an error term (Kyngdon, 2008). Thus, the difference between a merely ordinal and a fully quantitative structure in Rasch modelling is error. In every other form of measurement, eliminating error improves measurement but in Rasch modelling if errors are eliminated, quantitative measurement is impossible. At best this represents a conundrum, at worst, a nonsensical outcome.

Not Probabilistic Conjoint Measurement

Conjoint measurement is concerned with the measurement of any attribute, where it is a non-interactive function of two, or more, other attributes. It thus has a generality exceeding that of Rasch modelling which is concerned only with the constructs of

probability, item difficulty, and person difficulty. Because there is no articulation in the Rasch model of the ordinal and equivalence relations required by conjoint measurement, Rasch modelling and conjoint measurement are not mathematically equivalent. Therefore, it is erroneous to claim that Rasch modelling is simply a probabilistic form of conjoint measurement (Kyngdon, 2008; Michell, 2008).

Conclusion

Creating and/or modifying instruments in a way that achieves fit with the Rasch model is no doubt a considerable achievement and one that has yielded practical improvements compared with traditional psychometric instruments. However, data can fit many models and care should be taken in the interpretation of latent variables and assumptions underlying the Rasch model. Though similar to conjoint measurement, Rasch modelling is not a variant of conjoint measurement and, therefore, cannot be used to determine the quantitative structure of attributes. Hence, the conclusions drawn from Rasch modelling are subject to the same criticism as that of traditional methods. That is, without proof, the claim to truly measure variables remains speculative at best.

In contrast, the theory of conjoint measurement should rightly be seen as an historic breakthrough for measurement theory in psychology. In showing that measurement does not depend solely on the extensive structure of attributes, it lays the foundation for a measurement system predicated on the same principles as those in the natural sciences. It is a surprising and somewhat frustrating fact, then, to find so few resources being employed to extend and develop conjoint measurement. Though some psychophysical attributes have proven amenable to conjoint measurement, thus far, psychological attributes have not. Some have expressed optimism that this can change with the development of substantive theory (e.g., Borsboom, Mellenbergh, & van Heerden, 2003; Goldstein & Wood, 1989). A recent argument challenges this contention, however, suggesting that regardless of improved theory, the requisite level of experimental control needed for measurement in psychology is not achievable (Trendler, 2009).

The Empirical Task in Psychological Measurement

Trendler

The Basic Assumption of Quantity

That the possession of quantitative structure is the fundamental reason why some attributes are measurable, and others not, should now be clear. Verifying the existence of such structure can occur either directly, by satisfying the axioms of Holder (1901), or indirectly, by meeting the conditions prescribed by the theory of conjoint measurement (Luce & Tukey, 1964). Implicit in both of these approaches is the demand of the most basic condition of quantity, namely, that any two magnitudes of the same quantity are either identical or different (Michell, 1999).

As noted by Trendler (2009), verifying this condition is an empirical task. It cannot simply be taken for granted that equal levels of some manifest variable, say test performance, correspond to equal levels of a latent variable such as intelligence. Using an analogy with Ohm's law, Trendler demonstrates that the empirical task of quantification in psychology necessarily involves two related procedures. Firstly, manipulating a hypothetical construct and noting the effect of such manipulation on a dependent observable and secondly, meeting the requirement that no disturbances, either systematic or stochastic, impact on the observations.

Control and Manipulation

Trendler (2009) asserts that there is no doubt that psychological variables can be manipulated and controlled. However, he denies that variables in psychology can be manipulated and controlled to the extent that allows for accurate measurement to occur. For this reason, Trendler contends that no psychological variable has ever been measured. Analogous to the difficulty facing conjoint measurement, Trendler cites the interrelated nature of psychological constructs and the dynamic nature of the brain as insurmountable obstacles to empirical manipulation and control.

Via the example of reward manipulations, Trendler (2009) points out the problem with trying to attribute changes in test performance solely to changes in motivation.

Replication with the same person across time, or by comparing different people, implies holding all other possible influences such as attention, learning, and ability constant. However, to allow for valid comparisons requires that these other variables are already quantified. Utilising neuroscience methodologies fares no better, according to Trendler, because the brain is a non localisable system. That is, there is no way to "slice and dice the brain of a test subject" (Trendler, 2009, p.592) in a manner that allows for observations of direct causation that could be attributed solely to one particular psychological construct.

An Overly Strong Case?

At face value, Trendler's (2009) argument seems robust; however, it is not without criticism. One argument advanced by Markus and Borsboom (2011), concerns the condition of equivalence (Holder's first axiom), on which Trendler's argument is predicated. Markus and Borsboom (2011) point out that assessment of equality and inequality are fundamental elements of every measurement system including nominal and ordinal relations, and are not just particular to quantitative structures. Thus, if Trendler's argument is correct, then not only is quantitative measurement ruled out but, absurdly, so is every form of measurement in psychology.

A more generous reading of Trendler might presume that the assessment of equivalence is possible in psychology, just not for quantitative structures. However, this would seem to shift the burden of Trendler's (2009) argument from Holder's first axiom to other axioms which are not discussed by Trendler. The implication is that on one interpretation of the assessment of equivalence Trendler puts forward an overly strong case, but yet, on another interpretation where equivalence is possible Trendler puts forward no case at all (Markus & Borsboom, 2011).

Conclusion

Trendler's (2009) assertion that no psychological variable has ever been measured in psychology is clearly false. Fundamental measurement, the same as that used in physics, has been applied to the constructs of loudness and brightness amongst others by using the theory of conjoint measurement (Luce, Steingrimsson, & Narens, 2010). Similarly, the absurd implications highlighted by the criticisms of Markus and Borsboom (2011) might

make it easy, in one sense, to dismiss Trendler's claims out of hand. However, to do so entirely would be to throw out the baby with the bathwater.

Applied to latent constructs such as introversion or motivation, that have a less clearly defined biological foundation than either brightness or loudness, and where proving Holder's first axiom would seem exceedingly difficult, Trendler's argument is much more cogent. Existing in their natural, non isolatable state, and without requisite control, many psychological phenomena may elude quantification. If this is the case, the implications for psychology are quite serious. Notwithstanding the arguments presented on p.39, Kline (1998) notes that without quantitative data the philosophical justification for psychological methods such as structural equation modelling and factor and regression analysis is missing: "It may be that the task of the new psychometrics is impossible…If this is the case, then the truth must be faced that perhaps psychology can never be a science." (cited in Borsboom, 2005, p.85).

Conclusion

Technical versus Conceptual Progress

Since Fechner's earliest attempts to subordinate the mental and physical realms to mathematical determination via the evaluation of just noticeable differences, there can be no question that psychological measurement has increased by orders of sophistication. Statistical methodologies such as structural equation modelling, factor analysis, and meta-analysis are three prominent examples. In other ways, however, it is clear that there has been little progress. Though the advent of neurological imaging is often touted as a shining example of increasing technological and measurement sophistication, question marks hang over many of the correlations obtained between brain activity and personality measures (Vul, Harris, Winkielman, & Pashler, 2009). Similarly, the issue of whether psychological variables are quantitative remains a question that for the most part remains unexplored and unanswered.

The question of quantification in psychology was one that was initially overridden on practical grounds by followers of Fechner and subsequently transmuted into a statistical concern by Thorndike and Kelley. Then when confronted with criticisms of the Ferguson

Committee, Stevens created a new measurement paradigm. Operationalising the concept of measurement and welding the four scales of measurement to statistical methods of Fisher and Neyman and Pearson, ensured the widespread adoption of Stevens' definition and effectively buried the issue of quantification. This is an unparalleled example of a theory being accepted because it answered questions, rather than investigated because it raised them (Michell, 1999).

The Problems and Consequences of Traditional Measurement

The truth is that operationalising the definition of measurement resulted in a definition so loose as to be farcical. Paradoxically, it leads to a situation where almost anything can be called *measurement* but where no two measures can measure the same thing. No doubt there are examples where the traditional approach to psychological measurement has been shown to be practically useful in illuminating relationships between variables. However, by assuming, and ignoring, the issue of the quantitative status of the relations, such cases remain exercises of uncertain scientific validity. They raise questions about psychological phenomena but do not answer them.

The consequences of adherence to the dominant psychological measurement model has led some to conclude that psychology comprises only trivial exemplars of mostly inaccurate explanations of phenomena; that psychology produces transient descriptions of never-to-be-reencountered situations, that are easily contradicted with replication (Barrett, 2008; Wright, 1997). Some of those sharing similar sentiments have turned to Rasch modelling to improve traditional instruments. Though there has been some success, Rasch modelling is not without problems, paradoxically predicating measurement on error, and failing to prove axiomatically its claim of fundamental measurement.

A Quantitative Science?

The most significant development in psychological measurement has been the introduction of the theory of conjoint measurement. In providing an indirect means of proving quantitative structure, and taking advantage of the interrelated nature of psychological phenomena, the theory of conjoint measurement has shown how

fundamental measurement, equal to that in the natural sciences, is possible. Unfortunately, it has also been a most ignored development. The greatest obstacle concerns the lack of already quantified constructs from which to derive the quantitative structure of additional, related constructs.

Although such an approach is made possible by the interrelated nature of psychological phenomena, it does have a drawback, that is, obtaining the requisite levels of isolation and control necessary for accurate measurement. This is a condition that some believe cannot be attained and which has led others to question whether psychology can be seen as a science at all (Kline, 1998; Schonemann, 1994; Trendler, 2009).

Such a conclusion is unwarranted. There is no necessity for variables in psychology to possess quantitative structure. Similarly, there is no necessity for psychology to concern itself solely with quantitative structures. Non quantitative, or qualitative variables, can be studied in terms of classes and categories and psychology will be no worse for it. It will remain a science.

Future Options

At this juncture, there are three possible ways for psychological measurement to proceed. First, psychologists can continue using the dominant measurement tradition instantiated by Stevens. There may be some practical benefit in being able to loosely predict behaviour via the identification of correlated variables. At best, however, it is the option of an epistemological ostrich, burying the problems of quantification in the sands of practicalism.

Second, the application of the theory of conjoint measurement in order to further the discovery of psychological phenomena can be pursued. Philosophically, this is a far stronger alternative. However, there is no guarantee to how many further constructs can be shown to possess additive structure using this method and, as noted by Kyngdon (2008), it also requires the kind of substantive theorising that psychologists studiously avoid.

Third, new and alternative methods for studying psychological phenomena, which do not rely on assumptions of quantitative structure, can be utilised. Probabilities of occurrence, order relations, and structural analysis are germane methods that can be used

to identify and evaluate psychological phenomena. Indeed, the new method of Observation Oriented Modelling documented in chapter 5 employs all three of these characteristics to appraise psychological theory.

Conclusion

There can be no doubt that some psychological attributes are quantitative. Continuing to assume that all are, by utilising only linear statistical methodologies to measure psychological variables, is at best bad science and at worst, fraudulent. It is with rare exceptions that psychological phenomena have been shown to possess quantitative structure. Without a true revolution stemming from interest in the theory of conjoint measurement, or the resolution of the problems highlighted by Trendler (2009), psychology many variables in psychology will remain un-quantified. Option three, therefore, represents psychology's most viable class of method for tackling issues raised by the measurement problem.

Chapter 3: Null Hypothesis Significance Testing

History

Early History

The fundamental logical elements of what is today referred to as *null hypothesis significance testing* (NHST) were present in scientific papers as early as the 1700s (Kline, 2004). For example, in 1714 Daniel Bernoulli, one of the founding fathers of probability theory, conducted tests of significance on the randomness of planetary orbits. Similarly, in 1773 Laplace tested the hypothesis that comets come from outside the solar system (Ziliak & McCloskey, 2008). A systematic method did not emerge, however, until Karl Pearson developed the first modern test of significance, the Chi Square, in 1900 (Hubbard & Lindsay, 2008).

Although Pearson may have been the founding father, it is Sir Ronald Fisher who is credited for producing the scaffolding on which psychological statistics was founded, and to whom responsibility can be laid for popularising significance testing (Wright, 2009; Zilliak & McCloskey, 2008). The many editions of his books, *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935), have proved to be two of the most influential texts in psychology (Gigerenzer, 2000; Hubbard & Lindsay, 2008; Yates & Mather, 1963). Though today it is quite rare to see Fisher's name associated specifically with the NHST method now taught in psychology, it is even rarer to see the name of William Sealy Gosset mentioned, or credited, with the development of NHST (Gigerenzer, 2004, Hubbard & Bayarri, 2003).

Gosset

Beer and Mathematics

Born in Canterbury, UK, in 1876, Gosset studied at Winchester College and New College, Oxford where he graduated with a first class in mathematical moderations in 1897 and chemistry in 1899. A staunch pragmatist, Gosset was focused on the practical applications of mathematics to issues of substantive and economic significance (Ziliak &

McCloskey, 2008). Thus, it was while working at the Guinness Brewery in Dublin that Gosset saw the need for a small sample test to distinguish the quality of varieties of hops and barley. Gosset's subsequent investigations resulted in the publication of two revolutionary papers.

The t-Test

Using a mechanical, turn crank calculator, Gosset created the tables and formulae for what is now known as the z test, which he published as "*The Probable Error of a Mean*" and "*The Probable Error of a Correlation Coefficient*" in 1908. Though the forerunner of the t-Test, the ideas in the papers also laid the foundation for the creation of Monte Carlo simulation, the concepts of power, and what came to be called "the alternative hypothesis" (Ziliak & McCloskey, 2008). The t-Test itself didn't appear formally until 1922 when Fisher substituted n-1 for n as the sample size in Gosset's calculations, and Gosset's original tables were corrected and renamed as the Student's t-Test of significance (Ziliak & McCloskey, 2008). The name was a reference to the pseudonym, *Student*, necessarily used by Gosset on his earlier papers, as a condition of his employment with Guinness Brewery.

Copyright Cover-up

Though Gosset copyrighted his original tables in 1908, 1914, and 1917, Fisher copyrighted the t-Tables in his own name in 1925 when he published *Statistical Methods for Research Workers*. As noted by Ziliak and McCloskey (2004), though Gosset was mentioned, each successive edition gave less credit to Gosset for his contribution. Indeed, Fisher's later book *The Design of Experiments* makes no mention of Gosset at all even though the same tables appear. These omissions no doubt helps explain the rarity with which Gosset's name is recognised by researchers today.

Fisher

Agronomy and the Direct Probability

The value of Gosset's calculations to Fisher become readily apparent when one considers Fisher's position on the nature of methodology and scientific investigation. Believing inductive inference was the only way new knowledge was ever accumulated in science (Fisher, 1966), Fisher initially dabbled with Bayesianism (Zabell, 1992). However,

thinking the lack of a priori distributional information unacceptably subjective, Fisher quickly came to reject Bayes' rule as a sound method of scientific enquiry (Gigerenzer, 2000; Hubbard, 2004). Seeking a more objective methodology of inductive inference, Fisher renounced the inverse probability of Bayesianism, the probability of a hypothesis (H) given the data (D), Pr(H|D), and instead focused on the direct probability, Pr(D|H).

Fisher's view of inductive inference was strongly influenced by the practical problems of agronomists, who were concerned with the interpretation of agricultural field experiments on crops, and with whom Fisher had direct contact while statistician at Rothamsted experimental Station under Sir John Russell (Yates & Mather, 1963). Fisher recognised early on that the analysis of variance provided a powerful technique for explaining the individual differences in crop yields due to different fertilisation regimes, and it was from such work that Fisher's exhortation of the estimation of population parameters from small samples was born (Rodgers, 2010; Yates & Mather, 1963). The introductory chapter of *The Design of Experiments* best exemplifies Fisher's thinking in this regard: "We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression" (Fisher, 1966, p.3-4).

The Rare Event

In a general sense, Fisher's rationale was falsificationist, believing that no scientific theory could ever be proved with surety (Gigerenzer, 2000). As such, Fisher thought the best scientific investigators could hope for was a whittling away and rejection of plausible competing theories. Thus, Fisher advocated determining the probability of a result, in addition to more extreme ones, assuming a null hypothesis of no or limited effect or relationship. Any rare and unexpected observations then, constituted better inductive evidence against the null hypothesis. Or as Fisher succinctly stated: "Either an exceptionally rare event has occurred or the theory is not true" (Fisher, 1959, p.9).

Additional Considerations

Fisher's model of significance testing required researchers to know in advance the possible outcomes of an experiment and to randomise all variables so as to allow valid interpretation of results (Fisher, 1966). Fisher also prescribed that only the null

hypothesis was to be tested, with any alternative hypothesis or theory being too inexact to allow proper inference (Fisher, 1966). Similarly, Fisher exhorted experimenters to pay heed to sensitivity, the ability to discriminate between significant and non-significant findings, via the size of the experiment (Fisher, 1966).

Initially, Fisher also advocated the use of a 5% threshold as a convention by which to judge the significance of an experimental outcome. However, Fisher later recanted stating instead that experimenters should report the actual level of significance obtained (Gigerenzer, 2000). In addition, the publication of both significant and non-significant findings was encouraged so that cumulative frequencies of phenomena could be established (Fisher, 1966). Such relative frequencies were important to Fisher for he only considered a phenomenon established when experiments could be conducted that rarely failed to yield statistically significant results: "…it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis." (Fisher, 1966, p. 16).

Summary

Under Fisherian logic, a significance test is a tool of inductive inference. It is used solely to evaluate the probability of having obtained evidence against the null hypothesis, assuming the null hypothesis to be true. The null hypothesis need not be one of zero difference with the exact significance level to be reported, and lacking specified alternative hypotheses, attention to issues of experiment sensitivity are to be accorded keen interest. Finally, Fisher's hypothesis testing was to be used in cases where little was known about the problem at hand, with other tools of inference available as investigations progressed (Gigerenzer, 2000).

As noted by Gigerenzer (2000), Fisher sometimes changed and reversed his logic of inference at different periods. Thus, Fisher's writings possess an elusive quality and many have questioned Fisher's exact meaning and interpretation of key concepts. Gosset for example, eschewed the lack of consideration given by Fisher to the "pecuniary advantage" of experimentation (Ziliak & McCloskey, 2008). However, it was Egon Pearson and Jerzy Neyman who would famously become the staunchest critics, and favourite target of, Fisher's philosophy.

Neyman and Pearson

The Alternative Hypothesis

The Neyman-Pearson statistical paradigm (1928a, 1928b, 1933) is now considered to be the norm in classical statistical circles and it grew out of dissatisfaction with several elements of the Fisherian approach to hypothesis testing (Hubbard, 2004; Nickerson, 2000). The main difference between the two approaches, and the one that caused the most contention over the years, concerned the role of an alternative hypothesis. Whereas Fisher thought testing the null hypothesis alone was sufficient, Neyman and Pearson, in accordance with Gosset (Hubbard, 2004), believed the formulation of an alternative hypothesis to be required: "…in addition to H_0 [the null hypothesis] there must exist some other hypotheses, one of which may conceivably be true…This is quite important. The fact is that, unless the alternative H_a is specified, the problem of an optimal test of H_0 is indeterminate." (Neyman, 1977, p.104).

New Concepts

Rejecting Fisher's ideas about hypothetical infinite populations, Neyman and Pearson also believed results should be predicated on the assumption of repeated random sampling from a defined population (Hubbard, 2004). When coupled with the requirement for an alternative hypothesis, the concepts of false rejection (Type I error) and false acceptance (Type Π error) of the null hypothesis were born. Complementing these concepts was the idea of the power of a statistical test, the probability of rejecting a false null hypothesis (1-Beta). Though similar to Fisher's concept of sensitivity, the Neyman-Pearson concept of power was more complex, looking at more than just sample size, and was invoked prior to, not after, an experiment had been completed (Hubbard & Bayarri, 2003). Consequently, the Neyman-Pearson approach sees hypothesis tests as rules of inductive behaviour, concerned with decisions between alternative courses of action such that, in the long run, error is minimised (Neyman & Pearson, 1933).

Philosophical Differences

The philosophies underlying both the Neyman-Pearson and Fisherian approaches to hypothesis testing can be seen to be quite different, then. Whereas Fisher gave an epistemic interpretation to a significance test, believing significance indicated truth or

falsity of a particular hypothesis, Neyman and Pearson provided a behaviouristic rationale, with a significance test simply providing a rule for decision making regardless of the researcher's degree of belief (Gigerenzer, 2000). This was a point on which Neyman and Pearson (1933) were quite explicit: "We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis" (p. 290-291).

Summary

Neyman-Pearson hypothesis testing represents a markedly different approach to statistical inference from that of Fisher. Philosophically, the focus is on inductive behaviour rather than inductive inference. Conceptually, the method includes the addition of an alternative hypothesis, the assumption of repeated sampling from a known distribution, and the concepts of error and power. Indeed, the differences are so stark one might question how the Fisherian and Neyman-Pearson approaches could both be legitimately applied to the same problems.

The Inference Revolution

Changing Research Practices

The *inference revolution* is a term coined by Gigerenzer and Murray (1987) to describe the dramatic shift in research practices that occurred in psychology between 1940 and 1955. Prior to 1940, the dominant tradition had been a Wundtian one, with experimentation of single participants the norm. Beginning in 1940s America, however, this began to change to treatment group experiments in which group means were compared and led to the institutionalisation of one type of inferential statistics as *the* method of scientific inference guiding university curricula and journal editorial policy (Gigerenzer, 2000; Halpin & Stam, 2006). The swiftness of the statistical subjugation was best expressed by Kendall (1942) who dryly remarked that statisticians had "overrun every branch of science with a rapidity of conquest rivalled only by Attila, Mohammed, and the Colorado beetle" (p. 69).

Practical Pressures

Danziger (1990) saw the inference revolution as a result of pressure on American psychologists to legitimise their work by showing practical utility. At the time, the largest markets for psychological products were for educational and military application. Consequently, single-participant experiments were seen to be of little value. In contrast, the treatment-group experiment allowed users an efficient way to measure and compare groups of individuals in different treatment conditions (Gigerenzer, 2000; Gould, 1981). For such tasks, the null hypothesis testing of group means seemed tailor made.

The results were dramatic. For example, the percentage of empirical studies reporting only group data published in the *American Journal of Psychology* rose from 25% to 80% between 1915 and 1950 (Gigerenzer, 2000). Conversely, over the same time period, published individual-only articles plummeted from 70% to 17% (Gigerenzer, 2000). The figures for the acceptance of null hypothesis testing were even more dramatic, with Rucci and Tweney (1980) reporting only 17 articles using NHST procedures between 1934 and 1940. However, by 1955, more than 80% of the articles surveyed employed some form of significance testing. It should be noted that the changes were not isolated only to journal articles, with half of the psychology departments in the leading American universities offering courses on Fisherian methods, and making inferential statistics a graduate program requirement by the early 1950s (Gigerenzer, 1991).

Hybridisation and the Modern Method

It is an interesting, and recurring, theme that practical motivations drove the spread of null hypothesis testing and treatment group approaches from the field to the laboratory, and not the other way around (Gigerenzer, 2000). Faced with the two contrasting ideological approaches of Neyman-Pearson and Fisher, and coupled with the huge demand for practical application, academic authors started to graft the Neyman-Pearson concepts onto the skeleton of Fisherian logic (Gigerenzer, 2000; Halpin & Stam, 2006). Guildford's *Fundamental Statistics in Psychology and Education*, first published in 1942, is a germane example combining Fisher's experimental logic with the accept/reject dichotomy of Neyman and Pearson (Gigerenzer, 2000). The end result is that 50 years later a hybridised statistical methodology had become enshrined as *the* method for statistical inference in psychology (Kline, 2004).

Hubbard, (2004) has remarked that today null hypothesis significance testing follows Neyman and Pearson formally but Fisher philosophically. The modern hybrid logic, for example, employs a null hypothesis that is set up to be disproved, in accordance with Fisher, however, it is a hypothesis of strictly no difference, with rejection allowing acceptance of an alternative hypothesis in keeping with Neyman-Pearson thinking (Gigerenzer, 2004; Hubbard & Bayarri, 2003). Similarly, notions of power and error are borrowed from the Neyman-Pearson tradition but the Type I error probability is often referred to as the significance level, as in Fisherian logic (Hubbard, 2004). Finally, results obtained are interpreted epistemically and not behaviourally, a further nod to Fisher's paradigm (Gigerenzer, 2000).

Summary

As noted by Hubbard and Bayarri (2003), the conjoining of Neyman-Pearson hypothesis testing with Fisher's significance testing is a marriage of convenience that neither camp would have condoned. It is also one that many others have questioned, with Schmidt and Hunter (1997) and Schmidt (1996) calling for the modern NHST method to be banned from use in psychology. Rozeboom (1997) called it "the most bone-headedly misguided procedure ever institutionalised in the rote training of science students." (p. 335). Foreshadowing Rozeboom's comments, Paul Meehl (1978) believed it one of the worst things to ever happen in psychology and like several others, most notably Skinner, Meehl saw Fisher as the root cause: "Sir Ronald has befuddled us, mesmerised us, and led us down the primrose path." (p. 817).

Such condemnation is perhaps unwarranted. As illustrated above and highlighted by Gigerenzer (2000), Fisher saw significance testing as but one weak option amongst many for statistical inference and his original method bears little resemblance to the ritualised process used in modern psychology. There is, however, no shortage of other views on NHST with more than 400 references devoted to the topic (Kline, 2004). Indeed, the number of published journal articles criticising NHST practices has grown from less than 30 in the 1970s to more than 180 during the 1990s (Kline, 2004). It is to such criticism that attention is now turned.

Criticisms of Null Hypothesis Significance Testing

Assumptions

Modern Null Hypothesis Significance testing is predicated on several assumptions which, if violated, can lead to inaccurate calculations and erroneous inference. The most common of these assumptions mandate that samples are randomly selected, the threshold for significance is 5%, the dependent variable is a linear combination of the effects of independent manipulation, that observations and error are independent and come from normally distributed populations, and that these distributions are equal across conditions (Kline, 2004; Loftus, 1996). As noted by Kline (2004), these assumptions underlying NHST are much more restrictive than many researchers realise.

Sampling

The *p* values for test statistics used in NHST require random sampling from known populations and are crucial to making valid population inferences. However, most samples in social science are not randomly selected, being instead convenience samples from homogenised cohorts such as university students with generally little, or no, attempt made to specify the population of which the sample is supposedly indicative (Kline, 2004). The result is that p-values are understated, with statistical significance overstated because standard errors are too conservative (Reichardt & Gollob, 1999). The problem with such an approach is elucidated by Lisa Feldman Barrett: "The goal of psychology is to make nomothetic laws – laws that apply to all people… the question is how can you do that when you're sampling by convenience" (cited in Grice, 2011, p.92).

Normality

Assumptions of normality are also critical to NHST, with violations affecting both *p* values and power calculations regardless of whether group sizes are equal or not (Erceg-Hurn & Mirosevich, 2008; Kline, 2004). Micceri (1989) examined 440 large data sets from the psychological and educational literatures which encompassed a wide range of ability and aptitude measures (e.g., math and reading tests) and psychometric measures

(e.g., scales measuring personality, anxiety, anger, satisfaction, locus of control). None of the data were found to be normally distributed, and few distributions even remotely resembled the normal curve. Instead, the distributions were frequently multimodal, skewed, and heavy-tailed. Micceri's (1989) findings also cohere with other research that has identified similar asymmetrical and skewed distributions for commonly employed psychological variables such as reaction time (Erceg-Hurn & Mirosevich, 2008).

Whereas it was once thought that NHST was relatively insensitive to such violations, Wilcox (1998) has shown how even relatively small departures from normal distributions can lead to positive bias, incorrect calculations, and distorted inferences. Though there are versions of several tests, such as the t-Test and F-test that do not assume normality or homogeneity of variance, they are rarely used in practice with similar under utilisation evident for other modern statistical techniques such as bootstrapping and Winsorization (Erceg-Hurn & Mirosevich, 2008; Kline, 2004).

Homoscedasticity

Another important assumption underlying classic parametric tests is that of equal population variances which can be measured in terms of a variance ratio (VR). If two populations have similar variances, the VR will be close to 1:1. For example, if the variance of population A is 12 and the variance of population B is 10, the VR would be 12:10, or 1.2:1. However, when real data are analysed, the VR often strays markedly from the 1:1 ratio required to fulfil the assumption (Erceg-Hurn & Mirosevich, 2008).

Keselman et al. (1998) conducted a VR review of ANOVA analyses in 17 educational and child psychology journals. In studies using a one-way design, the mean VR was found to be 4:1 and in factorial studies, the mean VR was even higher at 7.84:1 (Keselman et al., 1998). Similar results have been evidenced in the *Journal of Consulting and Clinical Psychology*, the *Journal of Experimental Psychology: General* and the *Journal of Experimental Psychology: Human Perception and Performance* (Erceg-Hurn & Mirosevich, 2008).

Linear Effects

Assuming that the dependent variable is obtained by adding up effects from a combination of independent variable(s), interactions, and error is also problematic in NHST. Loftus (1996) contends that methods of data analysis often dictate the nature of

psychological theory. Assuming a general linear model, therefore, can bias research against more realistic or interesting hypotheses leading to a one-size-fits-all philosophy. Or as Loftus (1996) keenly states: "Off the shelf assumptions produce off the shelf conclusions." (p.164).

Dichotomous Thinking and Practical Significance

The outcome of a statistical test is dichotomous. Depending on the significance threshold, typically 5%, either the null hypothesis is rejected or the alternate hypothesis is rejected. Many authors have highlighted the problem with such an evaluation (Chow, 1996; Cohen, 1994; Rosnow & Rosenthal, 1989), noting that two similar experiments can have differing levels of significance and yet produce the same effect size[10]. Statistical significance, therefore, does not necessarily translate to practical significance. However, as Kline (2004) reports, experimenter confidence in research outcomes is strongly linked to levels of significance with sharp declines in confidence associated with values just above .05 compared with values just below .05.

Verification

One further problem with the assumptions underpinning NHST is that they are seldom verified or applied appropriately in practice. For instance, Keselman et al. (1998) reviewed more than 400 analyses published in psychology journals during 1994 and 1995 and found few studies validated the assumptions of the statistical tests employed. Similarly, Max and Onghena (1999) found corresponding levels of neglect in speech, language, and hearing research journals, whereas Glover and Dixon (2004) found only 35% of the tests employed in a range of psychology journals were utilised in a manner consistent with the logic of NHST. These examples lend credence to Kline's (2004) contention of a substantial gap between NHST as described in the literature, and its use in practice. The dire implication is that there are likely to be few cases in reality where NHST gives completely accurate results (Kline, 2004).

[10] Effect size is a general term that refers to a suite of statistical formula that calculates the strength of association between variables. Cohen's *d* and Pearson's *r* are but two indicators of effect size. (Cozby, 2004).

Meehl's Conjecture

It might be contended that the gaps identified by Kline (2004) represent problems with researcher application and not any deficiency intrinsic to the NHST method itself. There are, however, two other related assumptions that do highlight a serious inadequacy with the NHST method. Taken together they make up what has been come to be known as *Meehl's conjecture*.

Nulls of No Difference

By assuming a weak, zero difference, range null hypothesis, as opposed to a stronger, numerical point prediction, the psychological use of NHST is diametrically dissimilar to that in physics (Meehl, 1967). The problem with such an assumption is that it sets the hurdle too low for a test to surmount. In order for two groups, which differ on some independent property, to record identical dependent variable outcomes, either all average values of determinants of the output variables must be the same in both groups, or the pattern of differences of the average values must precisely counterbalance (Meehl, 1967). This is extremely unlikely, with almost any psychological variable differing between groups to some decimal place. The resulting inference is that psychologists are calculating hypothesis on premises they do not believe (Rorer, 1991).

Independence and Sample Size

As noted by Hubbard and Lindsay (2008), sample size has a huge effect on significance level by, amongst other things, increasing the power of a test. When coupled with Meehl's (1997) assertion that everything in psychology correlates to some extent with everything else, and with the unlikely assumption of a null of no difference, a *crud factor* emerges in psychological research (Meehl, 1997). The grave consequence of the crud factor is that the null hypothesis is almost always false.

Empirical Backing

Meehl's conjecture, which in the case of directional hypothesis predicts null rejection levels of 50%, has been demonstrated empirically. Using random, computer generated, directional, hypotheses and data from 81,000 individual MMPI-2 tests, Waller (2004) tested items for hypothetical gender differences. As noted by Gigerenzer (2004), of 511 items, 46% were found to be statistically significant with many item means 50-100 times

larger than their standard errors. These findings also cohere with Bakan's (1966) work in which more than 60,000 test results were compared on such arbitrary criteria as east versus west of the Mississippi river and all tests were found to be significant. However, it was Berkson (1938) who first highlighted the problem: "If, then, we know in advance the *p* that will result from an application of a chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all." (p. 526).

Logic

p(D|H0) ≠ p(H0|D)

It has been remarked by many authors that NHST procedures fail to tell researchers what they want to know (Cohen, 1994; Gigerenzer, 2004; Kline, 2004; Nickerson, 2000; Rorer, 1991). When testing H_0 one obtains the probability that the data (D) could have arisen if H_0 were true, $p(D|H_0)$. However, the inverse probability, $p(H_0|D)$, the probability of the hypothesis given the data, is what most researchers are seeking to quantify. This probability is not the same; a significance test does not provide a probability for a hypothesis.

To understand why, consider the form of the following syllogism which embodies the logic of NHST:

P1: If the null hypothesis is correct, the observed data would likely have not occurred.

P2: The observed data has, however, occurred.

C1: Therefore, the null hypothesis is probably not true.

Though this sounds logical it is actually formally invalid being equivalent to the syllogism: If a person is a New Zealander, he is probably not a member of parliament. This person is a member of parliament; therefore, he is probably not a New Zealander.

Lindley's Paradox

Another issue consonant with sample size was elucidated by Lindley in 1957. Lindley showed that for any level of significance, *p*, and for any non-zero prior probability of the null hypothesis, $Pr(H_0)$, a sample size could be found such that the posterior probability of the null, $Pr(H_0|D)$, is $1 - p$. That is, a null hypothesis rejected at the .05 level by a

Fisherian significance test could nevertheless have 95% support from a Bayesian viewpoint!

The rationale stems from the fact that no matter how small the *p* value, the likelihood ratio $Pr(D|H_0) / Pr(D|H_A)$ approaches infinity as the sample size increases. Consequently, for large *n*, a small *p* value can actually be interpreted as evidence in favour of H_0 rather than against it (Hubbard & Lindsey, 2008). This also no doubt explains Royall's (1986) example of well-known statisticians whose interpretations of *p* values in small versus large sample studies were totally contradictory. Though some argued that a given *p* value in a small sample study is stronger evidence against H_0 than the same *p* value in a large sample study, others argued precisely the opposite.

Another Bayesian Argument

A related implication of the above logic is an argument advanced by Bayesian statisticians. By convention, the null hypothesis is rejected when *p*(observed data | null hypothesis) <.05. However, doing so implies the conclusion that *p*(null hypothesis | observed data) is small. Indeed, using a Bayesian significance test for a normal mean, Berger and Sellke (1987) showed that for *p* values of .05, .01, and .001, the inverse probabilities of the null, $p(H_0|D)$, for *n* = 50 are .52, .22, and .034. For *n* = 100, the corresponding figures are .60, .27, and .045. Berger and Sellke (1987) went further, demonstrating that data yielding a *p* value of .05 resulted in a posterior probability of the null hypothesis of at least .30 for any objective prior distribution[11].

It might be argued by dyed-in-the-wool Neyman-Pearson frequentists that such an argument does not apply in long-run repeated sampling situations where α is a prescription for behaviours, rather than a means of assessing evidence like the *p* value in Fisherian and modern NHST logic. Sellke, Bayarri, and Berger (2001), however, devised a method for calibrating *p* values so that they can be interpreted as Neyman–Pearson frequentist error probabilities. They found that, *p* =.05 translates into a frequentist error probability of α (.05) = .289 in rejecting H_0, a result suggesting no evidence against H_0. Even α (.01) = .111. This, coupled with the Berger and Selke (1987) findings, undermines

[11] A test of a normal mean with symmetric priors with equal prior weight given to H_0 and H_A

significance testing as a method of generating reasonable measures of evidence (Hubbard & Lindsay, 2008).

Sequence Effects

A further logical impediment with the modern NHST method concerns the sequence by which results are obtained. Consider Fisher's (1925) famous tea tasting experiment in which a lady is asked to distinguish whether milk has been added to a cup of tea prior to, or after, the tea was added. With a null of no difference, that the participant cannot discriminate accurately, the expectation would be for 50% correct and 50% incorrect guesses. Presuming the participant actually guessed correctly on the first 5 of 6 trials (CCCCCI), a *p* value of .109 results, which is not significant at the 5% level (Hubbard & Lindsay, 2008). Now, imagine that a researcher decides to run the experiment until the participant makes a mistake, rather than having 6 specified trials. In this case, assuming the same sequence of outcomes as in the first experiment, the *p* value is .031, which *is* significant at the 5% threshold (Hubbard & Lindsay, 2008).

The difference arises due to the nature of how extreme results are calculated in each scenario. In the first fixed-trial experiment there are 7 different ways the participant could have correctly identified at least 5 of 6 cups (CCCCCI, CCCCIC, CCCICC, CCICCC, CICCCC, ICCCCC, and CCCCCC), thus giving a probability of $7(1/2)^6 = .109$.

By contrast, in the second sequential case a potentially infinite number of extreme cases could have occurred where the first 5 cups were correctly identified, i.e., 6 correct, 7 correct, 8 correct etc. The maths in this case gives a result of $(1/2)^6 + (1/2)^7 + (1/2)^8 + \ldots = (1/2)^6 / (1- 1/2) = (1/2)^5 = .031$. As noted by Hubbard and Lindsay (2008), this is nonsensical as the exact same data, obtained in the exact same sequence, should give the same *p* values. However, depending on the vagaries of researcher preference in experimental design, different results ensue.

Extreme Values

A similar, related problem is illustrated by Rorer (1991). In psychological statistics texts, it is common to draw NHST distributions as two competing point estimates, as in the first graph in Figure 1. However, with a null hypothesis of no difference, the alternative hypothesis becomes a directional one stating that every value greater than zero has the

same probability of occurring. Thus, as demonstrated in the second graph in Figure 1, it is possible to have a value in the rejection region for H_0, even though the probability of such an outcome is greater under H_0 than H_A.

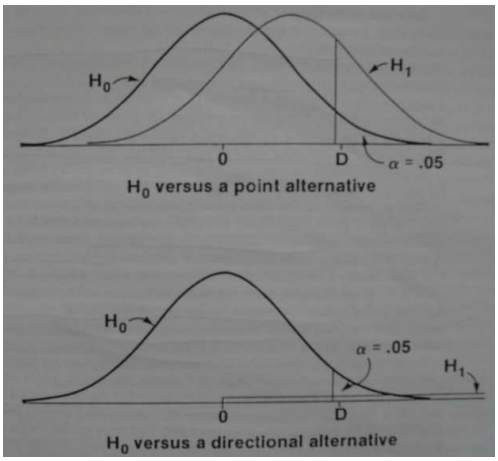

Figure 1: Null Hypothesis versus a point hypothesis compared to a null hypothesis versus a directional hypothesis. Reprinted with permission from Rorer, 1991, p.73.

The problem is most eruditely expressed by Jeffreys (1939): "What the use of p implies … is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure." (p. 316).

Sampling Uncertainty

Given the issues associated with the underlying assumptions of NHST, Meehl's conjecture, and the logical arguments presented above, it might be wondered whether the modern, hybridised NHST method has any use and/or support at all. A view characteristic of NHST supporters is advanced by Mulaik, Raju, and Harshman (1997) who assert that the sole task of NHST is to provide a measure of sampling uncertainty: "That's all a significance test provides, no more, no less." (p.73). However, there are strong reasons to question such a view in psychology.

Application

Neyman-Pearson hypothesis testing is a perfectly valid and well suited technique for making inferences and guiding decision making in domains such as quality control. In such situations, the acceptable deviation from null can be specified; both accept and

reject decisions are appropriate categories; the alternative courses of action can be clearly delineated; and any decision can be regarded as one of a series of such choices, such that one can minimize the overall loss (Bakan, 1966). However, as noted by many authors, such requirements often go unfulfilled in psychological research making accurate and valid inferences a rare and hazardous undertaking (Bakan, 1966; Cumming, 2008; Haig, 2011).

Power

A further challenge is issued by Schmidt and Hunter (1997) who, following Cohen (1962), cite the low average power of significance tests in psychology as problematic. Historically, the psychology literature has employed power in primary studies of between .4 and .6 (Cohen, 1962; Kline, 2004; Schmidt, 1996; Sedlmeier & Gigerenzer, 1992). With such numbers, even if a legitimate effect did exist, it would only be detected by researchers about half of the time. As Schmidt and Hunter (1997) remark, one might be better off flipping an unbiased coin to determine research results.

It has, however, been argued that such a conclusion is overstated (Mulaik, Raju, & Harshman, 1997). Power is after all a function of several factors[12] and thus, a study with a power level of .5 for detecting medium effects might have the power level of .8 for detecting large effects. Unfortunately, this argument is blunted when one considers that in many research domains in psychology the power is less than .5, and that typical effect sizes for psychological variables are of a small to medium magnitude (Cafri, Kromrey, & Brannick, 2010; Schmidt & Hunter; 1997).

Another argument contrary to the view of Schmidt and Hunter (1997) might be advanced along the lines that researchers should use sample sizes of sufficient magnitude to ensure high power. Once again, however, this is problematic. Cozby (2004) shows that with a medium effect size and power of .8, which still possesses a 20% chance of failing to detect an effect, the number of required participants is in excess of a 100, a number beyond that which is feasible to obtain in many cases (Schmidt & Hunter; 1997).

[12] Power is predicated on a combination of the type of statistical test, the sample size, significance level, and effect size, amongst other factors (Mulaik, Raju, & Harshman, 1997).

Replication and Confidence Intervals

In 1966, Bakan pointed out that the referent for all probability considerations used in NHST is neither in the population nor the subjective confidence of the investigator. Rather, it is a hypothetical distribution of experiments all conducted in the same manner, only one of which is actually observed. Thus, it is replication of the experiment that validates the inference model (Bakan, 1960), and as Steiger (1990) famously declared: "An ounce of replication is worth a ton of inferential statistics." (p.176).

This point is validated by Cumming (2008) and further belies the view that NHST is a valid estimator of sampling uncertainty. Taking repeated, n=32 paired samples, from a standardised hypothetical normal distribution using a power level of .52, Cumming (2008) conducted 25 separate mean t-Tests. Though the overall results were as expected, in terms of power, with 12 of 25 experiments producing a $p<.05$ result, the variation in p values was extreme ranging from $p<.001$ to $p=.706$. As Cumming (2008) stated, "Carrying out such an experiment—as much of psychology spends its time doing—is equivalent to closing your eyes and randomly choosing 1 of the 25 experiments… A *** result ($p < .001$) is first prize in this p value lottery, and you need almost the luck of a lottery winner to obtain it." (p. 288).

Many authors have remarked that confidence intervals provide a better assessment of sampling uncertainty than NHST. For example, a 95% confidence interval will, on average, capture 83.4% of future replication means (Cumming & Finch, 2001; Fidler et al., 2004; Estes, 1997; Loftus, 1996). However, it appears few researchers are aware of this (Cumming, Williams, & Fidler, 2004). This finding, coupled with the minimal utilisation of confidence intervals in psychological research (Fidler et al., 2004), no doubt explains the continuing belief in NHST as a valid method for assessing sampling uncertainty.

Summary

NHST is a woefully inadequate method by which to derive valid statistical inferences. Many of the assumptions, though theoretically plausible, are in reality too restrictive with the result that they are often not adhered to in practice. Moreover, other assumptions, such as the independence of psychological phenomena and null hypotheses of zero

difference, are so unrealistic as to be farcical. Meehl's conjecture illustrates the resulting problem: the null hypothesis is almost always, in reality, false.

Were that the only problem with NHST, it would represent a serious, no doubt some would say potentially fatal, concern associated with its use. However, as Lindley's paradox, sequence effects, extreme values and the several Bayesian statistical arguments all demonstrate, the logical foundation on which NHST is predicated is invalid. NHST can not tell researcher's what they want, and need, to know.

The argument that NHST is appropriately applied as a measure of assessing sampling uncertainty is similarly bankrupt. Its application in psychology is rarely viable and the lack of power, and moderate effect sizes, associated with most studies renders the argument even more suspect. Moreover, even if the erratic sampling results of Cumming (2008) could be dismissed, the truth is that confidence intervals are a much more informative tool for evaluating the degree of uncertainty in a sample than NHST. With the very rare exception of quality control situations, there is no legitimate reason for the continued use of NHST in psychology today.

Given the logical, conceptual and philosophical inadequacies of NHST, one might conclude that it would be a method long forsaken in psychological research. This would be a false conclusion however. As the next section demonstrates, NHST remains the most misunderstood, misused, and misapplied research methodology in psychology. Depressingly, it is also remains the most common.

Consequences of Null Hypothesis Significance Testing

Growth

A Growing Addiction

Growth in the use of NHST in published journal articles boomed during the 1950s as part of the inference revolution (Gigerenzer, 2000). Since then use of NHST has continued to climb. Hubbard, Parsa, and Luthy (1997) tracked the uptake of NHST in the *Journal of Applied Psychology* from 1917 to 1994 and documented an increasing reliance on

statistical significance tests. Referencing more than a 1,000 articles, NHST was found to appear in only 17% of articles during the 1920s but over 90% of articles in the 1990s.

The story is the same for other psychology journals. Picking an article at random from each year of publication for 12 APA journals, Hubbard and Ryan (2000) recorded the use of NHST from 1911 to 1998. On average NHST was used in 82.4% of the 8,001 articles sampled. However, the trend across decades was strongly non-linear with use exceeding 90% since the 1970s.

Saturation

Hubbard (2008) provides more recent figures. Sampling 1,750 papers in the same 12 psychology journals as Hubbard and Ryan (2000), Hubbard (2008) found the use of NHST averaged 94% between1990 and 2002. Indeed, *The Journal of Developmental Psychology* and *The Journal of Abnormal Psychology* both averaged more than 99%. Hagen (1997) highlights the consequence of such reliance: "NHST is … deeply embedded in our methods of statistical inference. It is unlikely that we will ever be able to divorce ourselves from that logic even if someday we decide that we want to." (p. 22).

Editorial Policy and Statistical Reform

Editorial Encouragement

The recognition of the deeply entrenched nature of NHST in psychology, and the shortcomings of NHST as a research methodology, has lead to calls for other tools of statistical inference to supplement the use of NHST. Most notably, the Task Force on Statistical Inference (TFSI) was convened by the Board of Scientific Affairs of the American Psychological Association in 1996 to evaluate and report on such methods. Along with suggestions on the use of power, experimental design, and graphical presentation, the TFSI recommended the use of confidence intervals and effect sizes in its final report: "It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p value or, better still, a confidence interval…Always provide some effect size estimate when reporting a p value." (Wilkinson and the Task Force on Statistical Inference, 1999, p. 599).

These sentiments echoed earlier recommendations in the fourth edition of the APA's Publication Manual where the reporting of effect sizes was encouraged (Kline, 2004;

Nickerson, 2000). These recommendation gained little traction, however, with effect sizes appearing in psychology journals between 1994 and 2000 in only 25% of articles, and even then few were actually interpreted (Kline, 2004). Similarly, between 1994 and 1997, Geoffrey Loftus, then editor of *Memory and Cognition,* asked contributors to report their results with error bars rather than relying solely on statistical significance tests. Once again, in the rare cases where Loftus' recommendation was heeded, authors invariably failed to use error bars for interpretative purposes (Cumming & Finch, 2001).

A Dismal Failure

More recently, Fidler et al. (2005) investigated the effects of the TFSI's, and Phillip Kendall's (1997), editorial recommendations by reviewing 239 articles in the *Journal of Consulting and Clinical Psychology*. They found little change in practices with effect sizes remaining little reported while confidence interval reporting peaked at a scant 17% in 2001. The same story is evident in education, public health, and counselling journals (Byrd, 2007; Fidler et al., 2004; Sink & Stroh, 2006). Finch, Cumming, and Thomason, (2001) best summarise these results. In noting that many important aspects of statistical inference in psychology remain the same today as it was in the 1940s, they concluded: "the cogent, sustained efforts of the reformers have been a dismal failure." (p. 205).

Confusions and Fallacies

A Statistical Aberration

Given the extreme, some might say pathological, reliance on NHST for statistical inference in psychology, one might think psychologists would possess an intimate and profound understanding of the methodology. Indeed, one might reckon it a rare thing for a psychologist to be caught unawares or answer a question regarding NHST incorrectly. Reality, however, shows this to be a far from accurate expectation.

Although Tversky and Kahneman (1971) demonstrated a cognitive bias in many statistical experts with the belief in the law of small numbers[13], it was Oakes (1986) who first unearthed evidence of statistical confusion in a pure psychology cohort. Asking 70 academic psychologists for their interpretation of $p < .01$ via a choice of 7 statements,

[13] This belief is that small samples are typical representations of populations, and that statistically significant results are likely to obtain with replication samples half the size of the original.

Oakes (1986) found only 8 (11%) gave the correct interpretation with almost 50% endorsing statements that *p* values indicated the conditional probability of either H_0 or H_A. Lest it be thought Oakes' (1986) results were a statistical aberration, similar findings have been established by many other authors (Gigerenzer, Krauss, & Vitouch 2004; Haller & Krauss, 2002; Hubbard & Bayarri, 2003).

Haller and Krauss (2002), for instance, provided 113 German psychology faculty with a statistical print out and asked them to answer questions concerning interpretation of the results, for example, whether the result absolutely disproved the null, provided evidence for the probability of making an incorrect decision, or gave reliability information concerning the probability of generating significant results in the long run. Depressingly, the best result from all groups was that of lecturers teaching statistical methods classes, 80% of whom were found to agree with at least one erroneous statistical conception (Haller & Krauss, 2002).

Common Confusions

The most common fallacies and statistical misconceptions endorsed or believed by psychologists from the studies listed above included:

- *The Magnitude Fallacy*: A *p* value is a numerical index of the size of an effect; thus, low *p* values indicate large effects. This is erroneous because it confuses the definitions of *p* and effect size.
- *The Replication Fallacy:* A *p* value of .05 means that same results will replicate on 95% of future occasions. As illustrated by Cumming (2008), just because a significant result is obtained it does not necessarily hold that it will be replicated consistently.
- *Valid Research Fantasy:* 1-*p* is the probability that the alternative hypothesis is true, i.e., a $p<.05$ gives a 95% probability of the alternate hypothesis being true. This is a gross misunderstanding; $p(D/H_0) \neq p(H_0/D)$, nor for that matter $p(H_A/D)$.
- *Odds Against Chance Fallacy:* The *p* value is the probability that the research results are due to chance. This is incorrect because *p* values are calculated on the premise that sampling error is the only thing that causes the sample statistic

to deviate from the null hypothesis. That is, the likelihood of sampling error is already taken to be 1.00 when a statistical test is conducted because H_0 is assumed to be true.

- *Inverse Probability Fallacy/ Bayesian Wishful Thinking:* A *p* value is the probability that the null hypothesis is true, thus, $p<.05$ implies $p(H_0/D) < .05$. As with the valid research fantasy, *p* values are conditional on the data, not on the null hypothesis; ($p(D/H_0) \neq p(H_0/D)$).
- *Statistical Significance Equals Scientific Significance:* This fallacy ignores the fact that statistical significance says nothing about the size/importance of an effect. As Meehl's conjecture makes clear, with a large enough sample size almost any arbitrary comparison is significant.
- *P and α Confusion: P* values and alpha levels (α) are the same thing. As noted by Hubbard and Bayarri (2003) the *p* comes from the Fisherian school of thought and represents a probability concerned with measures of evidence, whereas α concerns the long run relative frequency of error in the Neyman-Pearson paradigm.

Meaningless Applications

Such confusion and misconception has engendered a retardation of the cumulative development of psychological knowledge (Schmidt & Hunter, 1997; Meehl, 1978). By assuming significant statistical results are automatically veracious, meaningful, and important, effort has been diverted away from critical appraisal of data and the tasks of replication, two fundamental principles of scientific analysis (Kline, 2004). Additionally, the lack of replication, coupled with the obsessive focus on statistical thresholds has created a file drawer problem, whereby, studies that fail to reject the null are filed away and never submitted for publication (Meehl, 1997, 1967; Rosenthal, 1979). As Hubbard (2008) duly summarises: "The end result is that applications of classical statistical testing in psychology are largely meaningless." (p.297).

Summary

It may be seen then, that Hagen (1997) was more right than even he realised. Not only have researchers demonstrated an almost obsessive aversion to desisting with the practice of NHST, there has been a substantial rejection of supplementing the practice with other methods such as effect sizes or confidence intervals. As the research reported above attests, NHST remains at, or close to, saturation point in many journals, with some comprised entirely of studies using NHST methods. Concordantly, and despite the best efforts of many editors and boards of standards, the use of effect sizes and confidence intervals remains of trivial magnitude.

The utter lack of recourse to other methods of statistical inference by itself constitutes grounds for serious consternation, given the follies delineated earlier. However, this issue is greatly exacerbated by the atrocious levels of understanding of NHST evinced by many psychologists. The resulting faltering application has not only severely retarded the cumulative development of knowledge in psychology, it has also harmed "the usefulness of psychological research as a means for solving practical problems in society" (Schmidt & Hunter, 1997, p. 449).

Conclusion

Hybridisation

William Sealy Gosset was a major force in the creation of the various tools of statistical inference that are endorsed and used in modern psychology. Despite this, in usurping Gosset's statistical tables and calculations, and publishing two of the most influential statistics texts of the 20th century, it is to Sir Ronald Fisher that most recognition for significance testing procedures is now accorded. The subsequent criticism and extension by Neyman and Pearson, and the practical pressures of the inference revolution, acted as the genesis for the hybridised methodology used by many researchers today. It is, however, a bastardised sundry of statistical theory that has resulted in misconception, misunderstanding, and misapplication within psychology.

The Sterile Rake

Whereas many assumptions of NHST are restrictive, and are not verified by researchers, others are a travesty of realist philosophy. When combined with the invalid logical foundations of NHST, the results include the assaults of Meehl's conjecture, Lindley's paradox, and Rorer's extreme values argument. The much touted defence of NHST as a means of assessing sampling uncertainty is similarly laid waste by Cumming's (2008) replication research and Schmidt and Hunter's (1997) power arguments. Meehl's (1967) description of researchers employing NHST as sterile intellectual rakes who leave behind a long train of ravished maidens but no viable scientific offspring remains germane.

The Deleterious Effects

The use of NHST procedures in modern psychology has reached saturation point. Further, attempts by reformers to encourage the utilisation of other tools of statistical inference to supplement NHST have been abysmal failures. This situation has only been aggravated by the miserable knowledge evidenced by research concerning psychologist's understanding of NHST. Tryon's (1998) comments cogently illustrate the consequences of such findings: ". . . the fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in miniscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are undoubtedly substantial." (p. 796).

Conclusion

It should be of little comfort to psychologists that similar problems with the application of NHST are evident in the research literatures of economics, ecology, and epidemiology (Ziliak & McCloskey; 2008). Neither should the supplemental tools of effect sizes and confidence intervals be regarded as a perfect salvation for they are not also without issue in application (Chow, 1996). Or as Kline (2004) has remarked, placing candles on a cow pie does not make a birthday cake.

It might be thought that reverting to using either the Neyman-Pearson method or Fisher's original testing procedures would improve things. However, as Haig (2011) notes, neither is adequate for the role psychologists would have them perform and this no doubt explains, in part, the attraction in trying to combine them. The Neyman-Pearson approach fails to yield inferences with respect to any specific hypothesis and is limited to applications consonant with quality control situations which are rare in psychological research. Similarly, without an alternative hypothesis, and predicated on the null being true, Fishers significance testing cannot, therefore, also be a direct yardstick for adjudicating whether the null is false (Haig, 2011).

There are many alternatives and supplemental tools to NHST procedures for the analysis of data and derivation of statistical inference. Bayesianism, meta analysis, and the neo-Fisherian paradigm are but several of the examples. Though no clear alternative to NHST has emerged (Ferguson, 2009), and each contender has its own weaknesses and issues, this should not detract researchers from examining, evaluating, and employing them as means allow and circumstances dictate. Proper psychological research involves critical thinking, or to paraphrase Tukey (1969), it is detective work, not the ritualised application of any one technique.

Chapter 4: Interindividual Versus Intraindividual Research and the Granularity of Research Methods

An Idiographic-Nomothetic Debate?

Origins

Allport

It is commonly believed that the terms *idiographic* and *nomothetic* were introduced into the psychology lexicon by Gordon Allport in 1937 with the publication of his *Personality: A Psychological Interpretation.* Runyan (1983), Lamiell (1998), and Krauss (2008) are several authors crediting Allport with borrowing the terms from the philosopher Wilhelm Windelband (1894/1998): "The proposal to distinguish sharply between the study of general principles and the study of the individual case has taken many…forms. The philosopher Windelband, for example, proposed to separate the *nomothetic* from the *idiographic* disciplines." (Allport, 1937, p. 22; italics in original). However, the terms were actually used in psychology for the first time nearly 40 years earlier.

Munsterberg

It was in 1899, during a presidential address to the American Psychology Association, that Hugo Munsterberg first used the terms *idiographic* and *nomothetic* in a psychology context. In the address, published in *Psychological Review* a month later, Munsterberg (1899/ 1994) highlighted the difference between sciences which seek isolated facts and those which seek laws: "...and thus we have two groups of sciences which have nothing to do with each other, sciences which describe the isolated facts and sciences which seek their laws. A leading logician baptizes the first, therefore, idiographic sciences, the latter, nomothetic sciences" (p.231-232). The leading logician in the above is likely to have been Windelband who influenced Munsterberg in the 1890s while both were members of the Southwestern School of neo-Kantianism (Hurlburt & Knapp, 2006).

Stern

It is uncertain whether Allport was introduced to Windelband's thinking directly or indirectly. Though Munsterberg was Allport's first teacher at Harvard and Allport's brother, Floyd, was Munsterberg's research assistant, Allport was heavily influenced by another Munsterberg disciple, William Stern: "…from Stern in particular I learned that a chasm exists between the common variety of differential psychology . . . and a truly personalistic psychology that focuses upon the organization, not the mere profiling of an individual's traits" (Allport, 1967, p. 10). Indeed, many of Allport's general ideas show Stern's influence (Ghougassian, 1972). Although Allport may be credited with popularising the terms within psychology, especially personality psychology, the idiographic/nomothetic distinction in psychology can be seen to exist prior to Allport, thanks to Munsterberg (Hurlburt & Knapp, 2006).

Definitional Confusion

A Disorganised Discussion

Issues concerning definitions within science are very important as ill-specified meanings can lead to conceptual confusion and misallocation of research resources. Borsboom, Mellenbergh, and Van Heerden (2004), for example, have suggested changes in the term *validity* have hindered progress on validity research by focusing attention on differing, exterior facets of validity at the expense of the core concept. Similarly, it can be seen that issues concerning nomothetic and idiographic research in psychology have been plagued by definitional confusion. At various times, for example, the idiographic-nomothetic research debate has concerned the comprehensiveness and usefulness of personality traits (e.g., Allport, 1937; Block, 1995; Cervone, 2005), individual uniqueness (e.g., Allport, 1937, 1962; Higgins, 1990), quantitative versus qualitative research (e.g., Allport, 1962; Meehl, 1954), the status of psychology as a science (e.g., Eysenck, 1954; Holt, 1962; Krauss, 2008; Nunnally, 1967), and the study of individuals versus the study of groups (e.g., Allport, 1962; Bem & Allen, 1974; Lamiell, 1987, 2003; Runyan, 1983). As such, the discourse is better characterised as a disorganised discussion of slightly related terms than as a true debate (Krauss, 2008).

Windelband and the Emerging Discipline

In order to understand how the idiographic-nomothetic debate has evolved, it is necessary to contrast the original views of Windelband with later incarnations. Windelband first used the terms *idiographic* and *nomothetic* in 1884 in a speech marking his assumption of rectorship at the University of Strasbourg on its 273rd anniversary. For Windelband, no strict division between humanities and natural sciences could readily accommodate the emerging science of psychology: "…an empirical discipline of such significance as psychology is not to be accommodated by the categories of the natural sciences and the humanities: to judge by its subject, it can only be characterised as a humanity, and in a certain sense as the foundations of all of the others; but its entire procedure, its methodological arsenal, is from beginning to end that of the natural sciences." (Windelband, 1884/1998, p.11). It is within this context that the terms *idiographic* and *nomothetic* were later applied: "So we may say that the empirical sciences seek in the knowledge of reality either the general in the form of the natural law or the particular in the historically determined form (Gestalt)...The one comprise sciences of law, the other science of events; the former teach us what always is, the latter what once was. If one may resort to neologisms, it can be said that scientific thought is in the one case nomothetic, in the other idiographic." (Windelband, 1894/1998, p.13).

Lost in Translation

Of particular relevance in the quotations above are the connotations associated with the terms *natural sciences* and *humanities*. In German they are termed *Naturwissenschaften* and *Geisteswissenchaften* respectively, and Windelband saw them as two complementary components of empirical science*, die Erfahrungswissenschaften* (Lamiell, 1998). Thus, just as there could be scientific thought associated with *what always is*, so Windelband believed there could be scientific thought with *what once was.* However, in English a connotation of scientific or generalised knowledge for the humanities is lacking (Salvatore & Valsiner, 2010).

For Windelband, nomothetic knowledge represents what is true for each and every human or collective, e.g., the law of gravity. Idiographic knowledge is then concerned with unique events, entities, and trends, but with the level of analysis crucially left open. Such an investigation might focus on a single human being but it might also focus on an

entire race, a cultural custom, or a specific dialect within a language. In sum, Windelband's conception of nomothetic and idiographic hung on differences between constraints of time and context, not of the unit of analysis (Kraus, 2008; Lamiell, 1998; Salvatore & Valsiner, 2010).

The Real Issue

Inconsistent Application and Boundary Conditions

Though Allport, and for that matter Munsterberg, shared Windelband's view that both nomothetic and idiographic research were complementary and that psychology should employ both approaches, it is unfortunate how Allport choose to use the terms. Being more interested in examining the behaviour of single individuals, Allport typically used the term *idiographic* to refer to the study of individuals, and the term *nomothetic* to refer to the study of populations and groups (Krauss, 2008). Moreover, as with other authors, Allport's use was not always consistent and the terms *nomothetic* and *idiographic* were used interchangeably with *understanding* and *explanation*, *morphogenic* and *dimensional*, and the *particular* and the *general*, amongst others (Allport, 1962; Krauss, 2008; Runyan, 1983).

The consequence of such inconsistency has seen researchers arguing to and fro about where the boundary between idiographic and nomothetic research lies. Some have contended that idiographic research is defined by that which holds for a subset of the population and, therefore, that interaction and moderating effects are idiographic in nature. Others, however, have pointed out that interaction terms added to regression equations, standard staples of nomothetic research, can also accommodate such classifications (Krauss, 2008). Similarly, studies that compare the level of variables in a person have been framed as both idiographic when focused on number of participants but nomothetic when viewed by number of observations (Epstein, 1979).

A Spurious Distinction

Johann Wolfgang von Goethe wrote that: "the particular is ever subordinate to the general; the general must ever accommodate the particular." (cited in Lamiell, 1998, p.30). A similar sentiment can be seen in Windelband's original use. All science is idiographic in the sense of being *what once was*, that is every event is time and context

sensitive, but it allows for nomothetic generalisations when further instances are also seen as *what once was as well* (Salvatore & Valsiner, 2010). The idiographic-nomothetic debate in psychology can be rightly seen as a misnomer then, the terms *nomothetic* and *idiographic* having been misconstrued as adversarial when in fact they represent two sides of the same empirical coin.

Granularity and the Fragmentation of Research Methods

The issue of real import for psychology associated with idiographic-nomothetic definitions is that of granularity of research methods, i.e., at which level, and by which methods, research should be conducted to generate the best psychological knowledge. Contemporary psychology has progressively identified itself as a nomothetic science but where nomotheticity has been taken to mean ergodicity of psychological phenomena (Salvatore & Valsiner, 2010). That is, an individual's variability (intraindividual variation) has been assumed to be identical to the variation between persons within a given population (interindividual variation) (Grice, Jackson, & McDaniel, 2006; Molenaar, 2004).

The result has been a fragmentation of psychological science (Salvatore & Valsiner, 2010). In stark contrast with Windelband, and for that matter Allport's earlier harmonious conception, two diametrically opposed methodological research views have ensued. The first is concerned with the averages and aggregates of people and groups of people, interindividual variation (IEV), whereas a contrasting approach focuses on the study of events and dispositions within a single individual history, intraindividual variation (IAV) (Krauss, 2008; Molenaar, 2004; Salvatore & Valsiner, 2010). In the former case, regression analysis, correlations, t-Tests, and ANOVAs are the most common analytical tools; in the latter, visual analysis of single-case designs predominates (Saville, 2008).

The Dominance of Interindividual Research

The Early Balance

Today the dominant tradition in psychology takes a large-N, group, or IEV approach to conducting research (Blampied, 1999; Krauss, 2008; Molenaar, 2004; Saville, 2004). However, this was not always the case. Danziger (1990), has shown that during the period 1914-1916 the ratio of published intraindividual to interindividual research was

more than 2 to 1 in several leading psychology journals. Indeed, the *American Journal of Psychology* held an almost 3 to 1 ratio with IAV methods utilised in 70% of all studies whereas IEV methods were employed in only 25% of total publications (Danziger, 1990).

Decline of IAV Research

These numbers shifted dramatically, however, and by the 1950s IEV research comprised more than 80% of articles in the same journals (Danziger, 1990). The serious decline in published IAV research is best exemplified by the 31 to 1 ratio of IEV to IAV research exhibited in the *Journal of Educational Psychology* (Danziger, 1990). The situation remains similar today with Molenaar (2004) lamenting that only one research program within psychometrics is dedicated to purely IAV research, namely, Nesselroade's work on P-technique which employs factor analyses of a single-participant's multivariate time series to investigate individual personality (Molenaar, 2004).

Contemporary Figures

To illustrate the complete dominance of IEV methodology within psychology, several searches were performed using PSYClit, PSYCinfo and PSYCarticles databases. Search parameters were confined to empirical articles between 2000 and 2010 with a personality subject and returned more than 41,000 articles. Adding the keywords *idiographic*, *single-case*, and *intraindividual* to the same search parameters, however, resulted in only 71, 48, and 67 articles returned respectively. This would seem to add much credence to Lamiell's (1998) contention that modern trait psychology has become a demographical exercise exploiting a psychology vocabulary.

No Behavioural Bastion

The figures for trait and personality psychology are not a statistical aberration with journals traditionally dedicated to IAV research, notably journals emerging from the behaviour therapy revolution of the 1960's with its initial emphasis on the scientific study of individual cases (e.g., Shapiro, 1966) also showing a marked change in methodological focus. Forsyth et al. (1999) tracked the number of single subject and group designs in the journal *Behavior Therapy* since its inception in 1970. Spanning 27 volumes and 1690 articles, the review showed single-subject designs peaked in the early 1970s at more than 40% but after the late 1980s they never made up more than 20% of total published work.

By contrast, group designs had consistently accounted for more than 40% of all articles since the late 1970s.

Similar results were obtained in a second expanded study comprising the journals *Behaviour Research and Therapy*, *Behavior Therapy and Experimental Psychiatry*, and *Behavior Modification* for the time period 1974 through to 1997. Once again, all three journals showed the same decline in single-case designs. For example, since the mid 1980s single-case designs have never comprised more than 5% of all articles in the journal *Behavior Research and Therapy*. As a result, Forsyth et al. (1999) concluded that behaviour therapy had drifted from its ideological roots of single-case/single-subject methods toward the disproportionate use of group methods.

Summary

Though popularised by Allport, the terms *idiographic* and *nomothetic* were introduced into psychology by Munsterberg in 1894. Inaccurate translation and inconsistent use has meant these terms have ignited spurious debates within psychology. Issues ranging from individual and group differences, to differences in qualitative and quantitative research, to the status of psychology as a science, and the usefulness of personality traits have all been enveloped by the idiographic-nomothetic discourse. However, when viewed through the lens of Windelband's original conception such debate is seen as misguided, with nomothetic laws and idiographic peculiarities comprising two sides of the same scientific coinage.

Though somewhat obscured by the terminological verbiage, the real issue to emerge from the idiographic-nomothetic debate concerns the granularity of psychology's research methods. A methodological split in the psychological research tradition has occurred along interindividual (IEV) and intraindividual (IAV) lines. Whereas IAV methods dominated early psychological research, it is an IEV research paradigm that typifies research practice today, with psychological science being increasingly defined by the limitations of statistical procedure (Borsboom, 2005). In order to understand the evolution of psychology's research tradition further, a look back at early developments is now undertaken.

A Brief History of Psychological Research Methodology

Early Flux

The rigid tradition that has come to dominate psychological research since the inference revolution mentioned earlier is a direct result of developments occurring during the early half of the 20th Century (Danziger, 1990). Until the 1940s, psychology was characterised by an eclectic mixture of research approaches and methods. This was a direct consequence of the various techniques employed in psychology's inception that were borrowed and adapted from biology, physiology, neurophysiology, and experimental medicine amongst other fields (Blampied, 1999; Danziger, 1990). Broadly construed, however, there were two schools of research that dominated these early periods, the Wundtian and the Galtonian.

The Wundtian School

Borrowing methods from experimental physiology and grounded in an introspective rationale, Wundt sought to manipulate the conditions of internal perception via psychological experimentation, such that objective report and observation could occur (Danziger, 1990). Experimentation, however, was only one part of the Wundtian paradigm, with Volkerpsychologie[14] also accorded an objective, and no less important, role supplementing "physics in the investigation of the total content of experience" (Wundt, 1897/1990, cited in Danziger, 1990, p.38). As such, communication and the subjective elements of individual perception were important components of Wundtian research with the experimenter and participant frequently changing places to deepen understanding of perceptual processes.

The Galtonian Paradigm

The alternation of experimenter and participant roles was but one element of the Wundtian paradigm rejected by psychologists following a Galtonian rationale. Thorndike

[14] *Volkerpsychologie* is a German term with no clear English equivalent but which most closely refers to a social psychology based on historical, ethnographic and comparative analysis of human cultural products, particularly language (Danziger, 1990).

(1911), for example, saw role swapping as unnecessarily subjective, viewing such procedures as primarily for the benefit of the participant at the expense of experimenter. Whereas Wundt worried about the segregation of mental processes from social and cultural factors, proponents of the Galtonian school were interested solely in abilities divorced from everything else (Danziger, 1990). As noted earlier in the measurement chapter, comparison and relative standing amongst individuals, stemming from Galton's fascination with individual differences, was the prime methodological rational perpetuated and extended by Cattell, Thorndike, and other proponents of the mental measurement movement.

The Rise of Interindividual Research

Practical Usurpation

By the 1920s the Wundtian school had began to lose out to the Galtonian one and several factors were responsible. Firstly, educational administrators began to utilise and apply psychological principles to the measurement, comparison, and evaluation of students, teaching methods, and educational conditions (Danziger, 1990). Secondly, WW1 and the US army alpha and beta tests created a huge demand for psychological testing (Gould, 1981). Combined, these two factors represented a strong spur for practical applications of psychology research which began to infiltrate, and impact on, the psychological research domain. Although laboratory methods were not abandoned in favour of mental testing, increasingly, laboratory work began to focus on groups of participants rather than individuals and basic research began more and more to resemble applied research (Danziger, 1990).

Social and Statistical Motivations

This trend towards group research was given further impetus with the emergence of private research foundations, such as the Laura Spelman Rockefeller Memorial, which sought to use research to achieve socially desirable ends (Danziger, 1990). Fisher's publication of *Statistical Methods for Research Workers* (1925), which emphasised the estimation of population parameters from small samples based on Fisher's experience with the sampling of agricultural commodities (discussed in chapter 3), and the introduction of Stevens' four scales (discussed in chapter 2) were two developments in statistical

psychology that further perpetuated the growth of group based research. When these developments were coupled with demand from the U.S. Government for research during the Great Depression, a synthesis in psychological research began to form (Danziger, 1990; Saville, 2008). Combining the practice of random subject assignment of experimental research with correlational and sampling statistics driven by the mental testing movement, the study of individuals was pushed off the research agenda by a focus on attributes of groups and populations (Lamiell, 1998). By the 1940s, this focus had become the dominant tradition within psychology discipline, a situation Danziger (1990) called "The triumph of the aggregate" (p.68).

Dissenters and Decline

There were several notable dissenters who were critical of what psychological science had become. The earliest was Boring (1919), who opined that statistical fascination divorced from intimacy with fundamental observations led nowhere. Another prominent critic was Skinner, who in his 1938 treatise, *The Behavior of Organisms,* first outlined his radical behaviorist approach to the scientific study of behaviour - a position famously expressed in his later work: "...instead of studying a thousand rats for one hour each or a hundred rats for ten hours each the investigator is more likely to study one rat for a thousand hours'' (Skinner, 1966, p. 21).

Indeed, it was the success of the Skinnerian philosophy that led to a modest re-emergence of single subject designs in the 1950s and 1960s (Saville, 2008). Another contributing factor was the demand in applied research for studies that used fewer participants (Saville, 2008). Despite this re-emergence, and as illustrated above, even research with a decidedly intraindividual and/or behavioural orientation had given way to utilising interindividual methodologies by the end of the 21st Century.

Summary

Early psychological research methodology was characterised by an amalgamation of research practices and standards imported from other disciplines such as medicine and physiology. The Wundtian and Galtonian methodological schools emerged with the former focused on individual experimentation and socio-cultural elements and the latter concerned with individual differences amongst the attributes of people. Demand for

applied psychological research from educational and military sources drove changes in laboratory research away from the study of individuals toward groups of participants.

This shift was further advanced by private research foundations, statistical developments, and concern for social research emanating from the Great Depression. Despite criticism from several prominent psychologists and the rise of behaviourism during the 1950s and 1960s, interindividual research has remained the dominant research paradigm since the middle of the century. As such, contemporary psychology is characterised by a fascination with what is common to an aggregate rather than what is universal to all.

Other than historical, social, and cultural factors, one might wonder what other reasons may have contributed to the overwhelming popularity of IEV research displayed in the psychology literature today. Surely, if IAV methods were useful to researchers would they not be employed? Similarly, were there drawbacks to an IEV approach, would not less IEV research be evident in the literature? The next sections answer both these questions by evaluating the arguments advanced for and against IEV and IAV research.

An Evaluation of Inter and Intra Individual Research Methodology

Intraindividual Research

Generalisability

A common view of psychology is that its primary goal is the "the development of generalizations of ever increasing scope, so that greater and greater varieties of phenomena may be explained by them, larger and larger numbers of questions answered by them, and broader and broader reaching predictions and decisions based upon them" (Levy, 1970, p. 5). According to this view, generalisations, once obtained, can be applied at the level of the individual by a process of deduction. Without such generalisations, however, it is contended that no proper scientific investigation is possible. Psychology without generalisation to universal laws is not science, but only a hope of science (James, 1961), and psychology without generalisation to individuals cannot be of any material use (Blampied, 1999).

The most withering criticism of IAV research concerns its proposed lack of generalisability. IAV research has been labelled an antiscience view that discourages the search for general laws (Nunnally, 1978). Holt (1962) called it an approach that is impossible for gaining useful information within psychology and Nunnally (1968) opined it was a view that allowed for only chaos to prevail in the description of human nature. Bem and Allen (1974) best illustrate the prevailing consensus remarking that IAV research is commonly characterised as "a scientific dead end, a capitulation to the man-in-the street view that a science of psychology is impossible because everybody is different from everybody else." (p.511).

However, the stirring rhetoric is ill informed. As argued by Windelband (1898/1994), there is nothing inherently unscientific about studying individuals with valid statistical and non-statistical enquiry being as possible with singular subjects as with populations (Blampied, 1999; Molenaar, 2004; Runyan, 1983; Salvatore & Valsiner, 2010). The research by Ebbinghaus on memory, by Pavlov on respondent conditioning, by Thorndike on instrumental conditioning, by Yerkes on comparative cognition, and Broca's work on neuropsychology all stand as incontrovertible testament to the utility of single-subject studies (Blampied, 1999). Indeed, it is the study of individuals and a focus on direct and systematic replication that best exemplify a cumulative research tradition that affords robust, empirical generalisations (Lamiell, 1998; Runyan, 1983; Salvatore & Valsiner, 2010; Saville, 2008; Sidman, 1960).

Efficiency

Another popular related criticism of IAV research suggests that it is not only impractical but logically impossible to conduct properly. Murray (1938) is an early example, asserting that "If individuals are as dissimilar as Allport suggests, then every sparrow would have to be separately identified, named and intuitively understood." (p. 715). Similarly, Levy (1970) argued that if individuals were truly unique then psychologists would have to formulate as many theories as people in the universe. Concordantly, Tuerlinckx (2004) highlighted economic issues with IAV research suggesting it was easier to engage participants for one-time studies rather than those which necessitated follow up meetings and additional longitudinal considerations.

It is of course logistically impossible to study each and every individual and/or their behaviour. However, this criticism applies equally to IEV research. As Saville (2008) notes, sourcing and administration of the large number of participants required in IEV research can be problematic. Molenaar (2004) also illustrates how technology such as on-line electronic screening measures, automatic scoring procedures for audio-visual registrations, and computational protocol analysis has made IAV research increasingly efficient. Similarly, the argument advanced by Levy (1970) is readily seen as misconstrued. Every human will of course differ in some way from another; however, he/she will also be representative of all humans in other respects. As demonstrated in the following sections, the study of the particular can lead to generalised knowledge (Allport, 1937, 1962; Epstein, 2010; Lamiell, 1981; Molenaar, 2004).

Causal Attributions

A further major criticism levelled at IAV research concerns the nature of causal attributions. In part this stems from the criticism above, that little faith in causal attributions can be afforded given unique and substantial human variability (Levy, 1970; Murray, 1938). The nature of single-subject designs, particularly the repeated measurement elements, is the other major component of this criticism. As Tuerlinckx (2004) argues, a repeated mood questionnaire might be unproblematic but any repeated ability test or measure will be confounded by complications such as learning, memory, and carry over effects.

The causal attribution contentions of Levy (1970) and Murray (1938) is renounced as above. That is, given enough replication and studious examination, increasing confidence for general principles extracted from particular instances can be engendered. Tuerlinckx's (2004) argument is more cogent but there exist a range of design considerations in single-subject research that allow for accurate causal attributions to be inferred. Multiple baseline designs, statistical filtering techniques, and reversal designs are but several germane examples (Molenaar, 2004; Saville, 2008).

Benefits of IAV Research

Indeed, it is the repeated measurement of variables that is a key strength of IAV research. Perone (1999), suggests that measuring the dependent variable only once, as is common in IEV research, precludes the intensive interaction with data that facilitates a true

understanding of research phenomena. Repeated measures also allow participants to experience all treatment conditions and can help researchers to identify and control further extraneous factors that may be introducing unwanted variability into a study (Saville, 2008).

An additional benefit with IAV research concerns the nature of data analysis. Whereas IEV research typically uses test of significance and statistical aggregation for inferring causality, IAV research typically relies on visual analysis for experimental inference. This enables researchers to uncover patterns of coherence that might be otherwise overlooked with IEV research and develop more effective treatment measures (Cervone, 2005; Haynes, Mumma, & Pinson, 2009; Saville, 2008; Tukey, 1969). As noted by Parker and Hagen-Burke (2007), by being able to simultaneously detect curvilinear trends, repeating patterns or cycles in data, delayed or lagged responses following intervention onset, and within-phase changes in variability along with changes in mean levels and trend slope across phases, visual analysis displays an inferential power unequalled by any other analytic technique.

To understand how patterns of intraindividual variability may be missed in IEV research, consider an example given by Cervone and Shoda (1999) depicted in Figure 2.

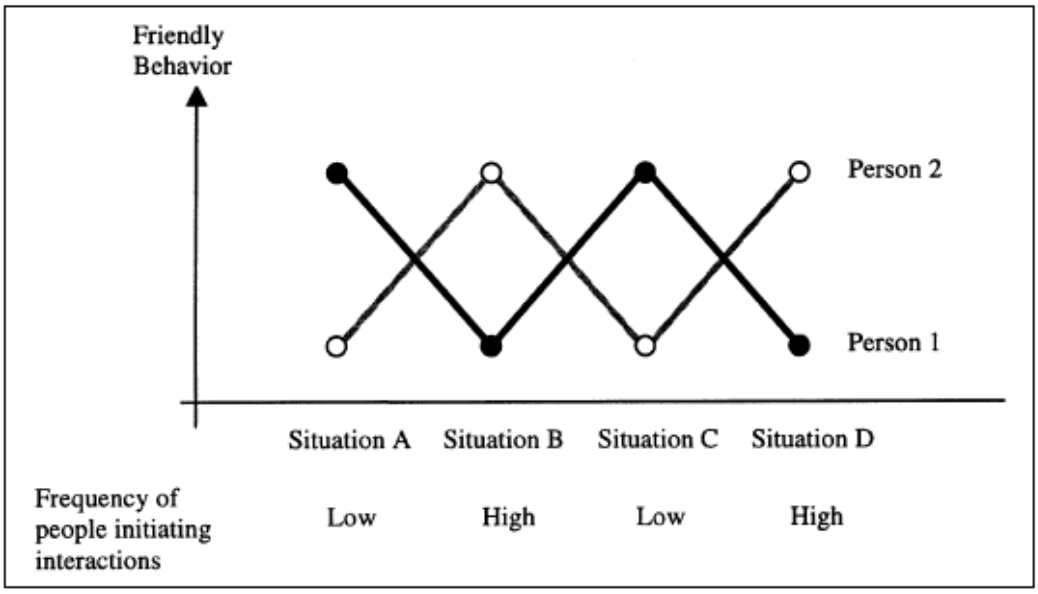

Figure 2: The loss of personality variability when considered in aggregate. Reprinted with permission from Cervone and Shoda, 1999, p.30.

Person 1 tends to be happy when left alone and becomes irritated by frequent social interaction, whereas Person 2 becomes irritated when being ignored. Given situations A and C which feature few social interactions, and situations B and D which include frequent social contacts, it is likely Person 1 will be unfriendly in Situations B and D, but not A and C, with the pattern reversed for Person 2. While both may exhibit the same degree of friendliness across the four situations when considered in aggregate, all evidence of consistent personality variability is lost.

Interindividual Research

The Prime Premise

A prime goal of psychological science is to discover functional relations among independent variables over and above the welter of environmental and biological variables influencing an organism at any given time (Barlow & Nock, 2009). The premise underlying modern IEV approaches is that such variability can be overcome with recourse to advanced statistical techniques which look through such *error* and extract, valid, general psychological laws from the study of groups of individuals (Barlow & Nock, 2009; Epstein, 2010; Krauss, 2008; Lamiell, 1998; Salvatore & Valsiner, 2010). Sidman (1960) sagely explicates this strategy: "The rationale for statistical immobilization of unwanted variables is based on the assumed random nature of such variables. In a large group of subjects, the reasoning goes, the uncontrolled factor will change the behaviour of some subjects in one direction and will affect the remaining subjects in the opposite away. When the data are averaged over all the subjects, the effects of the uncontrolled variables are presumed to add algebraically to zero. The composite data are then regarded as though they were representative of one ideal subject who had never been exposed to the uncontrolled variables at all." (p. 162).

Implicit within such an approach are several complementary assumptions. The first is that which is true of a sample is true of a population. A second assumption presumes that which is true at one time will also be true at another, while a third assumes that which is true of a group is true for individuals within that group (Cervone & Shoda, 1999; Krauss,

2008; Molenaar, 2004; Salvatore & Valsiner, 2010). There are, however, serious reasons to doubt all three assumptions.

Sampling

The case of the representativeness of samples of populations used in modern IEV research is highly suspect. As illustrated earlier in the chapter on null hypothesis significance testing, participants in the majority of psychological studies represent convenience samples comprising a homogenised cohort of university students, and researchers often make no attempt to define the population of which the samples are purported to be representative (Kline, 2004). Standard errors are likely to be understated, with the result that the statistical significance of such research is likely to be overstated (Reichardt & Gollob, 1999). Additionally, researchers are likely to generalise such results outside the bounds of such narrow samples (Saville, 2008). Ultimately, both of these practices can greatly misrepresent the conclusions drawn from the outcomes of a single IEV study (Wilkinson and the Task Force on Statistical Inference, 1999).

Cross-Situational Consistency

In a famous critique of personality psychology, Mischel (1968) showed that the cross-situational consistency of IEV research was often surprisingly low. The criticism ignited a *person-situation* debate that lasted throughout the 1970s and generated a sense of paradigm crisis (Cervone & Shoda, 1999; Shoda, 1999). Eventually, however, a trait revival was sparked by Epstein (1979) amongst others, who showed that stable situational personality dispositions resulted when behaviour was aggregated over a reasonable number of occurrences (Epstein, 1979; Cervone & Shoda, 1999). Thus, rather than detract from the process of aggregation, Mischel's (1968) criticism had provided another example of the benefits: "As a consequence, the scientific study of personality dispositions, which has been cast into doldrums in the1970s, is again an intellectually vigorous enterprise'' (Goldberg, 1993, p. 26).

The key element overlooked by IEV proponents is that Epstein's (1979) paper also laid the foundation for a repudiation of the assumptions underpinning IEV research. This is best illustrated by Epstein's comments in the discussion section of his paper: "This indicates that one can predict average behavior accurately from a similar sample of average behavior. However, the prediction is only actuarial and is a far cry from

predicting with confidence to individual instances of behavior… it should be noted that not everyone is equally predictable. This was demonstrated in the present article by the finding that within-subject correlations varied over a range that suggested almost, no stability in a few individuals and extremely high stability in others, with most individuals demonstrating a moderately high degree of stability." (p. 1123-1124). That is, although aggregation helped demonstrate situational consistency in some individuals, it was by no means universal. Since Epstein (1970), findings that people differ in response tendencies, rates of change, and patterns of variability across contexts, in ways that cannot be captured by mean response levels, have been demonstrated by numerous researchers (Cervone, 1997; Cervone & Shoda, 1999; Epstein, 2010; Hemenover 2003; Shoda, 1999; Vansteelandt & Van Mechelen 2004). Perhaps the strongest refutation of the IEV approach concerns the assumption that individuals within groups can be characterised by aggregated statistics.

Group and Individual Differences

As noted above, the universal and unquestioned adoption of group averaging of data in psychology emerged as a by product of statistical advances and practical imperatives that precipitated the inference revolution. The ideological roots of the application of statistical aggregation to evaluating psychological attributes are far older, however, and can be traced to Quetelet, who in 1844 argued that the normal curve that had been applied to astronomical errors could also be used to define l'homme moyen[15] (Kasser, 2006; Stigler, 1992). Though statistical aggregation is often an appropriate strategy for making inferences from samples to populations in domains such as agronomy or quality control, where data are so numerous to preclude examination of every instance, and where inter-individual variation in properties are usually not of any concern, there are good reasons to question statistical aggregation when applied to psychological phenomena (Bakan, 1966; Blampied, 2011; Salvatore & Valsiner, 2010).

As highlighted by Molenaar (2004), classical theorems of ergodic mathematics[16] illustrate that most psychological processes are non-ergodic. That is, the structures of

[15] *L'homme moyen*, literally translated means "the ideal human".

[16] Ergodic theory is a branch of mathematical statistics and probability theory.

IAV and IEV research are not asymptotically equivalent (Molenaar, 2004). Ergodic theorems assert that analysis of interindividual variation will fail to correspond to the pattern of intraindividual variation when: a mean trend changes over time, a covariance structure changes over time, or when a process occurs differently for different members of the population. Unfortunately, these are precisely the conditions that characterise much psychological research (Krauss, 2008; Molenaar, 2004; Salvatore & Valsiner, 2010).

A plethora of empirical findings back up the theorems of ergodic mathematics. For example, Borkenau and Ostendorf (1998) had 22 participants self report items indicative of the Big 5 factors of personality, daily, over a period of 90 days. Though substantial consistency was evidenced for the factor structure of longitudinal rotations that had been averaged across all participants, the five-factor model only fitted the intraindividual organisation of psychological dispositions for fewer than 10% of the individual participants. That is, for most participants the supposed 5 factors of personality gave way to 2, 3, 6 and even 8 factor structures (Borkenau & Ostendorf, 1998). Similar findings have been obtained by Cervone (2005), Cervone and Shoda, (1999), Epstein, (2010), Grice (2004), and Grice, Jackson, and McDaniel (2006).

To understand how intraindividual tendencies can fail to fit the five-factor model, consider someone who regularly experiences positive affect and acts impulsively whenever they are not feeling self-conscious. Positive affectivity is a facet of Extraversion, whereas impulsivity and self-consciousness are facets of Neuroticism. As Extraversion and Neuroticism are independent dimensions, the 5 factor model does not anticipate robust within-person correlations between positive affect and the other two attributes. Moreover, impulsivity and low self consciousness are at opposite ends of the Neuroticism dimension, and thus correlate negatively, not positively (Cervone & Shoda, 1999). This is doubly troubling, for not only may many models be inadequately fit to individuals but doing so might also obscure other items of interest within the data (Cervone, 2005; Epstein, 2010).

That IEV approaches fail to elucidate and explain intraindividual variation should come as little surprise. As Harré (1998) notes taxonomic, classificatory concepts are of an incorrect logical type to serve as explanations within the domain of psychology. This is due to the fact that most latent variables are conceptualised as unchanging static

constructs. Because these variables cannot covary with their supposed effects, the premise that a position on a latent variable distribution causes a subject's item response is fallacious (Borsboom, 2003, 2005; Krauss, 2008).

Summary

Despite the vitriolic rhetoric, the criticisms of IAV research are unfounded. The focus on direct and systematic replication, repeated measurement of variables, and visual analysis of data are the key advantages of IAV research that make it an efficient, scientific tool for investigating human behaviour and making causal attributions. Far from being a chaotic, non-generalisable, antiscience undertaking, the greater intimacy, understanding, and control of psychological phenomena afforded by IAV research leads to the construction of valid and generalised psychological knowledge. As Sidman (1960) has emphasised, "Experience has taught us that precision of control leads to more extensive generalization of data" (p.152).

By contrast, the generalisations generated by IEV research, which are so often accorded great significance within psychology, are premised on shaky ideological foundations. In generalising beyond the boundaries afforded by samples of narrow, homogenised, cohorts, IEV conclusions can be overstated and misattributed. Though aggregating can increase measures of situational consistency, casting individual variation into the statistical darkness known as error variance (Cronbach, 1957) obfuscates patterns of consistent, idiosyncratic personality dispositions. Ergodic theory backed by empirical research show that psychological processes are in the main non-ergodic; the individual cannot be described by the aggregate. As Harré (1998) succinctly notes: "To say that a chimpanzee is a primate does not explain anything about its characteristics" (p.80-81).

Conclusion

A Redundant Debate

Though Allport (1937) is often credited with introducing the concepts *idiographic* and *nomothetic* into the psychology lexicon, it was in fact Munsterberg (1899/1994) who first used the terms in psychology. The genesis of these concepts, however, can be attributed

to Windelband (1884/1998), from whom both Munsterberg and Allport borrowed the terms. Unfortunately, issues of translation and inconsistent distinctions have seen the commonly interpreted meaning of these terms drift far from the ideological mooring of Windelband's original conception. Far from being a harmonious and consistent application of two complementary modes of scientific endeavour, the terms *idiographic* and *nomothetic* have been erroneously associated with various antagonistic issues concerning the study of groups and individuals, qualitative versus quantitative research methods, personality traits, and the status of psychology as a science.

The Fragmentation of Psychology's Research Tradition

The most serious consequence for psychological science to emerge from the terminological debate has been the fragmentation of the psychological research tradition. Early psychological research methodology was borrowed and adapted from various scientific disciplines and characterised by a state of flux. Whereas the Wundtian school focused on introspective, perceptive, and sociocultural elements within a person, the Galtonian paradigm's prime concern was for mental measurement and comparison between individuals.

By the 1920s the Galtonian paradigm had begun to emerge as the dominant approach and its emergence was facilitated by educational and military demand for practical applications which shaped laboratory research practices. This trend was exacerbated with the demand for social investigation emerging from government and private research foundations, and with the advances afforded by revolutionaries such as Fisher and Stevens. By the 1950s, the study of IEV via the statistical aggregation of groups of individuals had become the authoritative research approach in psychology. To this day, IEV research remains, if anything, an even more dominant research tradition with journals historically associated with IAV research wandering from their conceptual roots by employing aggregated statistical methodology.

A Nomothetic Illusion

In striking contrast with Windelband's original intention, IEV research in psychology today is characterised by a reductive framework in which nomotheticity has been taken to

be synonymous with ergodicity. That is, IEV and IAV are presumed to be equivalent: over time, between samples and universal populations, and between groups and individuals within such groups. These assumptions are completely untenable as the theorems of ergodic mathematics and a legion of empirical research demonstrate. Psychological samples are not generally representative of populations, psychological phenomena are not static over time, and individuals vary in ways that cannot be captured by group statistics. In contrast, IAV research with its focus on replication, control, and repeated measurement of variables is uniquely suited to discovering the consistent, idiosyncratic patterns of human variability missed by IEV methods, and well as the accumulation of robust, generalisable, psychological knowledge.

An Addressable Imbalance

That IEV research does not, and cannot investigate, test, imply or evaluate causal accounts of individual behaviour is clear, for: "our theories are formulated in a within-subjects sense, but the models we apply are often based solely on between-subjects comparison" (Luce, 1997, cited in Borsboom, 2005, p.83). Both the logic and evidence presented above illustrate that there can be no justification for assuming that relations among individuals are consistent with relations within individuals; they are not the same. The defining characteristic of man *is* his individuality (Allport, 1937) and by "stripping the person of all his troublesome particularities, general psychology has destroyed his essential nature." (Allport, 1967, p. 549).

Meehl (1992) remarked that "no statistical procedure should be treated as a mechanical truth generator" (p. 152). The pathological and one-sided fascination with the aggregate at the expense of the individual can, at best, result only in impoverished understanding of psychological phenomena. These sentiments are endorsed by Loftus (1996) who opined: "What we do, I sometime think, is akin to trying to build a violin using a stone mallet and a chain saw. The tool-to-task fit is not very good, and, as a result, we wind up building a lot of poor quality violins." (p.161).

As Kluckhohn and Murray (1953) have expressed the issue, every human is in certain aspects like all other humans, like some other humans, and in other ways like no other human. If research methods truly are to be the mortar that binds the various psychology

sub-disciplines together (Stanovich, 2004), then the imbalance between IEV and IAV research must be redressed to accommodate investigation of all dimensions and levels of human variability. Failing to do so only further threatens the foundation on which psychology's quest for relevant, veracious, and cumulative scientific knowledge is premised.

Chapter 5: Observation Oriented Modelling

Overview

Outline

Metaphysical Misgivings

Observation Orientated Modelling (OOM) is a novel methodology for conceptualising and evaluating psychological data created by James Grice in response to the metaphysical issues plaguing psychology's major research tradition (Grice, 2011). Citing such luminaries as Meehl, Bakan, and Cohen, along with many of the arguments detailed above, Grice asserts that the predominately positivist research tradition that characterises contemporary psychology has thwarted accumulation of genuinely scientific knowledge about people. OOM is thus advanced as a unique alternative for explaining patterns of observations in terms of their causal structure, in stark contrast with the variable and parameter orientated approach of psychology's prevailing research paradigm.

Grice's (2011) view echoes the earlier sentiments of editors of 24 scientific journals who asserted that traditional, variable-orientated, sample-based, research strategies were ill-suited to accounting for the complex causal processes undergirding psychological phenomena (NIMH consortium of editors on development and psychopathology, 2000). Barrett (2008) best exemplifies the consensus criticism of these strategies: "Instead of simply *dealing* with this state of affairs as scientists might, which is to think about why it is proving so difficult and perhaps concentrate more on methods for establishing some decent levels of predictive accuracy of whatever we hold as "important," psychometricians have instead created a self-sustaining illusion that data-model-driven statistical complexity equates to more accurate science. Where is the evidence for this proposition?" (p.81).

Philosophical Underpinnings

In contrast with the positivist philosophy of traditional psychological research methods, OOM is premised on a moderate philosophical realism exemplified by Aristotle and Thomas Aquinas (Grice, 2011). Broadly described, such realism holds that that there is

an independently existing natural world which humans are able to successfully cognize via observational methods, and that such knowledge is a reliable guide to individual and social actions. Seven principles of OOM result, whereby primacy is given to real, accurate, repeated, observable events and researchers are encouraged to think through the lens of an integrated model that incorporates statistical outliers but avoids aggregation and population inferences, except in a limited fashion. In Grice's (2011) own words: "…observation oriented modelling shifts the focus of analysis away from computed aggregates such as means and variances onto the observation themselves. In other words, the focus is shifted to the people, specific behaviours, animals, things, events, etc., under investigation…The psychologist instead worries less about fulfilling untenable assumptions… and thinks more about the patterns of ordered observations relative to a competing perspective of chance" (p.40).

Methodology

Deep Structure

At the core of OOM are the *deep structures* of qualitatively and quantitatively ordered observations. Such structures are obtained by translating data elements into binary form. For example, the deep structure of biological sex can be represented "1 0" for females and "0 1" for males. Similarly, in considering a 5-point Likert scale with highly disagree, disagree, neutral, agree, and highly agree categories, a highly disagree response would be recorded as "1 0 0 0 0" whereas an agree response would be coded "0 0 0 1 0". Multiple observations can be then recorded in standard matrix convention:

$$\text{Gender} = {}_6\mathbf{Y}_2 = \begin{vmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{vmatrix} \qquad \text{Likert} = {}_6\mathbf{X}_5 = \begin{vmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{vmatrix}$$

Procrustes Rotation

In matrix form, deep structures can be manipulated according to the rules of matrix algebra which allows for addition, subtraction, and other logical operations such as *if*, *and*, *not,* etc. The primary mathematical technique employed in OOM analysis, however, is referred to as *Binary Procrustes Rotation*, a modified form of Procrustes rotation first published by Green (1952)[17]. In OOM, the observed categories such as people, behaviours, and actions are always assigned to the rows of deep structure matrices whereas the units of observations are assigned to the columns. The goal of the binary Procrustes rotation then, is to align the column (units) in such a way that 1s in the *target matrix* (Likert response) are maximised with co-occurrence of 1s in the *conforming matrix* (gender). A process of normalisation is then used to ensure values in the final, fully rotated matrix do not exceed 1, and preserves the scale of the original deep structures.

Classification Strength Index

In an ideal outcome, the transformed values of the conforming matrix will perfectly match the values in the target matrix. This is not always the case, however, with possible values ranging between 0, no match at all, and 1, perfect agreement. The largest value in each row is used as an indicator of the extent to which the rotation clearly discriminated between the various units in each observation and is termed the C*lassification Strength Index* (CSI).

Percent Correct Classification Index

Though the CSI reveals the discrimination between units afforded by the rotation, it does not indicate the degree of similitude between rows of the target and conforming matrices. Consider a case in which a row of a conforming matrix results in values of 0, .50, and .88 but where a target matrix possesses values of 0, 1, 0. Clearly, in this instance the classification is erroneous indicating that a better match was obtained for a different variable. The *Percent Correct Classification Index* (PCC) provides the number of correct matches obtained between the conforming and target structure rows with a higher value demonstrating better similitude between cause and effect (Grice, 2011).

[17] Hurley, Cattell, and Schoneman are also credited with developing and popularising the technique (Grice, 2011).

C-Value

The success, or not, of the classifications within an OOM analysis raises an important question, namely, what constitutes success? Though 100% is obviously the gold standard, should 66% be considered an impressive outcome? To help answer this question, the OOM software employs a resampling procedure that involves randomly shuffling the rows of the conforming deep structure, applying the Binary Procrustes Rotation, and recalculating the PCC value. This process is repeated numerous times (500-1000) and a probability statistic is generated, termed the *chance value* (c-value), that reflects the frequency with which resampling provides results at least as accurate as the initial observed data. Low c-values thus provide an indication of the uniqueness particular to a set of observations (Grice, 2011).

Multigram

Bakan (1966), citing Tukey (1962), pointed out that statistical procedures can "take our attention away from the data, which constitute the ultimate base for any inferences which we might make" (Bakan, 1966, p.436). Heeding these words, and in concert with the visual analysis methods commonly employed in IAV research, a *multi-level frequency histogram*, or simply *multigram*, is also utilised to aid the inspection, analysis, and evaluation of psychological data. A multigram is created using units of the target deep structure as columns with units of the conforming deep structure represented as rows (Grice, 2011). As can be seen in the example below, red bars are used to represent incorrectly classified observations and green to correctly classified observations.

Generalisation

As noted above, the primary concern in OOM is the focus on phenomena. As such, representative statistical aggregates for hypothetical population parameters are given less consideration, if not ignored completely. Generalisability of findings, in harmony with IAV approaches, is the result of replication and theoretical considerations: "It is important to point out here that the psychologist's scientific desire to generalise…is satisfied not through an appeal to population parameters but to replication and theory" (Grice, 2011, p.40).

Example

Matrix Output

In order to flesh out the concepts listed above, consider the two hypothetical matrices below that represent the deep structures of gender (target matrix) and depression ratings (conforming matrix) on a 7-point Likert scale ranging from -3, extremely happy to +3, heavily depressed. Combined, they are represented in a transformational matrix. For example, the top left hand figures, 2 and 3, of the transformation matrix represents the frequency with which 1s were found for males endorsing -3 and -2 ratings.

$$\begin{vmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{vmatrix} \qquad \begin{vmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{vmatrix} \qquad \begin{vmatrix} 2 & 0 \\ 3 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 2 \\ 0 & 3 \end{vmatrix}$$

Gender = ${}_{10}Y_2$ Depression = ${}_{10}X_7$ Trans = ${}_7T_2 = {}_7X'_{10}\ {}_{10}Y_2$

Normalising the transformation matrix by columns and then by rows results in the fully rotated matrix which is compared to the original target matrix. As can be seen below, they are identical indicating that 7 unit deep structure depression ratings could be conformed and reduced to their 2 unit gender deep structures. That is, gender is the causal factor precipitating depression in this model.

$$\begin{vmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{vmatrix} \qquad \begin{vmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{vmatrix}$$

Fully Rotated Matrix = ${}_{10}X_7\ {}_7T_2$ Gender = ${}_{10}Y_2$

OOM Statistical Output

In OOM, output from the results are as documented below. In noting the strong classification strengths of each observation, the high percent classification results, and the low c-value, it can be inferred that there is strong support for the model that gender causes depression. Visually, this is confirmed by the multigram in Figure 3 which shows clear separation of distributions.

Individual Classification Results

	Classification Result	Classification Strength	Target Deep Structure	Classified Deep Structure	Conforming Deep Structure
obs_1	C	1.00	Male	Male	Extremely Happy
obs_2	C	1.00	Female	Female	Depressed
obs_3	C	1.00	Male	Male	Happy
obs_4	C	1.00	Male	Male	Happy
obs_5	C	1.00	Female	Female	Heavily Depressed
obs_6	C	1.00	Female	Female	Heavily Depressed
obs_7	C	1.00	Male	Male	Extremely Happy
obs_8	C	1.00	Male	Male	Happy
obs_9	C	1.00	Female	Female	Depressed
obs_10	C	1.00	Female	Female	Heavily Depressed

Note. C = Correctly Classified, I = Incorrect, A = Ambiguous.

Randomization Results

Observed Percent Correct Classified : 100.00

Number of Randomized Trials : 100.00
Minimum Random Percent Correct : 40.00
Maximum Random Percent Correct : 100.00
Values >= Observed Proportion : 1.00
Model c-value : 0.01

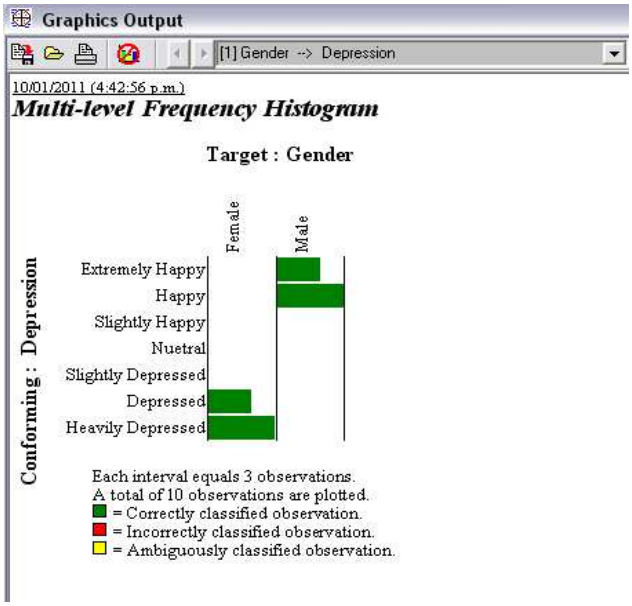

Figure 3: Multigram depicting depression results by gender.

Evaluation

Advantages

OOM and the Measurement Problem

The primary advantage of using OOM for the evaluation of psychological data is that it doesn't rely on restrictive and unrealistic assumptions. By employing binary translations of data and focusing solely on the structured patterns of causal phenomena, issues concerning the measurement problem, NHST, and IEV research are avoided. For example, whereas most traditional approaches to psychological measurement dubiously presume all psychological variables posses quantitative structure, which allow for analysis by linear statistical techniques such as factor analysis and structural equation modelling, OOM does not. In such cases, conclusions from OOM research attain a legitimacy that cannot be attained using other traditional measurement techniques (Grice, 2011).

OOM and NHST

The overly restrictive assumptions and resulting criticisms underpinning NHST are also evaded when employing an OOM approach to research. Assumptions such as linearity, equal population variances, and constraints pertinent to power analysis such as sample

size requirements are all germane examples that are irrelevant to OOM considerations (Grice, 2011). With fewer restrictions to violate, OOM affords a realistic, flexible rationale to modelling and appraising psychological phenomena that cannot be attained with prevailing research practices such as NHST.

OOM and IEV

Similarly, in dealing with phenomena at the individual observation level, drawbacks of IEV research approaches are sidestepped. Mathematical theorems concerning the ergodicity of psychological variables are respected. Additionally, in recognising and valuing outlying results, rather than interpreting such results as uninterpretable noise (Grice, 2011), and by employing methods of visual analysis, patterns of unique causal variability can be uncovered that are missed by IEV strategies.

Flexibility and Efficiency

Although OOM's strength is its ability to use discrete countable qualities, continuous qualities can also be analysed (Grice, 2011). When combined with the rich analysis options afforded by OOM, such as *model observation separation* and *pairwise rotation analyses*, OOM can replace Chi Square tests, t-Tests, correlation analysis, ANOVAs, MANOVAs, bivariate multiple regression and item comparison procedures of traditional statistical inference (Grice, 2011). By avoiding the activity required in learning and verifying assumptions, transforming data, and pondering statistical significance of results used in traditional statistical analysis, OOM can greatly simplify and economise the data analytic process (Grice, 2011).

Contentions

Matrix Restrictions

Though OOM is rightly seen as a flexible, efficient method for analysing the causal patterns within psychological data, it does have limitations that researchers need to be cognizant of. The first limitation concerns the nature of matrix algebra employed in OOM. At least one dimension (row or column) must be consistent to both the target and conforming matrices to allow the Binary Procrustes rotation to take place (Grice, 2011). For instance, though a 2x6 matrix could be rotated with a 10x6 matrix, a 4x3 matrix could not be conformed to a 2x5 matrix.

Similarly, it should be noted that rotations are asymmetric. That is, switching target and conforming matrices and rotating will give different results to a reversed ordering. Small number of units will result in ambiguous classifications more often when conformed to large number of units (Grice, 2011). For both these reasons, therefore, the selection and ordering of the units and variables of observation are a prime consideration.

Binary Multiplication and Division

Similarly, multiplication and division are operations that cannot be supported in binary form by OOM (Grice, 2011). In contrast with logical, addition, and subtraction processes, all multiplication and division of data elements must be performed before a binary translation occurs. In so doing, the assumptions of quantity underpinning such operations must be met. In this case OOM is as constrained as other traditional approaches.

Resampling and the C-value

As highlighted above, a key strength of OOM is the relative lack of assumptions intrinsic to it. That's not to say it has no assumptions, however. For example, three assumptions underpin the randomised resampling test, namely (1) that conforming observations are independent, (2) that target observations are independent, and (3) the conforming observations are independent of the target observations. Although minimal when compared with assumptions of NHST, verification of these conditions should take place, with violations of independence potentially resulting in erroneous inferences derived from the c-value.

It should also be noted that as the size of the data set diminishes so too does the value derived from the c-value, with fewer random permutations available from which comparison may be obtained. Even with large data sets and verified assumptions, however, an inferential conservatism is warranted. A number quantifying the probability of evidence should not be conflated with the strength of evidence (Haig, 2011). To do so is reminiscent of p-values and the effect size fallacy elucidated earlier in the NHST chapter.

Magnitude

A similar inferential caution is required in regards to issues of magnitude. Though OOM preserves the scaling of original data in binary transformation and rotation of matrices, it

is important that quantity considerations should also be attended to by researchers in addition to focusing on ordered patterns. For example, it is perfectly possible for two conformed matrices from two different rotations to possess the same structure. However, the magnitudes expressed by each matrix may be entirely different. Descriptive statistics are available within the OOM software and should be used where necessary to supplement conclusions drawn from OOM.

Conclusion

The Primacy of the Real

OOM is a new method for evaluating psychological data created in response to shortcomings identified with psychology's dominant research tradition. In contrast with this dominant positivist tradition, OOM is premised on philosophical realism. As such, OOM grants primacy to real, accurate, repeated, observations rather than abstract statistical aggregates, and allows for greater insight into, and better explanation of, patterns of observations in terms of their causal structure.

Conformity of Ordered Structures

Using binary translation of data elements, data matrices are created which are aligned to one another via a Binary Procrustes Rotation. The CSI reflects the degree to which units (columns) in the final, transformed, rotated matrix cohere with units of the original target matrix, whereas the PCC reflects the consistency between rows of the conforming and target matrices. A c-value indicates the degree of uniqueness of generated results against a competing perspective of chance. These statistical outputs, in conjunction with visual analysis of results provided by a multigram, along with replication procedures, are the primary mechanisms by which causal structure is inferred, and generalised within OOM.

Assumptive Parsimony

An OOM approach to the analysis and interpretation of psychological data holds many advantages over traditional research methodologies. The primary benefit concerns the dearth of assumptions that undergird OOM. With few postulates concerning the nature of

error, quantitative structure, linearity, and ergodicity, OOM overcomes many of the troubling aspects intrinsic to traditional approaches as documented in the measurement problem, the use of NHST, and the nature of IEV research. OOM, therefore, offers a robust, economical approach to data analysis that can replace many other statistical techniques employed in psychology's dominant, variable-orientated, sample-based research paradigm.

Data Intimacy

Though not without methodological caveats and restrictions, such as binary multiplication and division, most caveats relate to researcher utilisation of OOM rather than weaknesses fundamental to the method itself. Indeed, it could be argued in several instances that several of the restrictions, in fact, afford better research outcomes. For example, in adhering to the constraints of matrix algebra and the attendant evaluation of the units of analysis, researchers are forced to stay "close to the observations as they are ordered" (Grice, 2011, p.173). As a consequence, a greater intimacy with experimental data is achieved that can facilitate a better understanding of research phenomena (Perone, 1999).

Chapter 6: Conclusion

Issues in Contemporary Psychology

The Measurement Problem

Philosophical monism, quantification, and practicalism were three characteristics that united and typified early approaches to psychological measurement taken by pioneers such as Fechner, Cattell, and Thorndike. Nevertheless, criticism of such approaches abounded, reaching its zenith in the Ferguson Committee findings of 1940. Far from signalling the demise of psychological measurement, however, the convening of the Ferguson Committee can be seen as the crucible from which contemporary psychological measurement was born. By operationalising a representative theory of numerical assessment, and binding measurement to statistical procedures in response to the Ferguson Committee criticisms, Stevens revolutionised psychological measurement.

This revolution was unfortunate, for Stevens' formulation fails to meet the axioms of true quantitative measurement originating with Euclid and formalised by Holder. A similar conclusion can be drawn for Rasch modelling. Though, resembling, in style, the legitimate derived measurement practice of conjoint measurement theory, in substance Rasch modelling falls short of true quantitative measurement. The theory of conjoint measurement represents a significant achievement in mathematical psychology. However, it is in reality difficult to apply to many psychological attributes. This is an argument understood and expanded upon by Trendler (2009), who cites the lack of control afforded by psychological apparatus as the primary weakness retarding the accurate, quantitative measurement of psychological variables.

It should be noted that there have been measurement successes in the field of perception, such as loudness and brightness. However, the vast majority of psychological chracterisitcs have never been properly quantified and, thus, have never been properly measured. Though many psychological properties may in fact be quantitative, inferences from psychological measurement are of uncertain and/or dubious scientific utility without investigating and demonstrating direct or derived additivity. The results of failing to meet

this requirement are forcefully delineated by Barrett (2008): "The real-world consequences of this systematic aversion to properly considering the presumed [measurement] status of a psychological variable is that our journals are now filled with studies that are largely trivial exemplars of mostly inaccurate explanations of phenomena. Nothing seems to have changed since Lykken's similar observations of this phenomenon back in 1991." (p. 79-80).

Null Hypothesis Significance Testing

The current hybridised statistical methodology employed in psychological science is a result of pioneering work by Gosset, and a blend of statistical strategies of Fisher, and Neyman and Pearson. Such an amalgamation is, however, philosophically untenable and logically invalid. The assumptions underpinning the methodology are restrictive, ill-suited for application to many psychological phenomena, and often go unverified. The consequence is a fatally flawed technique that Bayesian arguments, Meehl's conjecture, and various other criticisms demonstrate.

Despite such deficiencies, NHST remains to this day the most popular statistical methodology employed in psychology and many social sciences. Attempts at reform, such as changes to journal editorial polices and concerted campaigns for abandonming NHST in favour of other approaches, have done little to dent the pathological over-reliance of NHST among psychologists. This situation is aggravated by orders of magnitude when research showing the paucity of understanding of NHST, and concomitant resulting fallacies displayed by researchers, is considered. Though NHST can be used as a means of assessing sampling variability, power considerations and the advantages of confidence intervals almost entirely preclude, or obviate, the legitimate application of NHST to sampling assessment within psychology.

Twenty years on from Lykken's (1991) imploration to give away "the nearly futile null hypothesis testing to which we have become addicted" (p.37), NHST remains the most popular, misunderstood, misapplied, and misused methodology in psychology. The questionable quality of much psychological research that relies solely, or primarily, on NHST for empirical inference is but another reason "why so much social science has

turned out to be no more than transient description of never-to-be-reencountered situations, easy to contradict with almost any replication" (Wright, 1997, p.35).

The Granularity of Research Methods

The distinction between IEV and IAV research within psychology is often characterised as a nomothetic-idiographic contrast. This is, however, erroneous resulting from mistranslation and incongruent utilisation of the terms *idiographic* and *nomothetic* originally conceived by Windelband. The triumph of the Galtonian paradigm over the Wundtian paradigm, synonymous with statistical advances and practical imperatives, has resulted in a fracturing of psychology's research tradition. Today, this schism is gravely imbalanced with the overwhelming dominance of group-based, statistical aggregates evident even amongst journals with a traditionally IAV foundation.

This imbalance is deeply troubling for as theorems of ergodic mathematics and empirical research demonstrate, only in restricted cases can the individual be accurately described by the aggregate. That people differ in patterns of cross-situational consistency, in ways that cannot be captured with IEV statistics, has been illustrated by numerous researchers. This contrasts sharply with empirical work highlighting the robust, cumulative, generalisations afforded by IAV approaches.

Depending on the level of analysis, humans can be seen to resemble all, some, and no other humans. In focusing almost completely on the general, psychology has done little, theoretically and empirically, to capture the idiosyncratic variability inherent within individuals. As such, an incomplete and impoverished understanding of human psychology currently exists. The point is again understood and underscored in Lykken's (1991) critique: "To the extent that our brains are running different programs, no one nomothetic psychological theory is going to be able to account for all of us." (p.17).

Conclusion and Solution

Continued Cult Science

The three issues with psychology's dominant research tradition examined in this thesis were all cogently highlighted by Lykken in his classic 1991 critique. As the preceding

chapters demonstrate, little has changed since its publication. The pathological fascination with "the sizeless stare of statistical significance" (Ziliak & McCloskey, 2008), coupled with the failure to deal with issues of quantification and the imbalance between IEV and IAV research means many areas of psychology are still practicing "cargo-cult science". Twenty years of continually slow progress on from Lykken, a lot, therefore, remains wrong with psychology's research tradition.

Levels of Analysis

A particularly troubling aspect with the foregoing is that with some overlap, each of these three issue can be seen to address different "levels" of psychology's research tradition. The IEV versus IAV issue is concerned with description, focuses on observations, and addresses matters of how best to perceive and evaluate data. The NHST issue is at a higher level concerning methodological application and inferences derived from the observations at the descriptive level. Similarly, the quantification issue can be seen to be at a still higher level focusing on meta-level philosophical issues underpinning the methodology. Thus, it is not just one level of psychology's dominant research tradition that requires attention but indeed every level.

Practicality Over Philosophy

That practicalism, buttressed by the misuse of mathematics, has overridden philosophical sensibilities explains much of the current situation in psychology. For example, the practical demand for psychological testing coupled with Stevens' scales of measurement, has seen the issue of quantification disappear from psychology's radar. Similarly, the practical pressures for mental and educational applications of psychological knowledge that initiated group testing and averaging, in direct opposition to the theorems of ergodic mathematics, has resulted in the serious IEV/IAV research imbalance and the consequent impoverished understanding of individuals. In the same manner, the practicalist imperatives behind the inference revolution and the hybridisation of Fisher, and Neyman and Pearson's statistical methodologies has seen the mindless, pathological application of NHST become the method of choice for statistical inference in psychology. As Danziger (1990) succinctly expressed it, psychology has demonstrated the "tendency for practical

technology to usurp the name of science" (p.128). Valsiner (2006) hihglights the consequences of such practical usurpation: "The result is predictable—a mindless accumulation of empirical publications in increasingly narrow 'research fields'. The latter are set by conventions rather than by theoretical needs. Psychology as science would probably suffer no loss if the overwhelming majority of empirical papers that are currently published never saw print." (p.604).

OOM Not Doom

It might be thought, from a reading of this thesis that no psychological measurement is possible, that IEV research has no value, that commonly used statistical methods are inadequate, and that, therefore, all psychological research is entirely without merit. However, this would be an incorrect inference. There is nothing intrinsically wrong with pursuing a sample-based, approach to IEV measurement that utilises aggregated linear statistical analysis *if* such an approach represents the best strategy for answering a given demographic research question. For other questions, however, other approaches may be required. Indeed, the KAPA model (Cervone, 2004), the theory of conjoint measurement (Luce & Tukey, 1964), and the Violence Risk Assessment Guide (Webster, Harris, Rice, Cormier, & Quinsey, 1994) are germane examples that renounced traditional research approaches because traditional research methodologies could not generate relevant empirical insight.

The challenges posed by the preceding issues result largely from the mindless, ritualised, one-size-fits-all application of traditional research approaches and methods to psychological research, without regard for relevant conceptual and philosophical considerations. Created in direct response to criticisms of the prevailing research practices, Observation Oriented Modelling is a flexible, efficient procedure for detecting and attributing causal relations amongst patterns of phenomena that can replace a variety of traditional statistical methodologies.

Being founded on philosophical realism, and focusing on phenomena at the individual level, OOM can reveal patterns of unique causal variability that are missed by IEV strategies. With few restrictive assumptions to violate, OOM affords a more realistic, flexible rationale to modelling and appraising psychological phenomena that cannot be

attained with prevailing research practices such as NHST. Similarly, in dealing with patterns of ordered structure, and without dubiously presuming all psychological variables posses quantitative structure, conclusions from OOM research attain a legitimacy that cannot be afforded using other traditional measurement techniques. OOM thus represents a unique, powerful addition to the researcher toolbox for helping to remedy "what's wrong with psychology" (Lykken, 1991, p.3).

References

(2000). NIMH Consortium of Editors on Development and Psychopathology. *Applied Development Science*, *4*, 66.

Allport, G. (1937). *Personality: A psychological interpretation.* New York: Holt.

Allport, G. (1961). *Pattern and growth in personality*. New York: Holt, Rinehart & Winston.

Allport, G. (1962). The general and the unique in psychological science. *Journal of Personality*, *30*, 405-422.

Allport, G. (1967). *The person in psychology: Selected essays by Gordon W. Allport.* Boston: Beacon Press.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423-437.

Barlow, D., & Nock, M. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, *4*, 19-21.

Barrett, P. (2003). Beyond psychometrics: Measurement, non-quantitative structure, and applied numerics. *Journal of Managerial Psychology*, *18*, 421-439.

Barrett, P. (2005). What if there were no psychometrics?: Constructs, complexity, and measurement. *Journal of Personality Assessment*, *85*, 134-140.

Barrett, P. (2008). The consequence of sustaining a pathology: Scientific stagnation, a commentary on the target article 'Is psychometrics a pathological science?' by Joel Michell. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 78-83.

Beins, B. (2010). Teaching measurement through historical sources. *History of Psychology*, *13*, 89-94.

Bem, D., & Allen, A. (1974). On predicting some people some of the time: The search for cross-situational consistencies in behaviour. *Psychological Review*, *81*, 506-520.

Berger, J., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p* values and evidence (with comments). *Journal of the American Statistical Association*, *82*, 112–139.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*, 526-542.

Blampied, N. (1999). A legacy neglected: Restating the case for single-case research in cognitive–behaviour therapy. *Behaviour Change*, *16*, 89-104.

Blampied, N. (2011). Single case research and the scientist-practitioner ideal. In G. Madden (Ed.). *Handbook of behaviour analysis.* Washington: American Psychological Association (in press).

Block J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin. 117*, 187-215.

Boring, E. (1919). Mathematical vs. scientific significance. *Psychological Bulletin*, *16*, 335-338.

Boring, E. (1920). The logic of the normal law of error in mental measurement. *The American Journal of Psychology*, *31*, 1-33.

Boring, E. (1957). *A history of experimental psychology.* New York: Appleton-Century Croft.

Borkenau, P., & Ostendorf, F. (1998). The big five as states: How useful is the five-factor model to describe intraindividual variations over time? *Journal of Research in Personality*, *32*, 202-221.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203-219.

Bridgman, P. (1927). *The logic of modern physics.* New York: Macmillan.

Byrd, J. (2007). A call for statistical reform in EAQ. *Educational Administration Quarterly*, *43*, 381-391.

Cafri, G., Kromrey, J., & Brannick, M. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research*, *45*, 239-270.

Campbell, N. (1920). *Physics, the elements.* Cambridge: Cambridge University Press.

Campbell, N. (1928). *An account of the principles of measurement and calculation.* London: Longman, Green and Co.

Cattell, J. M. (1890). Mental tests and measurements. *Mind, 15*, 373-380.

Cattell, J. M. (1893). Mental measurement. *Philosophical Review, 2,* 316-332.

Cattell, J. M. (1902). Proceedings of the Society for Psychical Research. *Psychological Review*, *9*.

Cattell, J. M. (1904). The conceptions and methods of psychology. *Popular Science Monthly, December*, 176-186.

Ceci, S., & Williams, W. (2007). Paul Wachtel was ahead of his time. *Applied and Preventive Psychology*, *12*, 13-14.

Cervone, D. (1997). Social-cognitive mechanisms and personality coherence: Self-knowledge, situational beliefs, and cross-situational coherence in perceived self-efficacy. *Psychological Science*, *8*, 43-50.

Cervone, D. (2004). The architecture of personality. *Psychological Revue*, *111*, 183-204.

Cervone, D. (2005). Personality architecture: Within-person structures and processes. *Annual Review of Psychology*, *56*, 423-452.

Cervone, D., & Shoda, Y. (1999). Beyond traits in the study of personality coherence. *Current Directions in Psychological Science*, *8*, 27-32.

Chow, S. (1996). *Statistical significance: Rationale, validity and utility*. Thousand Oaks: Sage Publications, Inc.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3*, 186-190.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145-153.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist*, *49*, 997-1003.

Cozby, P. (2004). *Methods in behavioural research.* New York: McGraw-Hill.

Craik, K. (1940). On the sone scale. *Advancement of Science, 1*, 335-340.

Cronbach, L. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671-684.

Cumming, G. (2008). Replication and *p* intervals: *P* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286-300.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532-574.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*, 299-311.

Cuthbert, B. (2007). Discontent in investigation: Plus ca change, plus c'est la meme chose? *Applied and Preventive Psychology*, *12*, 15-18.

Danziger, K. (1990). *Constructing the subject.* Cambridge: Cambridge University Press.

DuBois, P. (1966). A test dominated society: China 115 B.C. – 1905 A.D. In A. Anastasi (Ed.). *Testing problems in perspective.* Washington: American Council on Education.

DuBois, P. (1970). *A history of psychological testing.* Boston: Allyn & Bacon.

Epstein, S. (1979). The stability of behavior: On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*, 1097-1126.

Epstein, S. (2010). The big five model: Grandiose ideas about surface traits as the foundation of a general theory of personality. *Psychological Inquiry*, *21*, 34-39.

Erceg-Hurn, D., & Mirosevich, V. (2008). Modern robust statistical methods: An easy way to maximise the accuracy and power of your research. *American Psychologist*, *63*, 591-601.

Estes, W. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review*, *4*, 330-341.

Eysenck, H. (1954). The science of personality: Nomothetic! *Psychological Review*, *61*, 339-342.

Fechner, G. T. (1860). *Elemente der psychophysik.* Leipzig: Breitkopf & Hartel. (English translation by H. E. Adler, *Elements of Psychophysics,* vol. 1, D. H. Howes & E. G. Boring (Eds). New York: Holt, Rinehart & Winston).

Fechner, G. T. (1887). Uber die psychischen Massprincipien und das Weber'sche Gesetz. *Philosophische Studien, 4*, 161-230. (English translation of pp. 178-198 by S. Scheerer (1987). My own viewpoint on mental measurement. *Psychological Research, 49*, 213-219).

Ferguson, A., Myers, C., Bartlett, R., Banister, H., Bartlett, F., Brown, W., Campbell, N., Craik, K., Drever, J., Guild, J., Houstoun, R., Irwin, J., Kaye, G., Philpott, S., Richardson, L., Shaxby, J., Smith, T., Thouless, R., & Tucker, W. (1940). Final report of the committee appointed to consider and report upon the possibility of quantitative

estimates of sensory events. *Report of the British Association for the Advancement of Science*, *2*, 331-349.

Ferguson, C. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*, 532-538.

Feser, E. (2005). *Philosophy of mind: A short introduction.* Oxford: Oneworld Publications.

Feynman, R. (1986). *Surely you're joking, Mr. Feynman!* New York: Bantam Books.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119-126.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2005). Still much to learn about confidence intervals: Reply to Rouder and Morey (2005). *Psychological Science*, *16*, 494-495.

Finch, S., & Cumming, G. (2009). Putting research in context: Understanding confidence intervals from one or more studies. *Journal of Pediatric Psychology*, *34*, 903-916.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement*, *61*, 181-210.

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., & Goodman, O. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments & Computers*, *36*, 312-324.

Fisher, R. (1925). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Fisher, R. (1935). *The design of experiments.* Edinburgh: Oliver & Boyd.

Fisher, R. (1959). *Statistical methods and scientific inference (2nd ed.).* Edinburgh: Oliver & Boyd.

Fisher, R. (1966). *The design of experiments (8th ed.).* Edinburgh: Oliver & Boyd.

Forsyth, J., Kollins, S., Palav, A., Duff, K., & Maher, S. (1999). Has behavior therapy drifted from its experimental roots? A survey of publication trends in mainstream behavioral journals. *Journal of Behavior Therapy and Experimental Psychiatry*, *30*, 205-220.

Ghougassian, J. (1972). *Gordon W. Allport's ontopsychology of the person*. New York: Philosophical Library.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*, 254-267.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. New York: Oxford University Press.

Gigerenzer, G. (2004). Mindless Statistics. *Journal of Socio-Economics*, *33*, 587-606.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask. In D. Kaplan (Ed.). *Handbook on Quantitative Methods in the Social Sciences* (pp. 389-406). Thousand Oaks, CA US: Sage.

Gigerenzer, G., & Murray, D. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*, 791-806.

Goldberg, L. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*, 26-34.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, *42*, 139-167.

Gould, S. (1981). *The mismeasure of man*. New York: W.W. Norton & Company.

Green, B. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, *17*, 429-440.

Grice J. (2004). Bridging the idiographic-nomothetic divide in ratings of self and others on the Big Five. *Journal of Personality*, *72*, 203-41.

Grice, J. (2011). *Observation orientated modeling: An introduction.* New York: Elsevier (in press).

Grice, J., Jackson, B., & McDaniel, B. (2006). Bridging the Idiographic-Nomothetic Divide: A Follow-Up Study. *Journal of Personality*, *74*, 1191-1218.

Guild, J. (1938). Are sensation intensities measurable? *British Association for the Advancement of Science*, *108*, 296-328.

Hagen, R. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15-24.

Haig, B. (2011). The philosophy of quantitative methods. In T. Little (Ed.), *The Oxford handbook of quantitative methods.* New York: Oxford University Press (in press).

Haller, H., & Krauss, S., 2002. Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research*, *7*, 1-20.

Halpin, P., & Stam, H. (2006). Inductive inference or inductive behavior: Fisher and Neyman-Pearson approaches to statistical testing in psychological research (1940-1960). *The American Journal of Psychology*, *119*, 625-653.

Harré, R. (1998). *The Singular Self: An Introduction to the Psychology of Personhood.* London: Sage.

Haynes, S., Mumma, G., & Pinson, C. (2009). Idiographic assessment: Conceptual and psychometric foundations of individualized behavioral assessment. *Clinical Psychology Review*, *29*, 179-191.

Heath, T. (1908). *The thirteen books of Euclid's Elements*, vol. 2. Cambridge: Cambridge University Press.

Hemenover, S. (2003). Individual differences in rate of affect change: Studies in affective chronometry. *Journal of Personality and Social Psychology*, *85*, 121-131.

Hempel, C. G. *Aspects of scientific explanation.* New York: Free Press, 1965.

Higgins, E. (1990). Personality, social psychology, and person-situation relations: Standards and knowledge activation as a common language. In L. Pervin (Ed.). *Handbook of personality: Theory and research* (pp. 301-338). New York: Guilford Press.

Hinshaw, S. (2004). Commentary on Meehl. *Applied and Preventive Psychology*, *11*, 39-41.

Holder, O. (1901). Die Axiome der Quantitat und die Lehre vom Mass. *Berichte iiber die Verhandlungen der Koniglich Sachsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse,* 53, 1-46. (English translation of Part I by Michell, J., & Ernst, C. (1996). The axioms of quantity and the theory of measurement. *Journal of Mathematical Psychology,* 40, 235-252.)

Holt, R. (1962). Individuality and generalization in the psychology of personality. *Journal of Personality*, *30*, 377-404.

Hubbard, R. (2004). Alphabet soup: Blurring the distinctions between *p*'s and α's in psychological research. *Theory & Psychology*, *14*, 295-327.

Hubbard, R., & Bayarri, M. (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *The American Statistician, 57*, 171-182.

Hubbard, R., & Lindsay, R. (2008). Why *P* values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*, 69-88.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—And its future prospects. *Educational and Psychological Measurement*, *60*, 661-681.

Hubbard, R., Parsa, R. A., & Luthy, M. R. (1997). The spread of statistical significance testing in psychology: The case of the Journal of Applied Psychology, 1917–1994. *Theory & Psychology*, *7*, 545-554.

Hurlburt, R., & Knapp, T. (2006). Münsterberg in 1898, not Allport in 1937, introduced the terms "idiographic" and "nomothetic" to American psychology. *Theory & Psychology, 16*, 287-293.

James, W. (1961). *Psychology: The briefer course.* New York: Longmans & Green.

Jeffreys, H. (1939). *Theory of probability.* Oxford: Clarendon.

Kaplan, R., & Saccuzzo, D. (2005). *Psychological testing: Principles, applications, and issues.* Belmont: Wadsworth.

Kasser, J. (2006). *Philosophy of science (Volume 1).* Washington: The Teaching Company.

Kazdin, A. (2007). Methodological diversity can augment progress in psychological research. *Applied and Preventive Psychology*, *12*, 27-30.

Kelley, T., & Shen, E. (1929). The statistical treatment of certain typical problems. In C. Murchison (Ed.). *The foundations of experimental psychology* (pp. 855-883). Worcester, MA US: Clark University Press.

Kendall, M. (1942). On the future of statistics. *Journal of the Royal Statistical Society, 105,* 69-80.

Kerlinger, F. (1979). *Behavioral research: A conceptual approach.* New York: Holt, Rinehart, and Winston.

Keselman, H., Algina, J., & Kowalchuk, R. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*, *54*, 1-20.

Keselman, H., Huberty, C., Lix, L., Olejnik, S., Cribbie, R., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350-386.

Kline, P. (1998). *The new psychometrics: science, psychology, and measurement.* London: Routledge.

Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington: American Psychological Association.

Kluckhohn, C., & Murray, H. (1953). Personality formation: The determinants. In C. Kluckhohn, H. Murray, and D. Schneider (Eds.). *Personality in nature, society and culture.* New York: Knopf.

Krantz, D. (1964). Conjoint measurement: The Luce-Tukey axiomatization and some extensions. *Journal of Mathematical Psychology*, *1*, 248-277.

Krantz, D., Luce, R., Suppes, P. & Tversky, A. (1971). *Foundations of Measurement,* vol. 1. New York: Academic Press.

Krauss, S. (2008). A tripartite model of idiographic research: Progressing past the concept of idiographic research as a singular entity. *Social Behavior and Personality*, *36*, 1123-1140.

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, *18*, 89-109.

Lamiell, J. (1981). Toward an idiothetic psychology of personality. *American Psychologist*, *36*, 276-289.

Lamiell, J. (1987). *The psychology of personality: An epistemological inquiry*. New York: Columbia University Press.

Lamiell, J. (1998). "Nomothetic" and "idiographic": Contrasting Windelband's understanding with contemporary usage. *Theory & Psychology, 8*, 23-38.

Lamiell, J. (2003). *Beyond individual and group differences*. Thousand Oaks, CA: Sage.

Levy, L. (1970). *Conceptions of personality.* New York: Random House.

Lilienfeld, S. (2004). Taking theoretical risks in a world of directional predictions. *Applied and Preventive Psychology*, *11*, 47-51.

Lilienfeld, S. (2007). Academic clinical psychology in the 21st century: Challenging the sacred cows. *Applied and Preventive Psychology*, *12*, 1-2.

Loftus, G. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161-171.

Luce, R. D. (1987). Measurement structures with Archimedean ordered translation groups. *Order*, *4*, 165-189.

Luce, R. D., Krantz, D. H., Suppes, P. & Tversky, A. (1990). *Foundations of Measurement,* vol. 3. San Diego, CA: Academic Press.

Luce, R. D., Steingrimsson, R., & Narens, L. (2010). Are psychophysical scales of intensities the same or different when stimuli vary on other dimensions? Theory with experiments varying loudness and pitch. *Psychological Review*, *117*, 1247-1258.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1-27.

Lykken, D. (1991). What's wrong with psychology anyway? In D. Cichetti & W. Grove (Eds.). *Thinking clearly about psychology: Essays in honor of Paul E. Meehl, Vol. 1: Matters of public interest* (pp. 3-39). Minneapolis, MN US: University of Minnesota Press.

Markus, K., & Borsboom, D. (2011). The cat came back: Evaluating arguments against psychological measurement. *Theory and Psychology,* (in press).

Max, L., & Onghena, P. (1999). Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research*, *42*, 261-270.

Meehl, P. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103-115.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834.

Meehl, P. (1991). Why summaries of research on psychological theories are often uninterpretable. *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 13-59). Hillsdale, NJ England: Lawrence Erlbaum Associates.

Meehl, P. (1992). Factors and taxa, traits and types, differences in degree and differences in kind. *Journal of Personality*, *60*, 117-174.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. Harlow, S. Mulaik & J. Steiger (eds). *What if There Were no Significance Tests?* (pp 65-115). Hillsdale, NJ England: Lawrence Erlbaum Associates.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355-383.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. New York: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*, 639-667.

Michell, J. (2008). Conjoint measurement and the Rasch paradox: A response to Kyngdon. *Theory & Psychology*, *18*, 119-124.

Miller, G. (2004). Another quasi-30 years of slow progress. *Applied and Preventive Psychology*, *11*, 61-64.

Mischel, W. (1968). *Personality and assessment*. Hoboken: John Wiley & Sons Inc.

Mischel, W. (1983). Alternatives in the pursuit of the predictability and consistency of persons: Stable data yield unstable interpretations. *Journal of Personality*, *51*, 578-604.

Molenaar, P. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 201-218.

Mulaik, S., Raju, N., & Harshman, R. (1997). There is a time and a place for significance testing. In L. Harlow, S. Mulaik & J. Steiger (eds). *What if There Were no*

Significance Tests? (pp 65-115). Hillsdale, NJ England: Lawrence Erlbaum Associates.

Munsterberg, H. (1994). Psychology and history. *Psychological Review*, *101*, 230-236. (Original work published 1899.)

Murray, H. (1938) *Explorations in personality.* New York: Oxford University Press.

Narens, L., & Luce, R. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*(2), 166-180.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*, 97-131.

Neyman, J., & Pearson, E.S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika*, *20A*, 175-240.

Neyman, J., & Pearson, E.S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika*, *20A*, 263-294.

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, *231,* 289-337.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301.

Nunnally, J. (1967). *Psychometric theory*. New York: McGraw-Hill.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* New York: Wiley.

Pallant, J., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, *46*, 1-18.

Parker, R., & Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. *Behavior Therapy*, *38*, 95-105.

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst*, *22*, 109-116.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reichardt, C., & Gollob, H. (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychological Methods*, *4*, 117-128.

Rodgers, J. (2010). The epistemology of mathematical and statistical modelling: A quiet methodological revolution. *American Psychologist*, *65,* 1-12.

Rorer, L. (1991). Some myths of science in psychology. In D. Cichetti & W. Grove (Eds.). *Thinking clearly about psychology: Essays in honor of Paul E. Meehl, Vol. 1: Matters of public interest* (pp. 3-39). Minneapolis, MN US: University of Minnesota Press.

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284.

Royall, R. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician*, *40*, 313-315.

Rozeboom, W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416-428.

Rozeboom, W. (1997). Good science is abductive, not hypothetico-deductive. In L. Harlow, S. Mulaik & J. Steiger (eds). *What if There Were no Significance Tests?* (pp 65-115). Hillsdale, NJ England: Lawrence Erlbaum Associates.

Rucci, A., & Tweney, R. (1980). Analysis of variance and the 'second discipline' of scientific psychology: A historical account. *Psychological Bulletin*, *87*, 166-184.

Runyan, W. (1983). Idiographic goals and methods in the study of lives. *Journal of Personality*, *51*, 413.

Salvatore, S., & Valsiner, J. (2010). Between the general and the unique: Overcoming the nomothetic versus idiographic opposition. *Theory & Psychology*, *20*, 817-833.

Saville, B. (2008). Single-subject designs. In F. Buskist, & F. Davis (eds). *21st Century Psychology: A reference handbook* (pp 80-92). London: Sage.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.

Schmidt, F., & Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S.

Mulaik & J. Steiger (eds). *What if There Were no Significance Tests?* (pp 38-64). Hillsdale, NJ England: Lawrence Erlbaum Associates.

Schonemann, P. (1994). Measurement: The reasonable ineffectiveness of mathematics in the social sciences. In I. Borg & P. Mohler (Eds.). *Trends and Perspectives in Empirical Social Research* (pp 149-157). UK: Wiley.

Sedlmeier, P., & Gigerenzer, G. (1992). Do studies of statistical power have an effect on the power of studies? In A. E. Kazdin, A. E. Kazdin (Eds.), *Methodological issues & strategies in clinical research* (pp. 389-406). Washington: American Psychological Association.

Sellke, T., Bayarri, M., & Berger, J. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, *55*, 62-71.

Shapiro, M. (1966). The single-case in clinical-psychological research. *Journal of General Psychology*, *74*, 3-23.

Shoda, Y. (1999). A unified framework for the study of behavioral consistency: Bridging person x situation interaction and the consistency paradox. *European Journal of Personality*, *13*, 361-387.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. Boston: Authors Cooperative.

Simonton, D. (2004). *Creativity in science: Chance, logic, genius, and zeitgeist*. New York: Cambridge University Press.

Sink, C., & Stroh, H. (2006). Practical significance: The use of effect sizes in school counseling research. *Professional School Counseling, 9,* 401-411.

Skinner, B. F. (1966). What is the experimental analysis of behavior?. *Journal of the Experimental Analysis of Behavior*, 9, 213-218.

Smedslund, J. (2009). The mismatch between current research methods and the nature of psychological phenomena. *Theory & Psychology*, *19,* 778-794.

Stanovich, K. (2004). *How to think straight about psychology* (6th ed.). Boston: Allyn & Bacon.

Steiger, J. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173-180.

Stevens, S. S. (1935a). The operational definition of psychological terms. *Psychological Review, 42*, 517-527.

Stevens, S. S. (1935b). The operational basis of psychology. *American Journal of Psychology, 47*, 323-330.

Stevens, S. S. (1936a). Psychology: The propaedeutic science. *Philosophy of Science, 3*, 90-103.

Stevens, S. S. (1936b). A scale for the measurement of a psychological magnitude: Loudness. *Psychological Review, 43*, 405-416.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103***,** 667-680.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds.). *Measurement: Definitions and Theories* (pp. 18-63). New York: Wiley.

Stevens, S. S. (1967). Measurement. In J. R. Newman (Ed.). *The Harper Encyclopedia of Science* (pp. 733-734). New York: Harper & Row.

Stigler, S. (1992). A historical view of statistical concepts in psychology and educational research. *American Journal of Education*, *101*, 60-70.

Suppe, F. (1977). *The structure of scientific theories.* Urbana: University of Illinois Press.

Suppes, P., & Zinnes, J. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush & E. Galanter (Eds.). *Handbook of Mathematical Psychology,* vol. 1 (pp. 1-76). New York: Wiley.

Tennant, A., & Conaghan, P. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism*, *57*, 1358-1362.

Thorndike, E. L. (1904). *An Introduction to the Theory of Mental and Social Measurements.* New York: Science Press.

Thorndike, E. L. (1911). *Animal intelligence*. New York: Macmillan.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. In G. M. Whipple (Ed.), *Seventeenth Yearbook of the National Society for the Study of Education,* vol. 2 (pp. 16-24).

Thorndike, E. L. (1924). Measurement of intelligence. *Psychological Review*, *31*, 219-252.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, *19*, 579-599.

Tryon, W. (1998). The inscrutable null hypothesis. *American Psychologist*, *53*, 796.

Tuerlinckx, F. (2004). The idiographic approach: Where do we come and where do we go? *Measurement: Interdisciplinary Research and Perspectives*, *2*, 240-243.

Tukey, J. (1962). The future of data analysis. *Annals of Mathematical Statistics, 33*, 1-67.

Tukey, J. (1969). Analyzing data: Santification or detective work? *American Psychologist*, *24*, 83-91.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105-110.

Valsiner, J. (2006). Dangerous curves in knowledge construction within psychology: Fragmentation of methodology. *Theory & Psychology*, *16*, 597-612.

Vansteelandt K., & Van Mechelen, I. (2004). The personality triad in balance: Multidimensional individual differences in situation-behavior profiles. *Journal of Research in Personality*, *38*, 367-393.

Vul. E., Harris, C., Winkielman, P., Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274-290.

Wachtel, P. (1980). Investigation and its discontents: Some constraints on progress in psychological research. *American Psychologist*, *35*, 399-408.

Wakefield, J. (2007). Why psychology needs conceptual analysts: Wachtel's "discontents" revisited. *Applied and Preventive Psychology*, *12*, 39-43.

Walker, E., & Mittal, V. (2007). External and inherent constraints on progress in psychology: Reflections on Paul Wachtel's observations. *Applied and Preventive Psychology*, *12*, 44-46.

Waller, N. (2004). The fallacy of the null hypothesis in soft psychology. *Applied and Preventive Psychology*, *11*, 83-86.

Wampold, B. (2004). Sir Karl, Sir Ronald, (Sir) Paul, and the human element in the progress of soft psychology. *Applied and Preventive Psychology*, *11*, 87-89.

Webster, C., Harris, G., Rice, M., Cormier, C., & Quinsey, V. (1994). *The violence prediction scheme: Assessing dangerousness in high risk men*. Toronto, Canada: University of Toronto, Centre of Criminology.

Whipple (Ed.), *Seventeenth Yearbook of the National Society for the Study of Education,* vol. 2 (pp. 16-24). Bloomington: Public School Publishing.

Wilcox, R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology*, *51*(1), 1-39.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Windelband, W. (1998). History and natural science. *Theory & Psychology*, *8*, 5-22. (Original work published 1894.)

Wittgenstein, L. (1958). *Philosophical investigations* (G. E. M. Anscombe, Trans.). London: Basil Blackwell.

Wood, R. (1978). Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology*, *31*, 27-32.

Wright, B. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*, 33-45.

Wright, D. (2009). Ten statisticians and their impacts for psychologists. *Perspectives on Psychological Science*, *4*, 587-597.

Yates, F., & Mather, K. (1963). Ronald Aylmer Fisher. *Biographical memoirs of Fellows of the Royal Society of London*, *9*, 91-120.

Zabell, S.L. (1992). R. A. Fisher and the fiducial argument. *Statistical Science*, *7*, 369-387.

Ziliak, S., & McCloskey, D. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

Appendix 1

Holder's Axioms

A binary, greater than relation upon the levels of an attribute (symbolized by >) is

1. **Transitive** if and only if for any levels a, b, and c, if a>b and b>c, then a>c;

2. **Asymmetric** if and only if for any levels a and b, if a>b, then not (b>a); and

3. **Connected** if and only if for any levels a and b (a _= b), either a>b or b>a

(where a, b, and c are any levels of the attribute);

4. for every pair of levels a and b, one and only one of the following is true:

(i) a = b; (ii) there exists a level c such that a = b+c; (iii) there exists a level c such that b = a+c;

5. For any levels a and b, a+b>a;

6. For any levels a and b, a+b = b+a;

7. For any levels a, b, and c, a+(b+c) = (a+b)+c;

8. For any a and b, there is a c such that c = a+b.

9. for any a, there is a b such that b<a.

10. For every nonempty class of levels having an upper bound, there is a least upper bound (where for any levels a, b, and c, a+b = c if and only if c is entirely composed of discrete parts, a and b).

From Michell (1997)

Appendix 2

Axioms of Conjoint Measurement

Consider two sets of objects, V and X, where V = {t, u, v} and X = {x, y, z). These sets are disjoint as they do not share any common elements. They may be able to be measured if they relate to a third variable in certain ways. The elements of V and X can pair to form the set C. The elements of C are the ordered pairs (t, x), (t, y), (t, z), (u, x), (u, y), (u, z), (v, x), (v, y), (v, z), and hence C is the Cartesian product of V and X. C = [V x X, ≥] is a conjoint measurement empirical structure if and only if the elements of C satisfy the following axioms:

C1. *Weak order*. Given C = [V x X, ≥] and the ordered pairs (t, x) and (u, x), then V and X are weakly ordered if and only if:

- For t and u in V and $(u, x) \geq (t, x)$ then $u \geq t$.
- For X, $y \geq x$ is defined similarly.
- The relation '≥' is transitive and connected.

C2. *Independence*. The relation '≥' upon $V \times X$ is independent if and only if:

- For t and u in V and x in X then $(u, x) \geq (t, x)$ is implied for every element w in X such that $(u, w) \geq (t, w)$.
- For x and y in X and v in V then $(v, y) \geq (v, x)$ implies for every element s in V that $(s, y) \geq (s, x)$.

C3. *Double cancellation*. The relation '≥' upon $V \times X$ satisfies if and only if for every t, u and v in V and x, y and z in X then:
If $(u, x) \geq (t, y)$
and $(v, y) \geq (u, z)$
therefore $(v, x) \geq (t, z)$

C4. *Solvability*. The relation '≥' upon $V \times X$ is solvable if for any three of the four elements t and u in V and x and y in V, the fourth exists such that the inequality $(u, x) \geq (t, y)$ is solved such that $(u, x) \sim (t, y)$.

C5. *Archimedean condition*. Let there exist the elements t, u, v and s in V and x, y, z and w in X. If u - t ≤ s - v and y - x ≤ w - z, then for any natural number n, V and X are Archimedean if and only if n(u – t) ≥ s - v and n(y - x) ≥ w - z.

If all axioms C1–C5 hold, then for t and v in V, x and z in X there exist real valued functions ϕV and ϕX such that:
$(v, x) \geq (t, z) \leftrightarrow \phi V(v) + \phi X(x) \geq \phi V(t)\ \phi X(z)$

From Kyngdon (2008)