Chapter 4 The fundamentals of assessment: prime considerations

4.1 Introduction

This chapter investigates the prime considerations within intelligence assessment; namely mathematical, statistical and measurement predicates upon which is based psychological assessment and measurement and attempts to link these prime considerations with the core philosophy of dynamic assessment. Such a view entails a re-evaluation of basic premises within intelligence assessment. These prime considerations can be classed under the rubric of quantification, often referred to in the social sciences as "the quantitative imperative" (Michell, 1997, 1999; Niaz, 2005). A re-evaluation is deemed necessary even though assessment has continued unabated from its earliest inception right throughout the twentieth century and on into the twenty first century. From the outset is made explicit that the mathematical formulations, logical derivations, statistical conceptualisations and the theorems, proofs, axioms and subsidiary deductions emanating from basic and proven tenets are not in question.[1] This chapter can only be read within the context of the preceding chapters, as many arguments employed and highlighted in those chapters are equally pertinent to the issues discussed in this chapter. Looking back on basic epistemological and ontological issues and discussing in turn issues such as reductionism, relativism, the nature of intelligence, the place of assessment within psychology and psychology as a discipline within the humanities as well as studying this method of knowledge acquisition as a form of a greater whole entitled "science"; illustrates how this all converges on assessment and where it is currently situated today.

4.1.1 The quantitative imperative[2]

Understanding assessment thus necessitates a re-look at its prime considerations, its foundations, the nascent fertile grounds from which it sprung and grew into what it currently is; a noble and at times misguided effort at assigning numerals and applying variegated statistical techniques to what are purported to be reified constructs in the hope of fulfilling a utopian ideal of equality and success. It has had a chequered past and has often fallen far short of these ideals which, at times, have yielded precisely the opposite results. As discussed in chapter 3, the growth of psychology as a discipline occurred in varied contexts since its inception as formally credited science. Having assumed the mantle of robust firmly entrenched natural science rigour and method as well as assuming hermeneutic, humanistic tendencies emphasising the notion of cultural relativism; pointing out the deficits of the hypothetico-deductive method of analysis and growth and seeking to amalgamate in some sort of consilient re-dress of the whole plethora of what it means to gather knowledge and what it means to "know"; the path towards psychology-as-science is indeed circuitous. As part of this same journey, is the questioning of psychology's mathematical, statistical and measurement past to which attention is now turned. Psychology's eclectic array of research efforts and individuals has been cited as a main reason as to why the quantitative imperative has been incorrectly aligned with positivism (Michell, 2003) even though naïve positivist methodologies have been cited as one among many reasons as to psychometrics' less than sparkling reputation among some work forces at certain periods in time (Sehlapelo & Terre Blanche, 1996). Carnap was influential in putting forth his ideas within frameworks, stating that no framework could be judged as right or wrong, because it was not an assertion. The need to move science along in terms of bettering the whole enterprise necessitates various methods which seek to do just this, as long as the framework is useful. Stevens'(1946) scales have been incorrectly assumed to have emanated from a fully fledged and developed methodological framework. There was no fully developed framework but rather an attempt at a solution (Michell, 2003).

The goal of this chapter is not in anyway meant to construe the great edifice of mathematics, statistics and measurement as ill-founded and ill-used within the psychological arena. On the contrary, without much of what the afore-mentioned has offered psychology during it's history as recognised subject of inquiry,[3] a considerable amount of research would never have seen the light of day. What is contested though in this chapter is how these methods of inquiry are used within assessment and it will be argued that through thoughtful consideration of what in fact forms the foundation of assessment, dynamic assessment will be offered a place within the broader arena of assessment conducive to its own development. That dynamic assessment will replace any mainstream assessment is of course absurd and it's main mission has always been one of concurrent usage throughout its history alongside conventional assessment. The move away from observable entities (whatever might be construed as an entity, let for instance, a test score be such an example), to statistical abstractions or artefacts somewhat spurned early psychologists' work towards the infusion of statistical methods in order to aid in theory-building, thus moving

[1] The author is neither a logician nor a mathematician and as such will not even attempt to understand the finer grained truths emanating from such studies. However, the author can also not ignore the influence these aspects of study have had in the history of assessment and psychometrics.
[2] Michell (1999, 2003b). From the seventeenth to nineteenth century's distortion of the Pythagorean philosophy of measurement (Barrett, 2005).
[3] The author hesitates to use the word "science" here as a backlash of criticism is expected; although the use of the word as applicable in this context can be defended.

further away from what Michell (1999) refers to as "foundational issues of quantification" (p.104). Issues of inference and problematic measurement resulted in psychologists' increasing tendency to rely on statisticians' models bringing into question the validity and acceptability of the quantification of psychological constructs much later on. This situation now finds itself once again in a position of debate within the origins and development of dynamic assessment (and dynamic assessment's resultant reluctance to make decisions on static one-time assessment scores alone; Wiedl, Guthke & Wingenfeld, 1995). In order to systematise this inquiry into the prime considerations of the fundamentals of assessment the chapter is divided into four subsidiary sections, namely the;

- mathematical foundation
- statistical foundation
- measurement foundation and
- psychological assessment foundation

These considerations will be critically discussed from both historical and intelligence assessment points of view and the role they have played in the formation and continued use of assessment instruments across the globe. A main hypothesis is that dynamic assessment has not quite found its place within the broader intelligence assessment framework due to the misunderstandings of what in fact is meant by "measurement" and when assessed in this light, it becomes obvious that dynamic assessment is fundamentally, philosophically and psychologically a theory not aligned with traditional manners of assessment. This situation is made even more intractable when one considers the almost pure utilitarian value of many intelligence assessment tests (Barrett, 1998). In this instance, Barrett (1998) fervently states that intelligence as a trait construct has much pragmatic value but little causal theoretical backing (and hence little scientific value), mostly due to the enterprise of poorly thought out measurement. Figure 51 depicts the roles played by each fundamental level within assessment and how each level interfaces with every other level. The diagram illustrates the multi-directional flow of how measurement levels being predicated on a statistical foundation, itself predicated on a mathematical foundation, need to inform one another. Problems emanating at the measurement level as is often the case within behavioural assessment must be traced back to its statistical and mathematical foundations if any solution is to be found. Solving issues at only one level will not bolster the cause of re-looking assessment within psychology for instance. This is not to say that any level cannot function on its own, that is counter-intuitive and incorrect as evidenced in the argument that statistical manipulation cares nothing for the level of measurement those numbers are levelled at, "even if the numbers are the purest nonsense, having no relation to real magnitudes or the properties of real things, the answers are still right as *numbers*" (Hays, 1981 in Maxwell & Delaney, 1985). Statistics reflect the numbers not the constructs yet numbers indicate where we are, what we do, how much of it there is and is imperative to science, as science would be "impossible without an evolving network of stable measures" (Wright, 1997b, p.33). The job of adequate representation is the psychologist's job. The chapter concludes with suggestions as how best to realign the basic fundamentals within the theoretical framework of dynamic assessment within intelligence. The foundations as illustrated here are not to be viewed as subordinate to any of the remaining two levels, (this is not an hierarchical arrangement) as the statistical foundation might well proceed from the measures taken before statistical manipulation can proceed. This diagram merely serves to position the three realms of entities or foundations upon and from which psychological assessment "results" or scores derive. As Eves and Newsom (1965) state, a particular philosophy can be equated to a process of refinement and ordering of experiences and values and in so doing find relations which are normally considered disparate and find differences between things which are normally considered the same. Hence, a philosophy is essentially a description of a theory concerning the nature of something. Lazersfeld (1977) posits four main reasons why quantification (or as Pawlowski, 1980 refers to it, "quantophrenia") was becoming an increasingly important part of the social sphere as early as the seventeenth century:

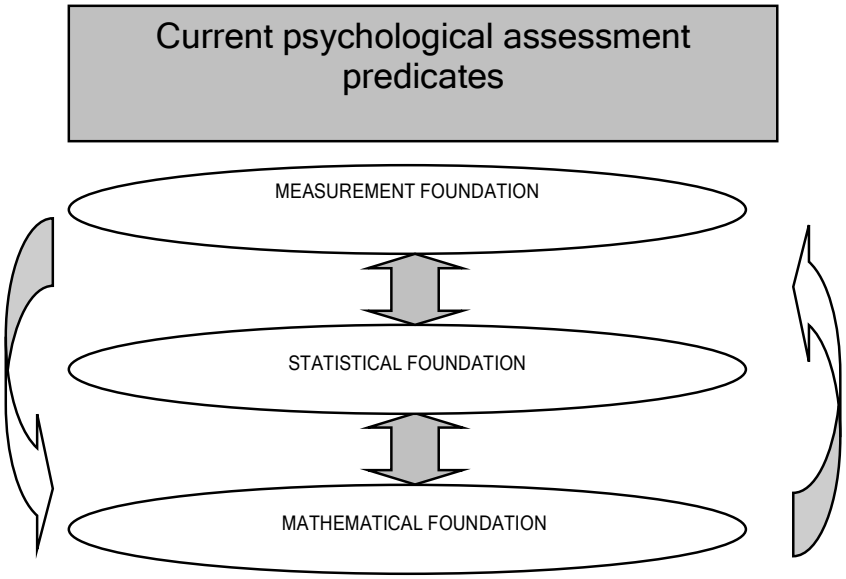
i. the rise of capitalism
ii. the prevailing Baconian spirit (see chapter 3)
iii. a pressure to derive accuracy similar to that of the natural science in social endeavours and
iv. the increasing role of public administration and planning due to bourgeoning population sizes – much to do with insurance and the role of money and taxation

Ramul (1963) furthermore adds that psychological quantification, measurement and ratings were being practised well before such physicalist notions of measurement were discussed within the psychological domain and cites early usage of ratings and measures within areas such statistics, vision, memory, attention and thought (in which the "velocity of thought" was already being pondered as early as 1750, a prescient notion of the speed of neural conductivity perhaps?). Most measures of psychological import go back only as far as the start of the eighteenth century and of such measures carried out, only a few were

conducted by persons considered psychologists.[4] The need for quantification in psychology proceeds along the following very narrow rationale (Schönemann, 1994):

- science is defined by its quantification via concatenation of its constructs
- any discipline wishing to call itself a science must adhere to this principle
- only if this is so can the discipline be called a science
- psychology does not posses constructs which are quantifiable via concatenation of these constructs, hence
- psychology is not a science

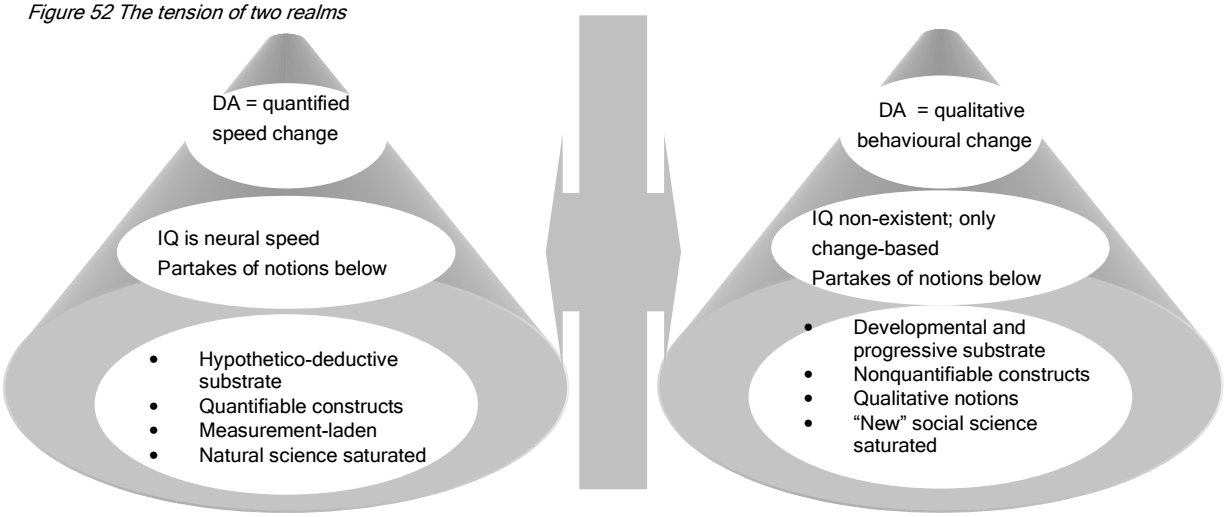
Figure 51 Assessment predicates



Measurement rhetoric has done a grave disservice to the psychological assessment enterprise and no manner of sophistication as exemplified by statistics and mathematical modelling can ever hope to rectify a problem which is clearly insoluble from this approach; the problem needs to be solved from another level entirely. Currently psychology exists within many realms, each purporting scientific accuracy, reliability and validity. These terms can lose their meanings very easily if misapplied in a variety of contexts. Psychological assessment should either reside in the domain of quantitative hypothetico-deductive development or it should reside in the nonquantifiable realm and progress in its own manner akin to the progress evidenced in the natural sciences. Neither realm is "right" or "wrong". To talk of the correctness of such realms is tantamount to misdirected and misinformed notions of what it means to conduct enquiries via truth-seeking mechanisms. Dynamic assessment lacks a fit or is at the very least a poor fit in the measurement models adhered to throughout the history of mental test theory. It is possible that it has found itself located in the wrong realm, which can hardly be the fault of dynamic assessment researchers and practitioners as they are merely fitting in with current mainstream concepts of what it means to practice psychological assessment. Figure 52 illustrates the tension of two realms.

[4] Ramul probably means that, although they were not formally designated as psychologists, they could for all intents and purposes be considered as such.

Figure 52 The tension of two realms



Dynamic assessment is currently placed in both realms, thus resulting in tension which has yet to be resolved. It must choose for itself a realm in which to lodge and grow. However, this is not only a fault of this sub-discipline but a pervasive trend within the whole of the psychological discipline. Michell's (2001) pointed criticism is levelled at the misguided efforts touted by the social science measurement effort which he envisions as "instances of the scientific method applied to psychology" (p.211). The manner in which psychometrics is taught, he says, subverts the scientific method. Once again, it is necessary to reiterate that the tools of a trade are not necessarily at fault; it is the incorrect tools which are being applied which is very much at fault. The difference inherent in scientifically experimenting *a priori* and instrumentally going about the practical work of extracting scientific concepts are the two tasks of quantification (Michell, 1997). Yet utilising the instruments before one has worked out the scientific basis for measures is "pretence of science" (p.359). There is thus a call for a change of tools-to-trade within this thesis which will hopefully go some way in alleviating the current misfit of appliance and trade. Utilising psychometric tests to test psychological constructs is not a proven mechanism and remains at best hypothetical (Michell, 2001). Rumblings of the soundness of psychometric measurement can be traced back to the first quarter of the twentieth century and is thus hardly a new concern for psychometrists (Maraun, 1998; Stevens, 1946). However, due to "big business" psychological measurement and the seeming lack of mathematically trained psychologists, the utilisation of measurement and the ever increasing sophistication of statistical techniques (Barrett, 2002, 2003; Blinkhorn, 1997) the characteristic lack of enthusiasm for test use by psychologists is evident (Maraun, 1998; Michell, 2005). The following citation partly sums up the essence of this chapter.

> "The post-Second World War methodological consensus in psychology combined a variety of elements thought to be necessary for scientific rigor, such as null hypothesis significance testing, [see below] Fisher's work on experimental design and analysis, [see below] and classical test theory [see below]. This consensus occurred at a crucial time in the history of psychology. Patterns of funding for universities and research were undergoing unprecedented changes in the USA and the effect was to set in concrete a methodological consensus that owed more to the values of window-dressing than to any values implicit in logical positivism" (Michell, 2003, p.16).

The issue of meaning-ladeness and measurement of a construct is one suffused with confusion (Barett, 2001). Utilising as measured construct is tautologous, as psychometricians assign to the supposedly quantifiable construct "intelligence" and then seek to measure . Upon locating it along a continuum of "less-to-more" is upheld as existing (Maraun, 1998) which is clearly absurd. From this point onwards, the robust and sound statistical techniques used become ever more detailed and inherently presumptuous in terms of manipulating in manners which falsely bespeak of its existence. Constructs are identified *a priori* as existing, but as to whether they do or not is another philosophical question altogether. Added to this is another erroneous Pythagorean assumption that all attributes are quantifiable (Barrett, 2005; Michell, 1999, 2003). Assuming that measures intelligence in some manner, this will need to be done within a context of rule-bound associations which are themselves products of human behaviour. Rules are not empirical facts or findings, but a constituted set of instructions to follow (Maraun, 1998).[5] Having identified supposedly quantifiable constructs, techniques further constrain interpretations of findings in such a manner as to lead to the acknowledgement of a methodological artefact and nothing more (Barrett, 1998). This turns the artefact into instantiated fact which is a leap not always scientifically condoned. What occurs here, states Michell (1997), is a gross

[5] As Barrett (2001) states " the IQ variable is created - which is constructed ad hoc from one or more of the constituent technical concepts. IQ is taken as a 'measure' of 'intelligence' by some, but not by others. This is because the meaning of IQ is conflated with specific technical concepts and then mapped onto an arbitrary concept of 'intelligence'. What's worse, the biometrical geneticists then try and use IQ as a proxy for intelligence in order to 'find' genes for IQ test scores" (p.37).

instance of "thought disorder" which is defined as blindly following on in a tradition of delusion of simply not acquainting oneself with methodological concerns as it pertains to psychological measurement. This, states Michell (2001), subverts the scientific enterprise and is an instance of psychometrics playing the role of a diseased method (pathology of science). Notwithstanding these strictures, there is also the very important point of having no identifiable and workable common unit or metric according to which to measure the purported construct (Barrett, 1998; Kline, 1998). Measurement is theoretically assured if, as Wright (1997) maintains, the following is adhered to:

- to measure is to infer (which is precisely the leap made within assessment, otherwise there would be no point)
- measures are obtained by stochastic approximations
- measures are one-dimensional
- measures are counted in abstract units which are of fixed sizes (note that the abstractions are of fixed size which need not necessarily indicate that the constructs being measured are of fixed size – for more see below)
- measurement results are not influenced by outside influences

Clearly, the path followed by the natural sciences works for the natural sciences and it could work for varying sub-disciplines within psychology, only if stringent rules are followed and implemented. There are areas amenable to such treatment and there are areas which are not amenable to such treatment (Barrett, 1998; Borsboom, 2005; Michell, 1997). It is the job of psychologists to piece together their discipline and sort out these issues and to determine the fit between theory and data or nomological network (Strauman, 2001). It is clear that psychology's challenge is to account for its scientific status in a manner unlike that of the natural sciences (neither observability nor levels of analyses will suffice as method of knowledge-gathering as psychologists work with inference from observation and theorising; Strauman, 2001). Maraun (1998) follows Wittgenstein's arguments in terms of rule-based measurement and argues against what he considers a conflation of conceptual and empirical issues when rendering the measurement issue within the context of construct validation theory, as espoused by Cronbach and Meehl (1955). Measures, states Maraun (1998), and how we measure are not empirical issues but logico-grammatical issues and attempts at solving conceptual issues via rule-based measurements (predicated on human judgment as to what constitutes these rules) cannot be equated. Cronbach and Meehl's (1955) predicates for construct validity are based on a nomological net of concepts derived from this particular philosophy of science and is applicable more so to the natural sciences than to the social sciences. The concept of validity within the natural sciences is redundant (Kline, 1998) as concepts are public domain (more obviously manifest) and detecting error in what is purportedly measured is far easier within the natural science domain (which is not to say that all constructs within the natural sciences are amenable to immediate comprehension). Laws (statistical or deterministic) are set forth (akin to rule-following) according to which observable properties are somehow related or observables are related to theoretical constructs or different theoretical constructs are related to one another. Empirical discovery and construct validation are not synonymous and the more of one does not imply one obtains more of the other. For instance poor scores on a maths test at school does not necessarily indicate lack of proficiency in maths years later. The score changes over time but the construct of underlying maths intelligence has not changed (at least this is the perception within mainstream assessment). Regarding intelligence, Bardis (1985) states "discovering a functional unity by means of correlation has nothing to do with inventing a faculty and attaching a label to it" (p.219). The limit of substantive utility has been reached and according to Barrett (1998), the purely methodological approach no longer suffices as means of measurement.

Think of *g* and one immediately knows of much research attesting to its (at least) statistical existence but substantive psychological theory has yet to account for it, not to mention trying to explain what it is in a psychological sense (Kyllonen, 1996). Barrett (2005) asks the pertinent question of whether modern psychometrics, which is now so dominated by psychological statistics, has lost its way in terms of concentrating on the substantive thesis underlying the statistical thesis. Recall Meehl's (1967) lament on the lack of fit between theoretical and substantive theories in terms of null hypothesis significance testing. A visual aid might assist in more fully understanding why this is so. Figure 53 illustrates the nebulous area occupied by a construct entitled "intelligence" and the well defined area encompassing a small section is the partially identifiable construct validation. Figure 54, adapted from Oberhauer (2005) whose model is delineated for working memory as construct can easily serve as generic model in this instance. His model illustrates succinctly the circular movement involved in theory-measurement-data-theory as process towards construct definition and validation.

Figure 53 Empirical construct validation and meaning-ladeness of true construct

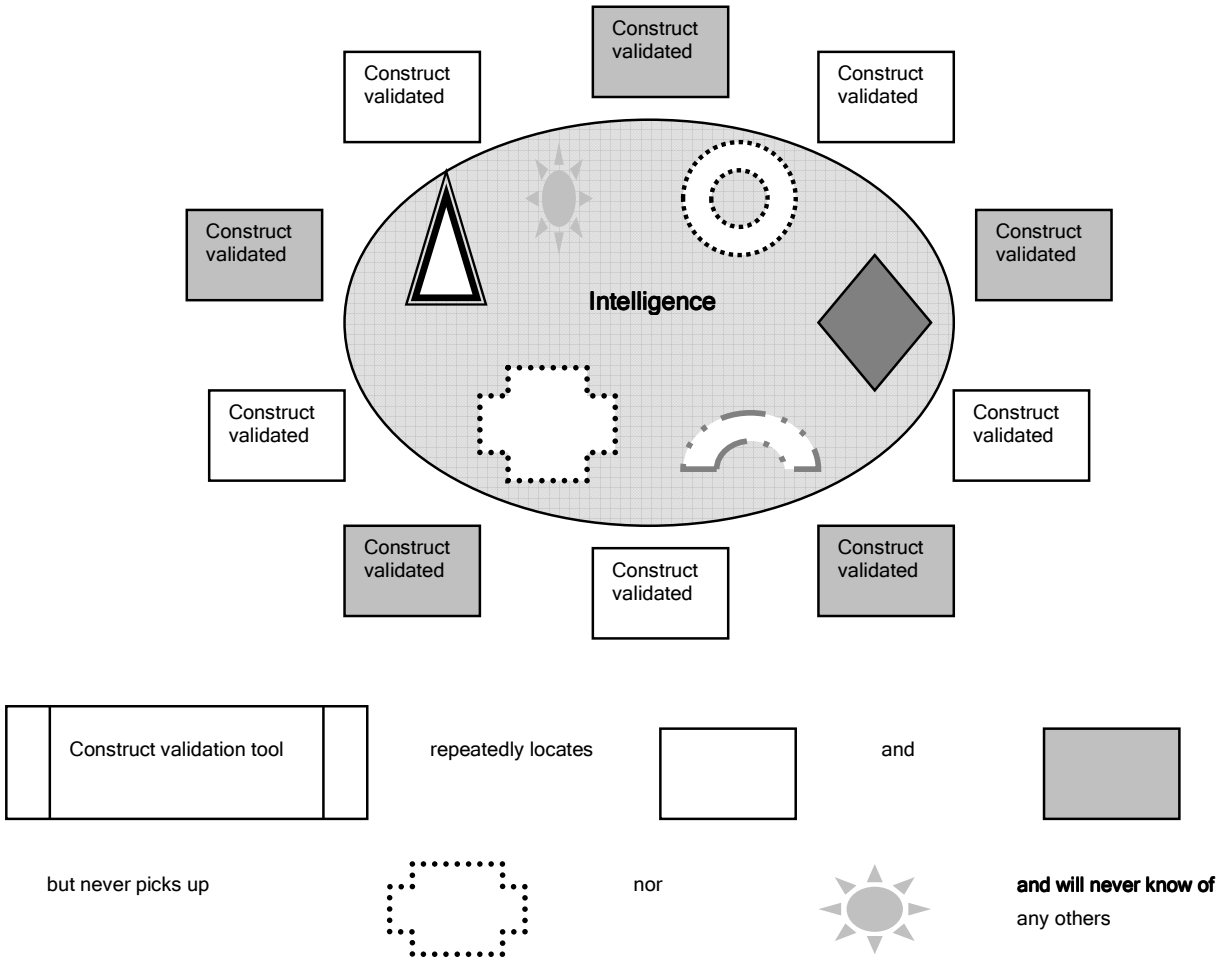
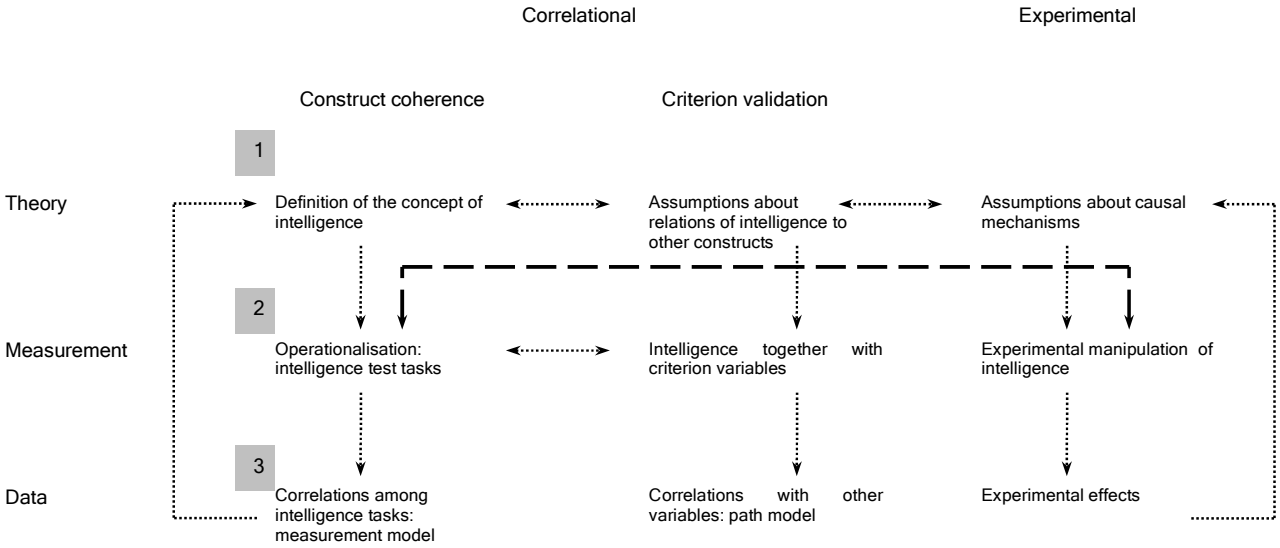


Figure 54 The vacillation between theory, measurement and construct validation (adapted from Oberhauer, 2005, p.394)



This figure is of course hypothetical and merely a thought experiment. The construct of "intelligence" is validated a number of times purely because the exercise happens to locate a number of similar constructs housed within the concept. The tools of the trade are precise enough to be able to locate a number of these correlating constructs and are thus able to conclude that the entire construct is possibly one large square shape. The reality is strikingly different! The areas "picked" up by tools of validation are indeed there but make up only a small percentage of the actual existing construct, which for the mean time, is beyond the reach of the technique. Probable models themselves are also merely tools for digesting information content which add little more in the way of credence to a latent attribute, assuming that such a latent attribute exists in the form we speculate about (Barrett, 2001). Building models predicated on constructs which have no doubt been validated as the above figure illustrates is precarious because a number of assumptions are being made, among others;

- that the construct in fact exists (it is possible that it likely exists in some form or another)**
- any tool utilised to search for the construct is purpose-built (with constructs being perpetually inferred; Utley, Haywood & Masters, 1992). This is tautologous as it is speculated that σ exists, a technique is brought in to search for σ which was custom built to find σ. Given this logic, it will in all likelihood find σ.[6] *A priori* considerations will lead a technique to search for an *a priori* concern (Williams, Zimmerman, Zumbo & Ross, 2003)
- the nature of σ is housed within the context of science progression itself lodged on the bedrock of any variety of philosophy of science schools
- searching for σ is loaded before we ever start the investigation
- σ is found and conclusions are drawn in favour of its empirical existence
- σ is rarely noted for being an extension of science practice dictated to by the above considerations

[6] Spearman having designed factor analysis to locate for his theory of intelligence a common factor of intelligence also went about developing classical test theory (Maraun, 1998; Williams, Zimmerman, Zumbo & Ross, 2003). This can hardly be considered a co-incidence. His endeavours never seemed to question the very utility of quantifying the very constructs he sought to measure (Michell, 1997).

- σ is defined according to the construct definition used. Any construct is as equally valid as any other provided one has followed the strict tenets as laid down by whatever model or school one happens to endorse. This unfortunately still tells us nothing about the actual underlying nature of the construct (Barrett, 2001).

** if σ exists, it may exist in the form which is amenable to extraction via techniques employed to extract this particular meaning from it.

In such a case, neural conductivity might well function as intelligence correlate. The correct tools should be employed to determine this. Tools are not to be borrowed from domains from which they were clearly not designed. Psychometric measures and physiological measures are bound to correlate but how sure can we be that this is not another instance of Meehl's (1990, 1997) crud factor? Measurement of characteristics follows directly on from definitions of both measurement and the characteristic at hand (Maraun, 1998). The logic of quantification in the social sciences usually runs as follows (and is perfectly acceptable practice bar one is in the correct domain in which one can apply such strategies); the first level of quantification would be one of assigning constructs numerics, in other words the metrification of constructs which proceeds with summarising statistics of these numerical counts and ending with the mathematisation of supposed theoretical entities (Meehl, 1998). As has been argued above, the logic flowing from this argument is sound enough given the correct circumstances in which to practice such dealings. However, the first premise is flawed if one considers that numerosity of entities supposedly existing is both right or wrong depending on how one views the situation. Numerical assignation as discussed in 4.2.2 below relates to this discussion on measurement, hence the need to include in this larger debate the mathematical foundation of measurement. Ross' (1964) delineation of the formal analytical view of theory constituents (section 3.11.1 in chapter 3) is similar in nature to his numerical assignation discussion pertinent to this chapter. Depending very much on the nature of a strict one-to-one isomorphic[7] relation between hypothesised constructs and their numerical counterparts, some schemes are more amenable to numerical assignation than others and this assumes a common unit across the discipline thus making it axiomatic (see section 4.2 on mathematical foundation below) (Barett, 2000). Most often, as is the case within psychology, the question as to whether one can even assign numerals to nonquantifiable traits is never even asked (Barrett, 2003). Hypothetico-deductive means of investigating nonquantifiable aspects exist and have existed for a considerable time (such as facet theory within intelligence research and cellular automata; Süß & Beauducel, 2005). The question then is, why such methods cannot be employed within psychometrics (Barrett, 2003).

Figure 55 illustrates Ross' (1964) paralleling of formal physical systems and number systems. Changes in the one system do not necessarily mean that similar changes will manifest in the other system. This is so due to the formal properties of both systems which can differ in ways not understood even though translation rules have been put in place. Figure 54 needs to be understood in tandem with figure 52 as well as figure 55 which is the more modern rendition of this tension between the physical-to-measurement concern. Figure 56 illustrates clearly the divide spanning the construct *g* which is an hypothesised notion only, although supported on numerous occasions, it is supported via a nomological network of mechanisms which were purpose-built to find *g*. Scores on measures of tests proposing to test subject matter exists as such. However, the link between the measure of *g* (as defined by and searched for by techniques developed to do just this) and scores on subject matter tests is highly contentious. Is there a mismatch between ability as predictor and task performance as criterion (Ackerman, 2005)? Chapter 2 detailed the author's preference for the notion of *g* and all that it entails philosophically. This is upheld as the position taken. There is no contradiction inherent here. *G*, if it exists, can most likely be probed via methods and tools amenable to such probing. Currently the wrong tools are being used. The deployment of theoretical models, schemes and other myriad conceptual frameworks of how the brain functions during intelligence tests are premised on statistical findings which are derived from mathematical artefacts themselves products of underlying mathematical models which guide the interpretation of psychometric test findings.

The assumption underlying this path of inference is that there seems to be a one-to-one correspondence or mapping of theoretical mathematical entity and actual brain functioning and processing. Granted that sophisticated neurological and less subtle invasive techniques were not yet developed during the heady days of Pearson's statistical development and analysis of data, a mathematical tribute and contribution to the measurement of human abilities was conducted in part by Galton and all their successive followers within the same tradition (Nunnally, 1978). Notwithstanding the great leaps forward in intelligence research the role of mathematical modelling to fit data at a behavioural level is perhaps currently overstated or perhaps slightly outmoded. Mapping local brain functioning on a one-to-one basis with models developed from this type of data would seem to be more parsimonious than what has gone before; advocacy of parsimony is not an underlying assumption within the preceding statement, as it is erroneous to think that in all spheres both behavioural and social sciences will indefinitely progress in a parsimonious fashion. Cognisance is taken of the fact that by the mere introduction of any one measuring technique immediately

[7] Isomorphism is more strictly speaking the relation regarded as a one-to-one mapping. Homomorphic mapping includes other elements not mapped onto the construct (Michell, 1999). Homomorphic measurements include various objects than can be assigned the same number.

delimits the area of investigation in terms of what can and cannot be observed and/or derived. No one single technique known currently to either sphere of natural or social sciences can indeed derive all aspects of a function or structure.[8] There are however methods and models available today which come closer to doing just this as opposed to the continous utilisation of some assumptions inherent within models followed from the early days of twentieth century psychophysics

Figure 55 Measurement theory and theory of positive real numbers (Ross, 1964, p.60)

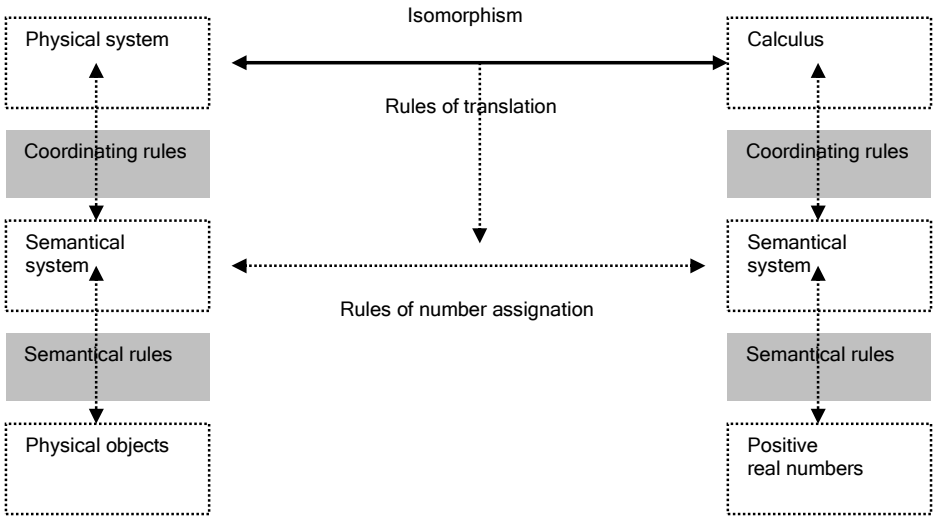
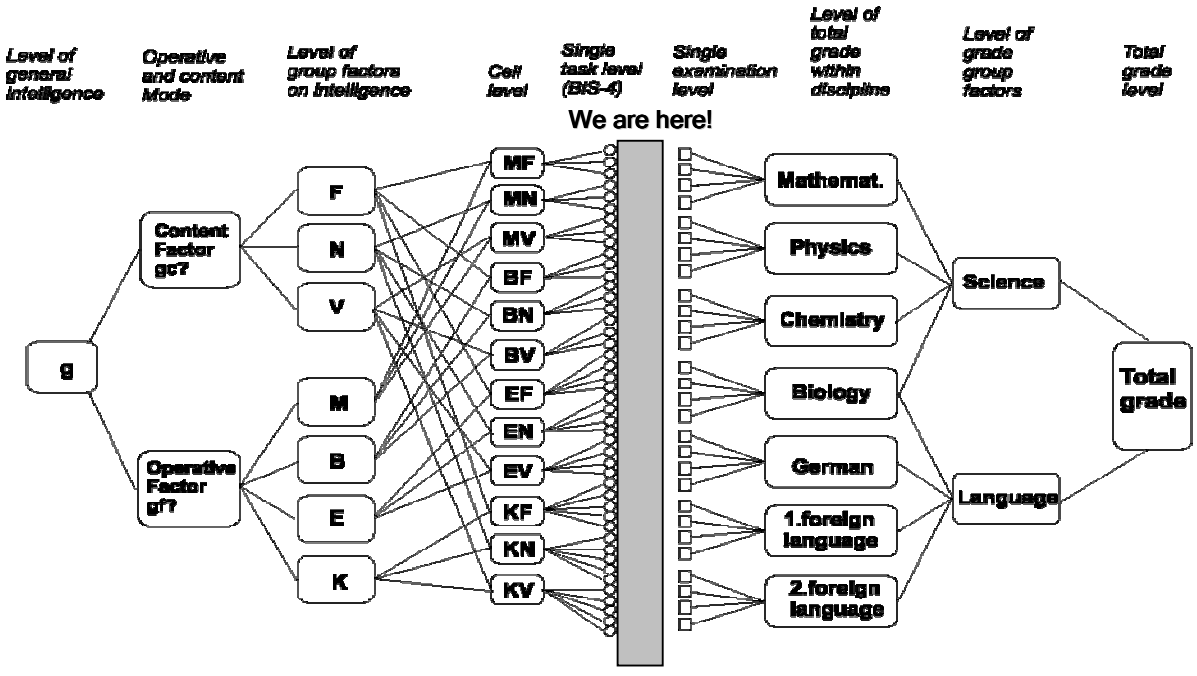


Figure 56 Barett's (2000, p.41) utilisation of Wittman's (1997) illustration



[8] Although this remains foreseeable within the ambit of scientists to rectify and it is believed that this can occur. Time will tell. What may initially look doubtful may well in time become mainstream practise.

4.2 Mathematical foundation

Prelude

Much of early psychological work was conducted within the confines of calculability and predicated on pillars of mathematical and statistical solidity.[9] It thus assumed great prominence within the broader field of natural philosophy, especially with the pioneering works within psychophysics, decision-theory, game theory, learning and information processing and fields dominated by the propositions of mathematical psychology (Coombs, Dawes & Tversky, 1970; Dalmedico, 1997; Estes, 1975; Luce, 1997, 1999; Ratcliff, 1998; Reber & Reber, 2001). It is within the concept of science that mathematics originated (Brouwer, 1983a; Russell, 1983) and functioned as "the handmaiden of scientific enquiry" in the nineteenth century (Maddox, 1998, p.9). As such, a link between chapters 3 and 4 is forged. However, the area of mathematics was not always conceived of as abstract and academic. Until the seventeenth century, it was considered as little more than a mechanical trade serving only to aid other trades (Uglow, 2003). This is very much the same historical path followed by measurement the history of which is not traced as originating with mathematics but within trade and construction (Wright, 1997). Mathematical models have been employed within psychological research since the mid-nineteenth century and have continued unabated since, finding for itself particularly fertile grounds in application since the 1950's (Coombs, Dawes & Tversky, 1970). Measurement theory necessitates some sort of reality (the psychological construct for instance) and the requisite mathematics (IRT or CTT modelling) to describe the reality or a model assumed to reflect the empirical situation (Wille, 1994). Gould (1998) maintains that the fascination humans have for numerical regularity is closely tied to our propensity to dichotomise objects in nature (see chapter 3) but that in certain instances this need for numerical patterning has resulted in "our overinterpretations, run[ing] so far beyond what nature could possibly exemplify, that we can only postulate some inherent mental bias as a driving force" (p.36). The enterprise of logic and mathematics, is of course, one of the more aloof disciplines calling itself a science as it focuses in on itself more often than appealing to outside issues (Quine, 1983b).

The link between Abraham Wald's decision theory,[10] Von Neumann's game theory[11] and the early work of Neyman and Pearson (Dalmedico, 1997; Kline, 2004) frames early to mid-twentieth century mathematics, statistics, science and social science studies. The axiomatic scientific method in which one proceeds from an undefined but accepted truth can be linked to the axiomatised system developed for mathematics. Probability (Boyer & Merzbach, 1991) in turn is utilised by the social sciences and psychology in how deductive thought envelops reasoning and how logic can be utilised to rationalise arguments and "prove" certain conclusions (reference here is made to the logicist school in mathematics). Mathematician Andrey Kolmogorov maintained that probability as a mathematical discipline should be derivable from axioms just as algebra and geometry were derived from axioms (http://en.wikipedia.org/wiki/Kolmogorov). In fact he did just this; axiomatising probability and thus satisfying Hilbert's call for the axiomatisation of mathematics (Boyer, 1991; Gratzer, 2002), the discussion of which is to follow. Theories of probability have played major roles within the psychometric discipline ranging from Bernoulli, Poisson and Bayes to more modern-day renditions of probability such as those encountered within item response theories which will be discussed below. The role played by probability within the natural sciences differs however where error has often been ignored or if corrected has subsequently utilised alternative models (Borsboom, 2005). Issues of change, probability and error preoccupy social scientists - perhaps we should reconsider our position on these aspects? The connection between "truth" and "proof" is, however, a loose one (Benacerraf, 1983b). The mathematical intricacies utilised within probability models for instance are a point in case as probability theory is widely employed as manner of inference within psychological research and without such models "it would not be possible to measure the reliability of inferences and decisions" (Stigler, 1999, p.2).

Mathematics was considered the pinnacle of scientific thought during the early twentieth century (Dalmedico, 1997), yet mathematics' edifice was starting to crack as a number of fundamental issues were being grappled with, some more successfully than others (Coveney & Highfield, 1995). Logic, as with mathematics, can be divided into three main periods of development, from classic Aristotelian logic through two thousand years till the algebraic or symbolic period beginning with Boole and ending with Hilbert and culminating with the metamathematical or modern period exemplified by Gödel (Rucker, 1987). The misperception on the part of psychologists and measurement specialists within psychology as to the sound edifice of mathematics is a gross misunderstanding on their part. Not only can foundational mathematics said to be grappling with a number of its own issues, but the wholesale transference to a realm in which is not always suited is tantamount to academic

[9] Itself a questionable notion as mathematics like any other formal discipline has its fair share of detractors opting for a more relativist notion of what it means to engage in mathematical science. Ethnomathematics, for instance, is an area devoted to the understanding of mathematical tendencies throughout cultures and throughout the ages (Ernest, 1994)

[10] Any one of a suite of theories seeking to describe for a system of decision making, how in fact decisions are made. These theories include information pertaining to probabilities, mathematical approaches based on game theory and probability theory as well as more subjective aspects such as attitudes and beliefs (Reber & Reber, 2001). Wald's work, among others has led to broader acceptance of Bayesian probability (http://en.wikipedia.org/wiki/Bayesian_probability). (See below for more on Bayes).

[11] Perhaps the most commonly known instance of game theory to psychologists would be the prisoner's dilemma. Game theory seeks to describe a system for the making of moves and decisions within simple to more complex games (Reber & Reber, 2001). Game theory postulates have been applied to areas such as interpersonal interactions, economics and international affairs.

sacrilege. This has resulted in another perception, espoused by Schönemann (1994) for instance where "mathematics seems to have been singularly ineffective in the social sciences" (p.151). Measurement theory (see below) is largely predicated on axioms and like mathematics, proceeds from such, as yet, improvable givens. Aesthetically pleasing axiomatised measurement systems, says Schönemann (1994), are what we are now left with. It is worthwhile focusing attention on Schönemann's (1994) list of reasons as to why mathematics seems to be ineffective in rendering itself a tool to the social sciences:[12]

- A sound body of factual knowledge was built up within physics long before axioms were constructed as point of departure. In other words proceeding from axioms to observations is not always the method followed within some areas of physics
- Only in a very select number of physics and mathematics fields has post-hoc axiomatisation been achieved
- Some areas within physics have proved quite successful despite having no axioms with which to guide them
- Scientific progress is more often made via induction based on observation and has not proceeded from deductions from axioms
- Induction is not a formal mechanical process but takes into consideration much else besides a narrowly construed set of determinations and lastly
- Science has often accepted results which have been contrary to traditional beliefs

Mathematics cannot necessarily be construed as the equivalent of the study of numbers as is evidenced within the area of geometry for instance and to say that mathematics is only quantification would be false (Russell, 1983). However one cannot ignore the measurement link between mathematics and its import into the realm of psychological assessment. Logical and mathematical "truth" such as it may be, cannot be equated with certainty, as Mill argued, such truths are merely based on a very large number of instances from which inductive inferences can be concluded (Ayer, 1983). What may seem to be logical, given the extremely large base of data on which to formulate such logical statements, is once again subject to finitist strategies to which humans are limited and included in this is the rate of computational power increase in the future. Once again there is no proven algorithm which would state whether such a system would maintain its logicism right up to the infinite end.[13] The answer may well lie within the Platonist (theoretical) realm, as yet inaccessible. Mill's view is synonymous with the fact that mathematical certainty can only be achieved via sensory observations (Lehman, 1979) and is severely contrasted to those propounding an *a priori* approach to mathematical certainty and truth.

On the other hand, hypothesis testing and the experimental method so much adhered to within psychology is tentatively related to the levels of chance and probability of certain outcomes and traces its roots to early experimental statistics. The latter was a branch of mathematics not given as much regard as the more abstract mathematical subjects then under study during the early twentieth century (Dalmedico, 1997) but was nevertheless seen as a branch of applied mathematics (Fisher, 1958). This section is not concerned so much with the fact of mathematics per se, but more so with its application within psychology and assessment which, when considering the overwhelming influence of statistics and measurement in psychology, pales into insignificance as a contributor to a perhaps ill-directed trajectory that the discipline is traversing. Measurement theory assumes statistical viability and validity and statistical techniques consequently have as their base of manipulation basic arithmetical procedures. This really is "the science that elaborates the abstract structures that all progressions have in common merely in virtue of being progressions" (Benacerraf, 1983a, p.291). Before measurement theory and statistical practice is looked at within intelligence research and allied areas, the mathematical underpinnings of these manipulations must be more closely scrutinised. In order to query the foundations of mathematics it is necessary to review the philosophy behind these foundations, an exercise some (Putnam, 1983a) might argue to be unnecessary. Unnecessary when one comes to identifying the course of mathematics and its practical application yes, but such an inward turn to basic epistemology and ontology is necessary in this particular instance as mathematical theories carry with them ontological assumptions (Lehman, 1979).

The enterprise of mathematics, as vast as it is, can trace its course back to the notion of numeracy and its roots are derived from the seemingly simple task of counting[14] and the necessity of measuring land, counting produce and people (Boyer & Merzbach1991; Bronowski, 1974; Clapham, 1996; Eves & Newsom, 1965; Ifrah, 2001; Livio, 2003; Martí, 1996; Moring, 2002; Omnès, 2005; Pascoe, 1992; Porter, 1997; Roberts, 1995; Sardar, Ravetz & Van Loon, 2000; Stewart, 1995, 1998; Williams, 1997). Counting of frequencies is one of the most utilised practices in psychological quantification (Michell, 1999) and precedes measurement (Wright, 1999). Mathematics' role is paramount in measurement and although measurement did not originate with mathematics, it has given to the discipline of measurement its tools and thus "provide(s) the ultimate foundation for better

[12] Schönemann (1994) is also at pains to question the utility of mathematics in the natural sciences as well!
[13] These are metaphysical arguments and one can safely proceed with work in the real world as we know it. Yet these types of questions always remain lurking in the background.
[14] As history illustrates, what is seemingly simple today was in the past considered unfathomable and the strides made were immense at the time. How could anatomists fail to see that blood circulated in the body for instance (Zimmer, 2005)? Of course many factors play in on any one situation within historical contexts, contexts which one is not always privy to and hindsight is in fact quite conceited (Dawkins, 2005). How could we fail to negotiate our way around the issues of assessment?

practice and the final logic by which useful measurement evolves and thrives" (Wright, 1999, p.73). As is to be expected, geometry and trigonometry, albeit not in any formally disciplined activity as such and having been utilised by the Mesopotamians, Egyptians and Greeks, preceded algebra by a good two thousand years. This method was originally employed by the Islamic countries (Moring, 2002; Russell, 1983). The rudiments of counting, order and magnitude are not unique to human beings (Boyer & Merzbach, 1991) and just as there is speculation as to a language centre in the brain so too is there debate circling the issue of a similar innate mathematical centre in the brain or "number module" (Butterworth, 2000) or at the very least innate neural circuitry already present in infants (Livio, 2003), a speculative adaptive mechanism (Demetriou & Valanides, 1998).

Our mathematical rendering of reality could be construed as the equivalent of how we think (Heyting, 1983b).[15] Counting and the use of mathematical concepts can also be framed within the cultural context in which they occur (Ernest, 1994; Hersh, 1994) but that it is an exclusively human innate propensity can be argued. The need to propose order on a system from without is a reflection of an almost obsessive need to arrange, order, manage and make manifest the underlying natures of things and as Brouwer succinctly states "the results of counting and measuring take so important a place, that a large number of natural laws introduced by science treat only of the mutual relations between the results of counting and measuring" (1983, p.77). Is this then not the *prima facie* case for psychological assessment? The need to propose order by fiat of measuring and counting? Is there perhaps not an alternative yet equally scientifically feasible manner in which to propose order?[16]

Regarding the twentieth century, the major highlights within mathematics are perhaps the most daunting in terms of their overwhelmingly large influence within areas other than mathematics. Events prior to 1931 were an effort to mechanise the reasoning processes involved in mathematics (Hofstadter, 1980). France, Germany, Britain and Italy were the foremost mathematical countries toward the end of the nineteenth century having been displaced somewhat by the United States during the latter half of the twentieth century, predominantly due to emigration during the second World War (Dalmedico, 1997). Prominent mathematicians during the late nineteenth and early twentieth centuries included Henri Poincaré (1854-1912) who, among many other contributions, introduced the concept of the group and utilised analogies emanating from non-Euclidian geometry as well as applying qualitative techniques to celestial mechanics (after work pioneered by Newton, Lagrange and Laplace[17]) and laid the foundations for topology (Levenson, 1997; Murray, 2004). David Hilbert (1862-1943) who, like Poincaré, was wide-ranging in various mathematical areas saw among other things the formation of a systematic axiomatic method for mathematics in 1899 (Chaitin, 2006; Coveney & Highfield, 1995; Dalmedico, 1997; Hofstadter, 1980; Maddox, 1998; Rucker, 1987; Stewart & Golubitsky, 1993). He did this by deriving a formal axiomatic model for Euclid's geometry and hoped to create for mathematics "certitude" of its methods (Hilbert, 1983; Murray, 2004) and created the new discipline of metamathematics or the theory of proof which was of course to be reanalysed by Kurt Gödel (Maddox, 1998).

Hilbert's foundationalist approaches towards mathematics, which melded in well with the positivism then reigning (and can therefore be contrasted with the Platonist foundation) (Kreisel, 1983), resulted in the point of departure for modern algebra as well as a school of logic and was firmly entrenched in set theory and axiomatics and algebra.[18] This resulted in various offspring among them empirical formalism whose dictates maintain that an objective subject matter does exist for the enterprise of mathematics stripped of all but the most rudimentary philosophy (Curry, 1983). Chaitin (2006) however offers a convincing counter-argument to the "fact" that mathematics is concerned with empiricism and states that, like physics, mathematics can be classified as quasi-empirical and that in order for mathematics not to become isolationist, the discipline should make axiomatic leaps whether or not it is open to provability. The notion of mathematics as abstract science which can relinquish the system of empirical checks is the most often presented front (Ellis, 1966; Woods, 2003) for its propositions are presented *a priori* and changes in the physical world have no bearing on its abstract nature in the ethereal realm; they are timeless propositions (Ellis, 1966). Whereas Euclid's work was constructed, Hilbert's axioms existed from the start (Bernays, 1983). Hilbert's work spread into areas as diverse as relativity, quantum mechanics, matrix algebra, group theory and theoretical physics (Dalmedico, 1997). What was at the time considered an almost perverse turn-around for mathematics came in the form of an affront to codified Euclidian geometry and Aristotelian syllogisms as many centuries were to pass before a renewed look at these axiomatic foundations took place (Hofstadter, 1980; Kreisel, 1983). The idea of non-Euclidian geometry was a shock to many as it highlighted the fact that mathematics as a tool was not only a utility for real world studies but was also a tool for more abstract and esoteric areas of concern; all encompassed within what was once considered reality; the true version of reality. The Kantian notion of intuitive derivation of reality was thus turned on its head (Eves & Newsom, 1965; Mays, 1966). Along with the discovery

---

[15] In keeping with the intuitionists' approaches towards mathematics (see below) mathematics is our mathematics, a squid's mathematics is theirs and a Zoggian's mathematics is theirs too (wherever and whenever Zoggians's happen to be). Which is the true mathematics? Is there a true mathematics? These are arguments and questions often raised in relation to the Platonists. Yet again, we are confronted with the eternal dichotomy, forever standing in the way of thought.
[16] Are our brains so hard-wired that to conceive of such an equally valid proposal lies forever beyond us?
[17] Himself playing a role in the development of population statistics and thus one of many forerunners of statistics in the social sciences (Lazersfeld, 1977).
[18] The mathematics taught in high school can be largely attributed to the Hilbert school of foundationalist mathematics.

of non-Euclidian geometry, the recognition of the existence of algebraic structure played a role in further cementing the development and acceptance of the axiomatic method in mathematical research (Eves & Newsom, 1965). What reality was at this time was now called into question. Along with an effort to formalise mathematics and to establish for it logical and axiomatic foundations, flaws were evidenced and although not destroying the neat foundations which had been laid, forever altered the course of how mathematics was to be thought about. Three major "crises" in mathematical foundations (or at least the philosophical interpretations emanating from these supposed crises) can be highlighted: the discovery of non-Euclidian geometry, the non-existence of a consistency proof for mathematics and no universally agreed upon solution to various opposing views within set theory (Putnam, 1983a).

4.2.1 The philosophical implications of mathematics[19]

In order to critically appraise the influence that mathematics has had on psychology and assessment in particular, philosophical issues, once again, need to be turned to. This ever-present philosophical concern which has permeated the discussions so far, attest to the need for a "return to roots" within psychological assessment. As with any topic within philosophy, mathematical philosophy is no less a daunting area of research and debates often overlap the boundaries between theoretical physics and mathematics. The application of mathematics to various areas of physics was doubtless necessary for the advancement of physics and in order to proceed with these areas of investigation, the "language of mathematics" had to be written and systematised as with any discipline calling itself a science (Gribbin, 2003). The universe was, as Galileo stated, written in the language of mathematics (Damasio, 2003) and physics is currently written in and described most accurately by mathematical language (Gardner, 2003), a language which is perhaps the most sophisticated of the sciences (Bronowski, 1974).[20] However, this summarised discussion will merely touch on the philosophical issues pertinent to mathematics as it pertains to fundamental issues of assessment and what it is means to measure.

Broadly speaking, the study of mathematics has highlighted the role played by philosophical inclinations and how proponents of various schools of thought have in fact conceptualised the place of mathematics and the type of answers it can expect from its varied areas of endeavour. As with prior discussions on consciousness, nature/nurture, intelligence and realism/relativism; epistemology and ontology are core to these various mathematical affiliations; what is the area of mathematics? What can be studied in a mathematical manner? What is "true" in a mathematical conjecture and how can something be proved conclusively? The notion of truth in mathematical reasoning especially, dates back to the early Greeks (Penrose, 1999). What is or is not amenable to mathematical rendering? Are there limits to which mathematics can aim? What is the proper area of mathematical concern and most fundamental of all what is mathematics? These are of central concern to mathematical philosophers which might have an unnerving effect on mathematical psychologists for instance, for they are the utilisers of a method which is itself still at odds as to its own nature. Mathematical psychologists along with statisticians employ sophisticated techniques in order to deploy such methods on the terrain of psychological assessment, an area itself replete with conjectures and lack of finality. Assessment then, is itself utilised in an area even more undefined and nebulous, namely intelligence research. In sum, the foundation structures are creaky, the pillars are unstable and the ideas upheld by these shaky foundations are not really understood.[21] Figure 57 details the overlapping areas of concern which is shared by philosophical mathematics and the scientific method and how they play forth into the area of psychological assessment concerns.

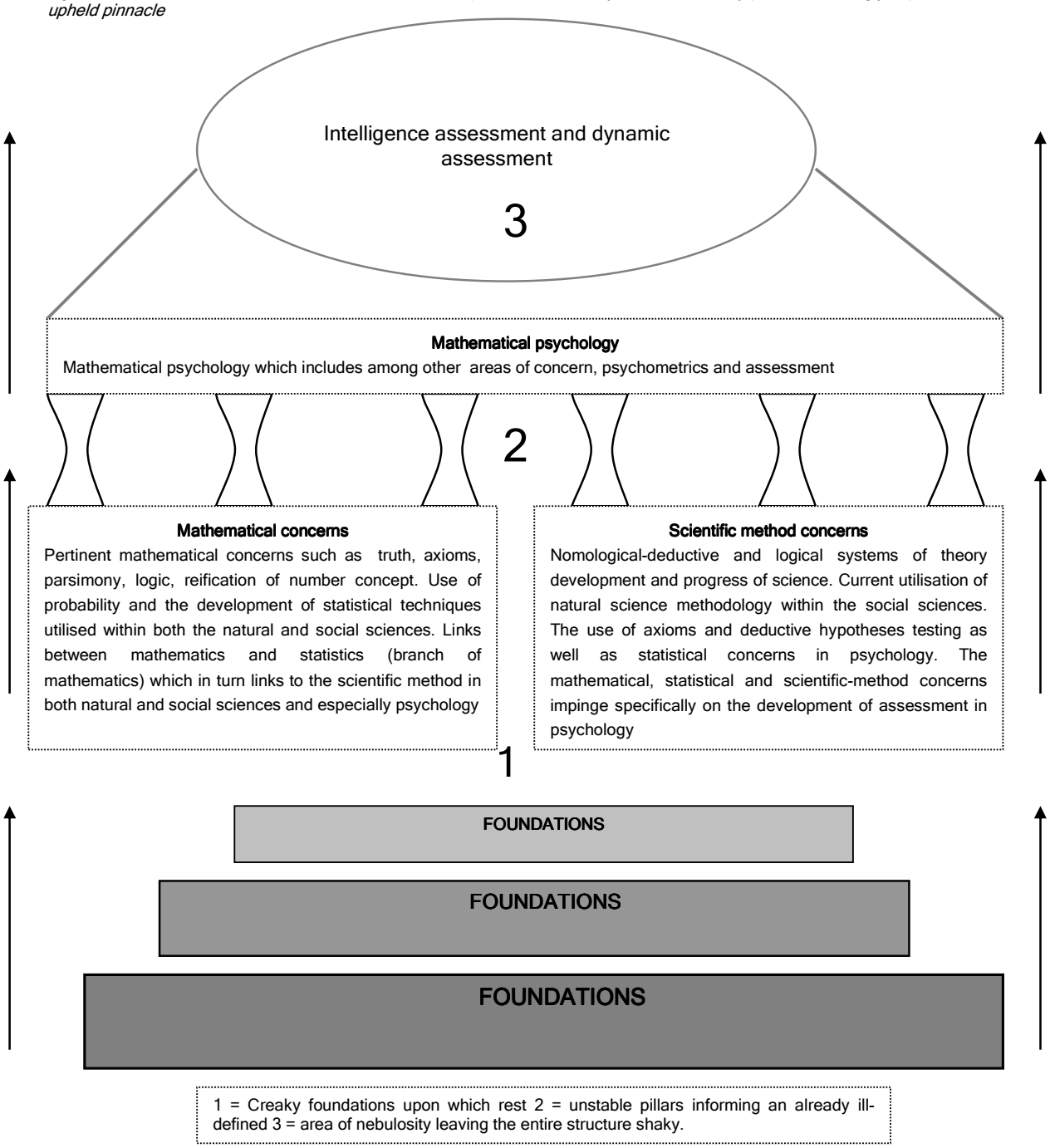
---

[19] Mathematics as formal discipline carries on just as surely as if there were no philosophical issues entangled in its history and has applications in everyday life, the likes of which we take for granted. So although the thesis posited here is that a necessary re-evaluation of fundamentals needs to take place, one also cannot deny that the practice is alive and well and it is often the case that many mathematicians may themselves have no particular viewpoint on the matter of foundational philosophies (Penrose, 1994).

[20] Although he refers to humans' need to impose order on the world in the context of the study of calendrics, Gould (1998) nevertheless cautions the reader not to be "oversold on nature's mathematical regularity". The image of mathematics being the universal language is to be tempered by what, why and how it is utilised by humans.

[21] How disconcerting then to think that most lay-people (those in fact seeking to benefit from assessment) are the most likely to think that mathematics is a "pure" discipline, that assessment is based on hardy and robust mathematical psychological techniques and that intelligence is a defined construct. None of which is true. This is not denying the incredible strides made in any one area, however the rate at which these areas are touted as being valid and reliable is misleading and at times unethical.

Figure 57 Three realms of shared concern; the base and pillars of which may not be as entirely pure as seemingly implied in the upheld pinnacle



1 = Creaky foundations upon which rest 2 = unstable pillars informing an already ill-defined 3 = area of nebulosity leaving the entire structure shaky.

Immanuel Kant's idea of the mathematical *a priori* was not entirely in keeping with Plato's (perhaps the first theoretician; Ayer, 1983; Livio, 2003) ideal realm (theoretical realm?) in which all things in their true and pure forms existed. Although one can see the allure and value of Plato's reasoning (Penrose, 1994, 2005; Scruton, 2004). Mathematical objects as such do not exist (Plato advocated that they did); in other words there is no transcendental realm in which mathematical objects are situated; rather our experience of empirical phenomena is grounded within the framework of *a priori* mathematical understandings (Kant) (Ifrah, 2001; Scruton, 2004).[22] Mathematics conveys "rigorous truths which reason is capable of discovering without need of experiment and yet truths which may always be confirmed by experiment within such broad limits as experimentation requires" (Cournot in Ifrah, 2001, p.355). This is a fundamental issue worth pursuing here as it highlights the place that mathematical reasoning inhabits. The appealing idea behind Plato's transcendental realm is that truth or proof exists to prove mathematical entities but which are forever independent of our knowing them directly (Livio, 2003; Scruton, 2004). David Hilbert, though not subscribing in full to the Platonist ideal of separate existing realm of number can be crudely considered a Platonist (Scruton, 2004). Those favouring the idea that proof was all there was and nothing beyond it were more inclined to think about the mathematical realm as existing only as and in proof itself and hence it was constructed via the mechanism of proof (Livio, 2003; Scruton, 2004). This is in keeping with the ideal as depicted by Kant. However problems soon emanate from such a position due mainly to the idea from logicism which, in this instance would maintain that a confirmatory proof or proof negating a logical claim is all that could be afforded within the proof where surely confirmation or negation would be necessitated. The situation unravels when neither proof is the offing. This would lead to a system which is either meaningless or neither true nor false.

The three most widely agreed upon modern philosophical foundations to the study of mathematics are the logicist foundation exemplified by the British philosopher mathematician pair Bertrand Russell (1872-1970) and Alfred Whitehead (1861-1974), the mostly French intuitionist (or constructivist) foundation led originally by the Dutch mathematician Luitzen Brouwer[23] (1881-1966) and the mostly German formalist foundation developed principally by German mathematician David Hilbert (Benacerraf & Putnam, 1983; Brouwer, 1983a; Clapham, 1996; Curry, 1983; Dummet, 1983; Ernest, 1998; Eves & Newsom, 1965; Fuchs, 1967; Lehman, 1979; Penrose, 1999; Wang, 1974). There are of course numerous arguments against the absolutist vision of any the three foundationist schools (Ernest, 1998) but their supremacy, value and role within the enterprise of mathematics cannot be understated. Mathematical exactitude reside on paper for the formalist, in the mental realm for the intuitionist and in the independent theoretical realm for the Platonist (Brouwer, 1983a, 1983b).

Regarding the origins of the logicist movement there are numerous mathematicians and their works which converge in on the logicist foundation such as Gottlob Frege's (1848-1925) first proposed system of the derivation of mathematics from a logical foundation (1884, 1893, 1903) who in turn had been influenced by Leibniz and Peano's attempts at application of flexible symbolism within mathematics (Ernest, 1998; Gödel, 1983a; Harnish, 2002; Woods, 2003).[24] This preceded Russell and Whitehead's independent opus on the very same issue and Frege took as support of this view Hume's assertion of the analyticity of mathematics (Carnap, 1983a; Frege, 1983, Mays, 1966; Scruton, 2004). At the time though, Frege contended that mathematics and logic were indistinct,[25] but this was largely ignored until the arrival of Russell and Whitehead's treatise (Boyer & Merzbach, 1991),[26] and even in Frege's description of the concept of number he was able to derive a definition utilising letters of the alphabet (Frege, 1983). Leibniz maintained the notion of logic as a science and the actual reduction of mathematics to logic was maintained by Dedekind and Frege. Mathematical theorems were later stated in logical symbols by Peano (Eves & Newsom, 1965). Russell and Whitehead's treatise utilised Georg Cantor's (1845-1918) idea of a one-to-one correspondence illustrating the idea of equinumerosity[27] (Cantor established set theory and developed the idea of infinite sets among other accomplishments) (Boyer & Merzbach, 1991; Gödel, 1983b; Scruton, 2004; Stewart, 1990; Wang, 1983), the work of whom was considered as sublime by Hilbert. Together, the ideas of Cantor, Hilbert and Poincaré made manifest the centrality of proof (Penrose, 1999) within mathematics.

---

[22] Although Kant's theory was largely overshadowed by the discovery of the exact feature which had made Kant's case: the axiom of parallel lines which had been turned in on itself to reveal another axiom in keeping with non-Euclidian geometry (Brouwer, 1983a). Ayer (1983) maintains that by stating that something is *a priori* is to level at it the notion of its being tautologous.

[23] This mode of reasoning goes back as far as Descartes and Kant and Brouwer and his followers can be considered as modern proponents of this school. Similarly, Aristotle can be viewed as the originator of the logicist approach (Beth & Piaget, 1966; Eves & Newsom, 1965).

[24] Frege maintained that Socratic rhetoric was just for show or "colour and shading" and that each linguistic sentence could in reality be reasserted as objective content (Ernest, 1998). When one thinks about this, it is not difficult to understand how and why Frege may have come to such a conclusion. One can, after all, proceed in error in such rhetoric and come out looking positively victorious. Perhaps Frege had in mind a way in which logical deduction would be the ultimate arbitrator in such cases. Looking at arguments logically and then reducing them to symbolic notation is reminiscent of the hypothesis procedure within psychology. The link between such endeavours becomes ever more manifest. It would seem that quantitative psychological methodology has quite a history behind it, one that is perhaps not often thought of.

[25] Until Frege's as well as Russell and Whitehead's works, mathematics had traditionally been aligned with science and logic aligned with Greek (Russell, 1983).

[26] Incidentally Gottfried Leibniz (1646-1716) had already employed logical means of proofs but had erred in this regard.

[27] "The concept *F* is equinumerous with the concept *G*" and in so doing Frege was able to show, in a logically argued case, that "we have reduced one-to-one correlations to purely logical terms and can now offer [this] definition" (Frege, 1983, p.143). Is this a foreshadowing of rules of correspondence utilised within representationalist measurement?

The logicist school did not promulgate the foundations of mathematics as consisting entirely of logical predicates, but reduced mathematics to logic in addition to theories of sets and properties (Benacerraf & Putnam, 1983; Quine, 1983a). In so doing the bridge between mathematics and logic was highlighted; a gulf which had until the works of Frege, Russell and Whitehead's treatise been considered unbridgeable. Although it often appears that the logicist school did in fact streamline mathematics to logical predicates (Carnap, 1983a). The main differentiating feature between mathematics and logic is the finite predicates of the latter and infinite predicates of the former (Heyting, 1983b). The Italian mathematician and leader of the modern Italian formalist school Giuseppe Peano (1858-1932) whose symbolic logic had as a consequence the development of important notation (Brouwer, 1983a; Clapham, 1996) was also responsible for the axioms of integers and his work towards the axiomatisation of mathematics in general (Eves & Newsom, 1965) and all that was left now was to define the fundamental concepts utilised within these axioms (Russell, 1983; Scruton, 2004).[28]

The logicists' entrenchment of logic as foundation of mathematics was flawed, however, due primarily to the fact that set membership proved to be the foundation of mathematics as well and not only logic which had among other repercussions certain insoluble paradoxes, the most well-known of these paradoxes known as Russell's paradox. Such seemingly self-referential paradoxes (there were other paradoxes) were the irritations for which attempts at banishment culminated in Russell and Whitehead's Principia Mathematics (Hofstadter, 1980). Russell's paradox had been anticipated as a paradox by Cantor (Penrose, 1999), an attempt for which resolutions were offered (Clapham, 1996; Scruton, 2004), which leads on to the set theory. The logician Frege's original ideas in set theory (1884) in part inspired the works of two mathematicians who helped lay down seven axioms for set theory and which are known today as Zermelo-Fraenkel set theory (after Ernst Zermelo; 1871-1953 and his contemporary Abraham Fraenkel; 1891-1965). These axioms are not logical truths but rest on intuitive foundations in which sets themselves are the primitives of mathematics which is *a priori* (Scruton, 2004). Set theory posits that numbers themselves are sets (Carnap, 1983b) in contradiction to the nominalist idea of number, where numbers were not objective things but only meaningful in terms of what they expressed or the context in which the numbers were placed (Minsky, 1988). The number four "4" is not anything in particular other than a representation of the set of "fourness", this conception of a set of four is really a Platonist idealisation of what four really is or at least what it means (Parsons, 1983; Wang, 1983). The set containing the set of empty elements is itself a set of one set and hence is not zero but one (Penrose, 1999). This particular rendition of number is of course just one such derivation for the concept of number (another scheme includes Alonzo Church's lambda calculus and there is also another scheme which envisages numbers intuitively without the need to capitulate to the concept of set at all; Benacerraf, 1983a). Frege's (1983) definition of number was that, as an "objective object" it only made sense once it was contextualised within a sentence of some meaningful setting (Ernest, 1998), a sentiment echoed by Russell (1983). He added to it by emphasising the human tendency to continually define primitives in terms of yet more primitives and so on ad infinitum, but knowing that our capabilities are finite we are unable to continue upon a path of infinite definitions and so stop at the most logical point, hence the emphasis on logicist leanings. An interesting feature within measurement theory in psychology is the use of intensive measures as extensive measures via conjoint measurement. In a manner mathematics' primitives find a measurement theory counterpart in non-derived extensive measures.

Scruton (2004) furthermore adds that one of the set theory foundation axioms, namely the "foundation axiom" states that there are no ungrounded sets[29] which Scruton levels as a partial attempt at the solution of Hempel's paradox of confirmation, so named because there is nothing illogical about the manner in which laws are confirmed. That is, laws can be inductively confirmed by instances in which something is said to be as well as something which is said not to be.[30] Von Neuman, as mentioned above, also provided an alternative to the Zermelo-Fraenkel set theory and so numerous were the branches of mathematics and applied mathematics that Von Neuman and Norbert Wiener (1894-1964) were both involved in many varied aspects of it continuous growth, which was starting to overlap more and more into the social sciences (Boyer & Merzbach, 1991). Hilbert's formalist school of mathematics (programme) as well as the logicist school could not hold up to scrutiny and perhaps the largest blow to the formalist school was Gödel's incompleteness theorem (Gardner, 2003; Penrose, 1999).

In essence, Hilbert's formalist agenda was to locate for any area of mathematics a number of axioms with defined rules and procedures allowing any defined reasoning in that area to be incorporated. Any mathematical proposition in this system would be consistent and complete and established for non-finitary mathematics a finitary construction (Von Neumann, 1983). What is

---

[28] Five axioms from which all arithmetic can be derived. The first three axioms present three primitives (or numbers) and so allows the fifth postulate to prove theorems about all numbers by considering only these three (Hempel, 1983). This led to the logicists' attempts to define the three primitives and to thus illustrate that the postulates are derivable via logic from the definitions, which is precisely what Frege and Russell independently set out to do utilising Cantor's one to one correspondence idea (Scruton, 2004).

[29] Which means that there are no sets which contain members which contain members which contain members and so on ad infinitum.

[30] To make this clearer: "all post-boxes are red" would be inductively confirmed if every street corner evidenced a post-box which was red. The logical equivalent to "all post-boxes are red" is a sentence stating "all non-post-boxes are not red". But this leads to an absurd state of affairs for one could quite happily spend the rest of one's life confirming that post-boxes are red by making completely inane comments. To get rid of this annoyance within set-theory then, the foundation axiom states that no such statements would be allowed. Here then is a link between the foundations of mathematics and the logic of scientific method (Scruton, 2004) which was discussed at length in chapter 3.

formalistic about this agenda is the fact that any prepositional meaning per se is irrelevant as it merely represents symbols.[31] Gödel showed that no specifically defined proposition (or its negation) could be proved within any formal system[32] and hence the notion of complete truth as understood by the formalist school is in fact incomplete. The formalist approach towards mathematics was anathema to the view endorsed by more Platonic-minded mathematicians in which, as stated above, an ethereal realm of mathematics existed (Penrose, 1999) (although in a manner of speaking, one can support the view that this be viewed as the realm of theory and was perhaps just framed in this manner by Plato and his cave depiction). Penrose (1994), as a self-confessed Platonist, highlights the need to understand the implications of Gödel's theorem (Gödel himself a Platonist), which does not result in "unknowable" truths lurking forever out of the grasp of human thought, but that solutions to problems can be known, they are simply not dependent on formal rules and contingent axioms thought up by human beings (or any other organism for that matter) (Coveney & Highfield, 1995; Ernest, 1998).[33] In other words, there does exist a Platonic realm of sorts, not a physical place, but rather a state of truth or knowingness for which solutions for insoluble and intractable problems do exist but for which solutions await discovery (Benacerraf & Putnam, 1983). The Platonic realm makes its presence felt within measurement theory more specifically within classical test theory's positing of a true score, which as a notion or idea is yet to be fully understood within its own theory (Borsboom, 2005; Borsboom & Mellenbergh, 2002). Is the true score in actual fact the construct score? The issue of hypothetical and substantive construct via its numericised score is a recurrent theme within this thesis and is found repeatedly in themes discussed in various dynamic assessment models in chapter 5 (sections 5.2.1 and 5.2.9).

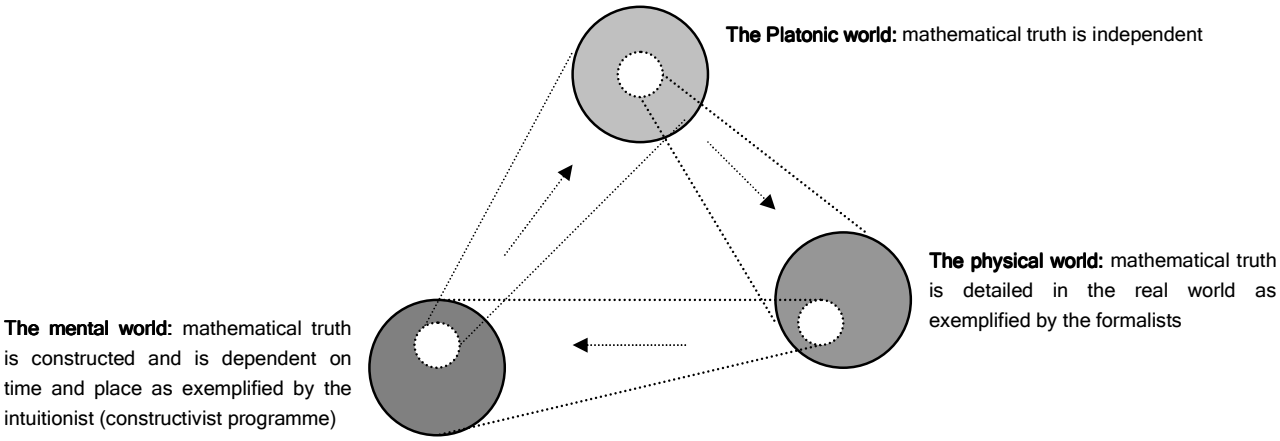
This is in contradistinction to the idea of mathematics conceptualised by the constructivists (intuitionists) in which mathematics is a creative activity for any undecideable proposition, in terms of human understanding of rules governing any proposition. Its truth too is undetermined and because there is nothing else which is relevant to the issue, truth is simply not available in the current system and hence there is no theoretical realm in which the truth resides. Rules and systems and any truths governing these systems are bound to time and place and there is no truth independent of human thought; no transcendental realm (Heyting, 1983a). The Platonists disagree and maintain that any correspondence between propositions and their mathematical counterparts are timeless and not context-bound. In other words just because something happens to be insoluble during a specific time and place it does not necessarily mean that a solution does not exist (Benacerraf & Putnam, 1983).[34] Could one not equate the constructivist programme with a relativist one and the Platonist programme with an objective one? In fact the similarities and contrasts between the mathematical foundationalists and early philosophers of science hinge around very much the same type of issue. Namely the level of reality to be studied and the realm of the unknowable (empiricist vs. Platonist mathematicians or philosophers, which in similar vein to the behaviourists in psychology, allowed no place for the existence of abstract entities, i.e. sense data was the only viable type of data) (Ayer, 1983; Carnap, 1983b). In a sense then, the affiliation towards the Platonist way of viewing mathematical reality could be construed as being at odds with the positivist programme in general and thus inconsistent with the affiliations stated in chapters 2 and 3. How is the author to defend this inconsistency?

Other than offering an account of attitudes towards the reality of mathematical entities on the one hand and looking at positivist tendencies within the global practice of science on the other these views are perhaps inconsistent. Others would disagree with the very need for the foundations of mathematics to be studied in the first place, as there really is no need for it as the discipline can get along without it all the same (Bernays, 1983; Putnam, 1983a). Putnam (1983b) maintains that a balance be struck between metaphysical realism (intuitively grasping at the Platonic realm) and scientific verificationism (empiricism) in an attempt to "solve" issues between varying philosophical positions within foundational mathematics. Hence, if this is construed as inconsistent then it either needs more thought or is trifling enough not to be of concern. Figure 58 illustrates the interwoveness of the three realms and figure 59 illustrates the case with the real life construct "IQ".

---

[31] This idea is reminiscent of the idea behind Ramsey sentences (see chapter 3) in which redundant terms are done away with in order for theories to proceed unencumbered. This is also reminiscent of Frege's attempt at logical sentences stripped of all but the bare logical essentials.
[32] The author is reluctant to enter into mathematical details, the meaning of which will by and large be non-sensical to a non-mathematician, such as the author! Penrose (1994) is of the opinion that this is just lazy and one way of taking the easy way out. This is conceded.
[33] The Platonic concept makes its appearance again in the section discussing measurement below.
[34] This Platonist argument can be taken further. There may very well exist solutions to problems about which humans are not cognisant. There could be potentially other types of space-time continuums of which we are simply unaware for which there exists mathematical uncertainties about which we are unaware and for which there exist solutions, to which we are also unaware. So, in a manner, the Platonist "take" on mathematical certainty and philosophy is quite compelling. The constructivist rendition of mathematical solubility, it would seem, is too dependent on time and place. Once again a relativism of sorts raises its head.

Figure 58 The relation between various worlds and how each world takes a small part of its predecessor with it into the next world with attenuations as to the mathematical foundations (Penrose, 1994, p.414)



The Platonic realm includes proofs and solutions (as well as proofs to the contrary) about any known as well as unknowable mathematical topics, even those not humanly knowable for they exist independently of our knowing them or even of our not knowing them. In fact it is entirely plausible that if any such truth were evidenced we would not necessarily be capable of recognising this truth as having obtained "or of getting ourselves into a position in which we can so recognise it" (Dummet, 1983, p.105). Yet the mental world can imagine a good deal that was once unknowable and now knowable but can only begin to fathom the depths which the Platonic or theoretical realm encompasses. The physical world is a given and the mental is to a large extent dependable and an extension of this physical world, yet conjectures in the mental world can play back forth into the physical world as is evidenced with theoretical physics. The idea or notion of a Platonic world is really just a metaphor to aid in our understanding that neither the formalist nor intuitionist renditions of what mathematics is, can come to any sort of "truth" when confined to the definitions of their own respective programmes.

Paradoxically within empirical science, mathematical truth is not as easily proven as other disciplines within science have shown yet its application within empirical science is without question (Hempel, 1983). Traditionally the route followed by any science in determining for it the truth or falsity of an hypothesis was to prove the validity or lack thereof of the hypothesis, this was discussed at length in chapter 3. Mathematically speaking though, the parallel to verificationism within the scientific method is the method of proof (Gardner, 2003; Putnam, 1983b). Mathematics as empirical science however does not avail itself of quite the same characteristic. Empirically validating that 6 + 2 = 8 is a complicated affair if the task is to be undertaken in a similar manner as utilised in physics for instance. Stating an hypothesis such as 6 + 2 = 8 and relying on empirical observations to confirm such a relation is almost impossible. The nature of the *a priori* in mathematics can be accepted on the basis of a tacit understanding of the meaning behind what is attributed to the concepts of the numbers "6" and "2 " and "8 "as well as the operator "+". What is necessitated by these figures is clearly understood by those utilising them.[35] Students will understand that when given the figure 6 it will be necessary to add to this another figure, namely, 2 and to describe the resultant relation emanating from having put the two together. This is usually taken for granted, but the implicit understanding can be explicitly complex, especially if proofs have to worked out. The price that is paid for utilising this *a priori* analytical truth is that the statement says nothing about facts. It is not a factual system at least not in the real world of factual information. We do not inhabit the Platonic realm and do not posses all theoretical truths (Hempel, 1983). Numbers are abstract "measures" as stated at the start of this chapter (Ifrah, 2001); counting live stock, measuring the angles of structures and employing techniques to aid in the construction of buildings as well as understanding the movement of stars in the sky.

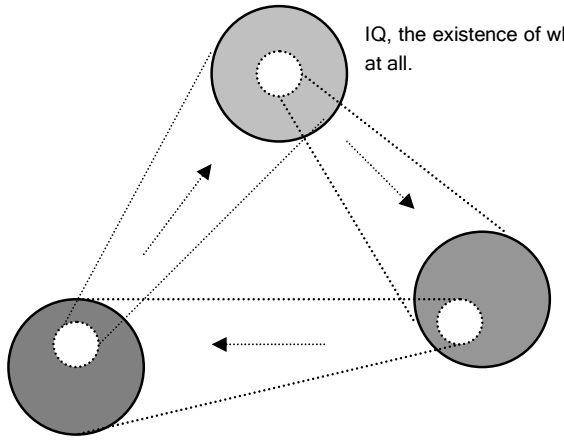
The move from measurement of objective things (lines) or entities (angles in a triangle) to the measurement of unknowable entities (intelligence) is a move perhaps not deeply considered as problematic. Things are attributed meaning by means of words which are later substituted by numbers representing these words, which are in turn substituted by more abstract symbols until such time that symbols are employed for regions within the mathematical realm for which no conceivable physicality exits (Ifrah, 2001). The level of abstraction reached within any area of mathematics does not impel the need to study mathematics for the sole sake of enhancing the mind, as was stated by Karl Jacobi but because mathematics can aid in the understanding of the world in which we live (Fuchs, 1967). That it has proven to be precisely this is telling of modern conceptions of the interplay

[35] The closest the author has come to describing the strange nature of symbols is to refer to Stewart's (1995) analogy in which he states that "mathematicians are forced to resort to written symbols and pictures to describe their world - even to each other. But the symbols are no more that world than musical notation is music" (p.ix).

between measurement and psyche. "6 + 2 = 8" is hanging up in the air somewhere, lodged in the ethereal realm of truths yet to be discovered; the statement is not anything in particular, is neither exists here nor there; it is simply a statement describing the relations between two things. What things? It is meaningless to ask. It is in fact quite amusing to ponder on this very issue for it is something we take for granted in our everyday dealings with mathematical and numerical entities. It is very abstract and extremely vexing as a discipline with which to try to come to grips. Have psychologists misapplied the services that mathematics has to offer or have they misinterpreted the results from mathematical manipulations of psychological data? The error does not lie with the enterprise of mathematics but with the skewed interpretation by psychologists' own understandings of measurement.

The assessment of intelligence from neuroscience applications (measurement of neural conductivity) is one method of assessing what is purported to be intelligence; the assessment of intelligence by measuring the results on a test supposedly assessing for intelligence is another method of assessment; statistically deriving results from sets of answers on a test is yet another method of adjudicating the level of intelligence. Psychology has employed all these methods in what many consider to be successful methods. The interwoveness of biological understandings of the brain with the measurement accuracies from mathematics and the statistical refining of results has jettisoned the enterprise of assessment into a new era. Yet we are no further along the path of what it is that constitutes intelligence in the first place. Are the applications at fault? Surely not. Are the techniques at fault? Surely not. Then what is the problem? As Hempel (1983, p.291) puts it: "in the establishment of empirical knowledge, mathematics has the function of a theoretical juice extractor: the techniques of mathematical and logical theory can produce no more juice of factual information than is contained in the assumptions to which they are applied". One can likewise extrapolate these sentiments to the arena of assessment.

Figure 59 The relation between the three worlds utilising IQ as reified yet ill-defined measurable construct



Similar sentiments govern figure 59 as they do figure 58. Although not intended by Penrose (1994) to be utilised as a description of psychological reality, the case for assessment rests firmly on the assumptions predicated by mathematical measurability and reification. The intuitionist school (which broke upon the mathematical scene in 1924) can be considered a break away from formalism in which the rules governing the system of sets and membership of such sets are more emphasised than the sets themselves which are not necessarily believed to be in existence as such (in terms of the Platonic meaning of the entity existing). Followers of this intuitionism (entitled such due to its supposed mirroring of the human mind; Penrose, 1999) could trace their philosophical roots back to Aristotle who had deviated somewhat from the ideals proposed by Plato. Aristotelian syllogisms cannot however serve as foundation for the enterprise of mathematical science because it does not provide a complete analysis of reasoning within mathematics (Beth & Piaget, 1966; Poincaré, 1983). A mathematical idea was said to exist only once the mental counterpart of the same idea became manifest, hence one would construct the idea which led then to the alternative label for this approach.

One of the main contentions of the intuitionist school was the rejection of the Aristotelian law of the excluded middle[36] which briefly states that the negation of something is logically equivalent to the affirmation of its opposite which forms the basis of the

[36] Perceived by Hilbert as a heresy of sorts (Coveney & Highfield, 1995).

reductio ad absurdum rule often deployed in pointing out egregious errors of the logical type;[37]something either is or is not (Boolos, 1983; Coveney & Highfield, 1995). This school also denied the existence of a certainty until such said certainty could be proven to be true or false. For example, until the proof to Fermat's last theorem was concluded the issue of the theorem being true or not was denied an existence, thus making mathematics contingent upon time and place; a dependence which some found illogical and unappealing (Penrose, 1999). One can clearly see the stark disparity between the intuitionist and Platonic modes of reckoning about the reality of mathematical entities.[38] Aristotelian conceptions of methodology emanated from three postulates; deduction, self-evidence and reality. The third postulate is particularly emphasised with regard to the opposing intuitive view as espoused by Descartes and Kant where there is a discernable shift from the domain of the real to the domain of the constructed (Mays, 1966).

4.2.2 The link made manifest

Mathematics[39]*" (verb)* the study of number, form, arrangement, and associated relationships, using rigorously defined literal, numerical, and operational symbols'' (Reader's Digest Dictionary, 1988). Such is the unpacked definition of what it means when one practises mathematics of sorts. Notwithstanding the enormous range over which this definition currently extends, for these purposes, this definition is ripe with key words pertinent to psychological assessment. How successful such a venture has been within the domain of psychology is questionable (in terms of the use of number, form and arrangement in reference to non-numerical aspects such as IQ, colour and liking/disliking of a painting for instance) and as dynamic assessment very much forms part of the repertoire of assessment tools, it too is not exempt from criticisms levelled at it.

The reality of an object and its counterpart or shadow in the mathematical realm is easily stated in mathematical terms, such that two apples as readily seen by the eye, is rendered mathematical as $2x$ (x representing an apple). The method utilised to establish this transfer from the real physical realm to one of corresponding mathematical reality is deceiving at first glance, as the transfer back into the realm of the physical from one of mathematics cannot always proceed as easily (if at all) (Brown, 2001a). Measuring an object (an apple) and measuring its properties (colour or taste) are not synonymous concepts nor ventures; likewise measuring a rank ordered statement is not necessarily the same as measuring its properties (what the statement is saying, or what it means to a person). Using the aforementioned example of rank-ordered statements: quantifying *rank* (its nominal order) and quantifying *meaning* (falling outside Steven's original conceptions of measurement scales) reside within very different realms; realms which are either not delineated correctly or adequately enough, or, which are simply not taken cognisance of at the outset. Are psychological assessors to abandon all known methods of assessment which fall within the realm of the physical as exemplified in the scales of measurement? And if this is done, what exactly is going to replace this method? Or is the mere fact that such a question need be asked already indicative of our reliance on an outmoded appraisal of what it means to assess within the social sciences and specifically within psychology and dynamic assessment?

Dynamic assessment is currently and has for many years been burdened with the unresolved issue of not being able to bring meaning to quantifiable change (quantifiable as understood within the physical scales of measurement) and numerous attempts at ordering or bringing meaning to such change has merely resulted in the re-issue of the same physical mode of quantification to the problem. Various angles of explanation are considered when trying to explain away what it means to have undergone change (typically from a pretest to posttest scenario) but the angle of explanation is not what should be highlighted. It is the fact that any angle of tentative explanation is proceeding from an incorrect stance from the outset. The application of tools within the physical reality's realm is simply not applicable to change occurring in a totally unique realm (the above-mentioned differences between physical and mathematical reality and how they are measured). Phenomenological tools of change assessment are themselves inadequate in discriminating the nature of change, for although not as sharply described as are more rigorous tools (techniques having as their foundations the scales of measurement) phenomenological tools are themselves apt to lean towards meanings residing on scales of sorts (although they can never be charged with such explicit liaison with scales of measurement as can more quantifiable tools).

---

[37] One can see a thread linking the logical/mathematical with the varied philosophies of scientific method, proof, argumentation, deduction and how something is considered to be true or not. The link then, although tenuous upon first observation has turned out to be more than just tentative. For what is considered logically deductive (logic and mathematics) can be transplanted into the realm of science (scientific method) and transferred almost wholesale into the social sciences where as its last stop it is considered pure, complete and valid. Three notions, which as it has turned out, are not quite what they seem to be. The establishment is currently saddled with a monumental historical edifice of repute, yes, but ill-defined and uncomfortably positioned within psychological assessment. The first thing to do then, in an attempt to re-evaulate this issue, is as mentioned a number of times already, to revert back to the origins of assessment. This thread is convoluted, knotty and spread thinly in some regions.

[38] And the entire edifice of mathematical psychology is predicated upon such a lower order edifice. Granted, once the ascent towards assessment is begun, such shaky foundations and their consequences become less marked. What is the point then of illustrating mathematics' less than pure foundations? It is inconceivable to the author at any rate that these consequences would not have some sort of ramifications somewhere at the top.

[39] The assertion may well be made that language too is a notational system representing reality as it is filtered through our brains and if this is the case mathematics is merely another notational system. However, a possible retort to this would be that just as with language, mathematics too is limited in its representational capacity to reflect a true reality (no subtle leanings are made towards relativism here).

Logical positivists and empiricists would not condone forays into metaphysical realms, realms which cannot withstand the scrutiny of physical reality as transposed into mathematical reality (yet it is undeniable that many metaphysical research programmes did just this; and which have subsequently led to many discoveries later on such as the germ theory of disease, the wave-particle nature of light and so on; Nickles, 2001). Blending methodology from neuroscience into psychological assessment requires that use be made of the real and physical realm and so instantiating the need for scales of measurement. Is this argument/debate thus rendered null and void? Perhaps the right questions are not being asked or at the very least are being ill-phrased. A crude analogy can for the moment be employed to further examine the type of frustration being dealt with here: it is often stated that ''words are not enough'' by which one can extract the meaning that language is incomplete or too limited in its vocabulary and/or manner of expression to do justice to what is being felt by the person stating it. Akin to this analogy is also the disputatious contention that thinking cannot co-occur without language (cognisance is taken of instinctive actions which are behavioural manifestations of hard-wired actions and intentions).[40]

Just as charged particles were originally considered mathematical entities and not field points, so too can IQ points be considered mathematical entities but not real ''brain stuff'' (Brown, 2001a), for intelligence is not IQ and IQ is not intelligence (Grigorenko, 2004a). In fact one is far closer by stating that an IQ score may represent phenotypic values and not, as envisioned in the early twentieth century, as a genetic value. Emanating from incorrect conclusions from heritability studies the perception that this paradigm "measures" "intelligence" is incorrect and in keeping with the theoretical-substantive criterion or differentiation, "the heritability paradigm refers only to studies decomposing observed phenotypic variance in *human traits* into its *estimable* (not measured!) components" (original emphasis) (Grigorenko, 2004a, p.54). The fact that atoms are today discernable[41] as physical objects (Shao-Horn, Croguennec, Delmas, Nelson & O'Keefe, 2003) does not necessarily mean that such a path is in store for the ever-elusive IQ score. Two major philosophical issues which impinge heavily on change measurement within dynamic assessment's pre-posttest score scenario concern the nature of *representation* and the nature of *empiricism* (observable and non-observable entities) an issue touched on in chapter 2.

Concerning *representation*: When objects are represented mathematically, the representation may extend only as far as the object itself and not to its properties as discussed above. Is an item on a psychometric test measuring the item's ''objectness'' or is it measuring what it represents? Numerical assignation may imply one of many things, among others;

- quantitative reality as is seen in real life scenarios, for instance five key strokes ///// may reflect the number of sticks on the ground taking on a structural feel to what it means to represent a number (Brown, 2001a, p.263); a more or less one-to-one correspondence with reality. Here the mathematics is descriptive of reality
- quantitative description of the reality, for instance ''the green book has 232 sentences'' here describes what colour the book is as well as the numerical representation of the number of sentences; nowhere does ''232'' appear in the book, it merely represents a framework for interpreting the properties of the book (Brown, 2001a);[42]a second-order correspondence with reality; here the mathematics is representative of reality

Concerning *empiricism:* This attends to what is and is not directly observable[43] (Trout, 2001a). The weight of an object, the elasticity of an object and the particular phase in which a material may happen to be are all directly observable, measurable and quantifiable. Any mathematical modelling imposed on these objects which seek to determine behaviour may proceed along the lines of description. The change in attitude towards the liking or disliking of a presidential candidate is readily quantifiable in terms of a first-order level descriptive exercise. However, the second-order representational mode[44] of quantification is elusive. Quantifying the answers to questions concerning a political candidate (when a scale of one to five is used for instance) is straightforward. The crucial and most often ignored part of this exercise is the transference of a behavioural concept (in this instance attitude) to a number. A behavioural phenomena and a numeric quantity are not synonymous. The fact that these two ''notions'' were ever brought together has lead to the particular situation with which researchers are currently saddled. The original and at-the-time inspiring idea to numerically represent on any scale of measurement any behavioural aspect (Francis Galton largely initiating such an enterprise) has perhaps long since outgrown its usefulness and should have matured beyond a

---

[40] The word 'intention' however is laden with meanings of forethought, for to be intent upon something, one can surely not be thoughtless for it is an intent that something occurred as planned.
[41] Although the answer as to what it is that brings mass to particles (the elusive Higgs particle or the god particle) will have to wait for results from experiments conducted with the Large Hadron Collider (Muir, 2005).
[42] Just as numerics represent abstract or real objects, varying radix systems of numerics may too well be used. For instance the hexadecimal numeric for the decimal number 232, is equal to E8. The point emphasised here is that the actual numeric representation may be arbitrary, the fact however that it imbues specific meaning is not quite as trifling a matter.
[43] Note that atoms are not directly observable by the human eye, but of course this does not mean that they are not directly observable; they are however *not* inferred and this is the distinguishing feature between observable and non-observables.
[44] Halford and Wilson (1980) elaborate on what they refer to as second-order isomorphism which is the representation of representations. Here the link with the discussion on cognition in chapter 2 is highlighted. The problems concerning representation pervade both the fields of measurement theory (see section on measurement theory below) and cognition alike. They are both "concerned with providing a valid model of the world, or of some segment of it which happens to be important for a particular purpose" (p.359).

simple one-to-one correspondence of behaviour to number.[45] This quantification quandary of twentieth century positivism was a direct spin-off from logical positivism evidenced during the early twentieth century (Mouton, 1993a).

The issue of concern here is not the need or lack thereof for mathematics (for mathematics' role in aiding communication or in fact *being* the communication channel within the realms of physics for instance is not in question) but its use within a different realm. This is the psychological realm which does not facilitate the use of representation *as well as* the realm of physics which is an issue that needs to be addressed. That it *is* used and *can* be used is not in question; that it can be better used is however debateable. The origins of the practise of ''psychophysics'' as it was known in the late nineteenth century were firmly entrenched in the mathematical norms of the day in which physics and other aligned natural sciences had found a rich repository (although itself still developing) of techniques deemed helpful in elucidating physical concepts. Why not impose upon the behavioural realm such a vast array of techniques which would no doubt aid in securing for psychology a firmer and more acceptable platform in accordance with the strict objectivity of the day? (Savage & Ehrlich, 1992). This is indeed what played forth in the development of psychophysical measurement (Gould, 1997a; Leahy, 2000; Nunnally, 1978; Sahakian, 1981).[46] Even if it was conceded that psychological constructs be measured on the continuum of nominal-ratio scales, it still stands that many behavioural sciences concepts do not avail of ratio scale attributes (Smit, 1996) nor interval scale attributes although test developers strive to attain such levels, interval level measures often yield scores with ordinal properties[47] (Cliff, 1991b; Murphy & Davidshofer, 1998; Rigdon, 1998). Rasch (1980) was keenly aware of this when he stated that the ordering of responses on a test does not mean that abilities are being measured on a ratio scale.[48] Person ability and item difficulty are conjointly solved allowing for ratio-level analysis, hence a stochastic measurement model attempting to avoid the "foolish practice" of ambiguous and contradictory results (Wright, 1997a, p.114).

In his original probabilistic model, Rasch stipulates that ability and difficulty are solved simultaneously and that the probability is described as a ratio of both properties and not either on its own. They are thus calibrated on a common scale (Woodcock, 1999). In fact Rasch (1980) utilises an apt example from physics to make clear his exposition on this point. Maxwell's analysis of the relation between force and mass/acceleration is studied and compared to parameters of a psychological model (Rasch, 1980). In ascertaining the law of force being equivalent to the product of mass and acceleration, the idea of force is formalised simultaneously (mass x acceleration).[49] Parameters follow laws, not the observed entities (behaviour) (Rasch, 1980). Lidz (2003) makes use of an example from age equivalence scores stating that these types of scores are neither ratio nor interval and that blanket comparisons of such scores across ages cannot be done, very much the old adage of a difference of 6 months at age 3 is not quite the equivalent of this same difference at age 18. Added to this is the crucial difference between an object and its properties which seems to have been denied, overlooked or simply not considered an issue worth much attention otherwise early proponents of psychological measurement would not have been so easily lead to assume that all is measurable. The current critical theme is not one of deciding the truth or falsity of intensive or extensive[50] scales of measurement but whether measurement using these scales is indeed applicable at all. Notwithstanding the fact that the entire edifice of psychological measurement is now predicated on these scales of measurement (Kerlinger, 1981), it is contended that this fact is now playing a role in the stale-mate reached within dynamic assessment research.

4.2.3 Summary

The enterprise of mathematical discourse is a flourishing one and is not touted here as being in any manner responsible for what may at times be considered misguided efforts on the psychological measurement front. However, the need to impress upon psychological measurement researchers the foundationalist concerns within mathematics is considered timely. As may often be the case, mathematics is viewed with awe and due acknowledgement of its seemingly unbiased, logical and abstract nature and the whole machinery is often viewed as pristine. In many instances this view is correctly espoused especially in areas which are ideally suited to its methods and results such as theoretical physics. Applied in various real-world contexts it remains, doubtless, unsurpassed in its abilities to bring answers to questions heretofore unanswerable. The context of psychological assessment is an area of concern within the larger arena of subjects to which mathematics can be appointed as problem-solver. The point of concern does not, however, lie with the mathematical method but with how the method is utilised and deployed within the field of human assessment. The foundations of mathematics are riddled with puzzles and conundrums and as with many disciplines it is

---

[45] Or ''numerals'' if nominalists are to be kept happy.
[46] This is not to say that quantification of behavioural phenomena cannot edify to a certain extent when practised in other contexts less influential in terms of affecting people's lives; such as the quantified approach that Murray (2004) utilises in his recent tome on ''Human Accomplishment''. Here, quantification of attitudinal and other qualities may not seem as out of step as at first glance may appear.
[47] Presumably due to the paucity of ordinal statistics and an ordinal-based psychometrics (Cliff, 1991b).
[48] There are models though which question the need to be so reliant on interval and ratio-scaled measures (Chin, 1998).
[49] Note that the extensive measures of force and mass are themselves compound ratios (Asimov, 1993).
[50] Extensive measures (interval/ratio level measure) concatenate units of measures and thus make possible assertions the nature of which include statements such as "*a* is twice the length of *b*"; whereas intensive measures (ordinal level measure but excluding categorical or nominal measures) simply state that "*a* is larger than *b*". Exclusion of intensive measures resulted from the belief that only empirical concatenation resulted from measurable entities or objects (Savage & Ehrlich, 1992).

still to resolve a number of key issues. The veneer of sublime abstraction is a surface which can be easily scratched to reveal cracks and creaks of its own. Mathematics' creaks of course, is not the main consideration within this argument though but plays forth into how it is accepted within other related areas of concern. The last bastion of purity (as is often considered by naïve social scientists) is not as pure and unadulterated as may originally be the perceived case and the maintenance of such a view will only hamper any progress to be made within the quantified fields in social science disciplines. The next section discusses the role of statistics within assessment.

Much early pioneering work within psychology was made functional due to certain mathematically informed predicates upheld by psychophysicists in the fields of vision, learning and information processing. Mathematical psychology espoused certainty and accuracy to a defined level not often seen elsewhere within the psychological domain. Undoubtedly this lent credence to the discipline in the guise of its adorned scientific mantle. Mathematics as a formal discipline had been couched within science and grew markedly in the eighteenth to early twentieth centuries. Coincident to this development was the parallel development of many formal scientific disciplines largely confined to the scientific framework in which theory developed according to certain notions of accepted scientific practice. Psychology as burgeoning science was among these formal developments and so partook in the many accepted notions of the day of what was meant by scientific progress and development on certain fronts. The importation of mathematical rigour was to be expected. That the utilisation of this "pure" technique was considered to be relatively faultless (in comparison to may other disciplines) attested to its acceptance within social domains. It is hardly surprising then that mathematical psychology played forth into various early sub-disciplines within psychology.

The development of formal systems of theory development can be considered as going hand-in-hand with the subsequent development and solidifying system of rigorous proof within mathematics. In an attempt to codify the enterprise of mathematics, strict formalism was seen to be the answer to the proper and objective progress of mathematical science. It is not surprising that the laying down of formal tenets for the development of scientific disciplines was also echoed in the similar progression within mathematics and in onto the area of psychology, specifically psychological assessment. However, this pure and abstract nature of mathematics was to be dealt severe blows in the early twentieth century in which it was evidenced that mathematics, like any formal discipline, has its own share of shortcomings and unresolved issues. None has been more striking than the crushing deflation of the formal axiomatised system of mathematics in the mid-twentieth century. The direct link between mathematics and assessment is not an obvious one but an established link is evident between the two realms and it is considered an important link due to the problematic issues with which psychological assessment is currently saddled. More pertinent to this study's contention is the philosophy undergirding mathematical discourse. Ideas stretch as far back as Plato, Aristotle, Euclid and much later through to Kant and although considerably revised, much of the core concepts still remain within current debates. The link between positivism in science practice and logical positivists can clearly be seen. The co-evolution or co-development of both the enterprise of science and mathematics has perhaps been underplayed by historians of science but becomes manifest and is jettisoned to the foreground when viewing the historical development of the psychological discipline as science. It is not surprising then, that mathematical certainty has worked its way into the efforts and areas of interest which currently hold sway within the social sciences.

The lack of the basis of certainty, the calling into question of the foundationalist assumptions of mathematics and the flagrant attack on certain basic assumptions within mathematics has had repercussions within the discipline, perhaps more so on a philosophical than applied front. The basis upon which measurement resides is then not as solid as once supposed. Are psychologists aware of these issues? Can anything be done to amend the present trajectory of psychological assessment which is predicated on outmoded notions of accuracy and solidity? This brief discussion on mathematics as one element within assessment has attempted to bring to the fore critical issues pertinent to the future development of assessment. By viewing and understanding the background of one of its bases, perhaps assessment can re-assess its course to come.

4.3 Statistical foundation

Prelude

The brief discussion on the reality of pure mathematics ties in closely with the similarly brief rendezvous with statistical endeavours in psychological assessment. Social science's rationalisation of the continued utilisation of statistical techniques in order to buffer and support notions of scientific credibility may start to wear thin if the promised goods are simply not being delivered (accurate identification of potential and prediction of success within psychological testing). As was the case with the view taken at the outset on mathematics above, it cannot be denied that the multitude of statistical techniques developed specifically for use within psychology have been works of wonderment. As with the development of mathematics within science so too did social and behavioural statistics develop within a scientific framework (Stigler, 1999) once again underlining the importance of a need to view the development of science within the social sphere (see chapter 3). How far can quantitative techniques propel psychological science though? It is perhaps not so much that the techniques are in any way flawed as much as it is the application of the techniques which are brought into question (Brown, Hendrix & Williams, 2003; Chow, 1998b; Huberty, 1993). Brown, Hendrix and Williams (2003) question the utility of inference within psychology and ponder the resultant

use of descriptive statistics as sole statistical agent. The difference between statistical hypotheses and the related theoretical hypotheses is also an area which causes confusion among researchers attempting to "prove" theories (Abelson, 1997b) based on the significance of (i) the statistical hypothesis and (ii) maintaining that the result indicates future replicability as well as (iii) making claims on the theory via only one test. Dynamic assessment, as with its intelligence assessment counterpart, are dual testing strategies facing the same statistical treatment but due to dynamic assessment's change-based philosophy and reliance on more qualitative methodology it need not necessarily maintain its position alongside mainstream intelligence assessment. Traditional statistical inferential testing did not set about to prove the probability of obtaining a result as much as it provided a method for ruling out chance (Borsboom, Mellenbergh & Van Heerden, 2003; Harlow, 1997; Howell, 1987). It set out to detect enough evidence to suggest that the effect being tested for is in the direction hypothesised (Harris, 1997). The main contention within the social science literature as it pertains to the utilisation of statistics, is that hypothesis testing has in some measurable way become synonymous with statistics (Nester, 1996). This trend is hardly surprising given the amount of attention that is paid to significance testing and subsequent hypothesis generating research designs within psychology. The major focus of this perennial interest and debate is the emphasis placed on hypothesis testing and not the mathematical statistical argumentation behind numerous and decidedly beneficial statistical techniques. As Nester (1996) states in connection with the continued flawed use of hypothesis testing by professional statisticians "continued association with hypothesis testing is not in our own best interest" (p.408).

As argued in chapter 2, the need for cross-disciplinary approaches towards the study of intelligence may be one of a few feasible methods to be pursued in the twenty-first century. A similar situation regarding the increased use and sophistication of robust statistical and methodological techniques in the area of intelligence and potential might also boost increased access to information about this constructs. Chapter 3 highlighted the trajectory followed by science and social science theory within a scientific framework and when considering the edifice upon which the whole enterprise of science is built, it is hardly worth contesting the role played by statistics and statistical methodology in pushing psychological assessment even further to the fronts of knowledge acquisition. However, the misperception that significance testing is the only manner of hypothesis testing (or Fisher's equating of null hypothesis significance testing with scientific hypothesis testing) and moreover that it is the only true way in which a science can progress, is erroneous (Schmidt & Hunter, 1997). There are other methods (discussed below) which can determine significance of hypotheses other than the traditionally accepted one of null hypothesis significance testing.

When utilised with sound understanding and judgement, many statistical techniques are decidedly helpful within the social science (Abelson, 1997a, 1997b; Chow, 1998b; Harlow, 1997; Nickerson, 2000) and will most likely be around for some time to come (Huysamen, 2005). Why then the need to assess its role in this section? The main reason is its present incompatibility with certain perceptions of what dynamic assessment is and how this manner of assessment seeks to assess change. Various attenuations of statistical methods exist to address these issues (see the section on measurement below) but the nagging question of the need to statistically quantify change persists, not just in the area of dynamic assessment but also within the social sciences as a whole. "It seems safe to assume that many [research psychologists have not had] a lot of exposure to the mathematics on which NHST [null hypothesis significance testing] is built" (Nickerson, 2000, p.246), a sentiment shared by Estes (1997). The general lack-lustre repertoire of superficial robust results within the social sciences can be traced back to the pressure placed upon these subjects to do justice to their existence as independent valid and reliable disciplines. The blame of NHST proliferation has been laid squarely on the "pernicious" hypothetico-deductive method of scientific inference (Rozeboom, 1997, p.336). What role has and does social science statistics play in assessment within the social sciences? This issue is now looked at.

4.3.1 Statistical issues

The core issue relating to the statistical foundation level of this analysis as pertaining to psychological assessment and experiment is the two-fold problem of statistical significance as exemplified by the *p*-value approach of Ronald Fisher (1890-1962) and hypothesis testing as exemplified by the fixed-alpha level approach of Jerzy Neyman (1894-1981) and Egon Pearson (1895-1980) (Huberty, 1993). Also included is the possibly underutilised and undervalued approach of probability theories, among them Bayesian probability, first introduced in the social and behavioural sciences in 1963 (Rupp, Dey & Zumbo, 2004), which is a culmination of other probability methods and techniques among them the binomial probability distribution, after Jacob Bernoulli who applied the binomial distribution as an inverse probability for understanding the observed event[51] (Wright, 1997b). Bernoulli's 1713 treatise on the theory of probability was the first comprehensive account of probability (Boyer, 1991). Inverse probability allowed raw observations to be cast as consequence of the particular stochastic process within stable formulation. But in order to make the step from raw observation to inference one needs to identify the underlying stochastic process through which an inverse probability can be defined and Bernoulli's distribution was the easiest and most widely available tool to do just this (Wright, 1997b). Simeon Poisson's distribution, like Bernoulli's is also a discrete distribution but the outcome is not limited to

[51] The Bernoulli distribution is a discrete probability distribution which assumes a value of 1 (success) with a concomitant probability p and a value of 0 (failure) with its probability $q = 1 - p$ (http://en.wikipedia.org/wiki/Bernoulli_distribution; Everitt & Wykes, 1999).

a choice of 1 or 0; moreover it is a limiting instance of the binomial distribution (Boyer, 1991). Poisson's compound distribution is the "parent of all useful measuring distributions" (Wright, 1997b, p.34) and the natural parameter for the Poisson distribution is a ratio of an object's location and the measurement unit of the instrument used. This formulation preserves concatenation and divisibility resulting in different units implying the same location (Wright, 1997b). As will be discussed below, Fisher took another step in 1920 by developing the likelihood version of inverse probability in an attempt to construct maximum likelihood estimation (utilised within IRT models) (Wright, 1997b). Is it interesting to note the development within probability statistics throughout the seventeenth to twenty-first centuries and how researchers have plied their tools upon the trade making it feasible to use numbers in a manner aiding inference. Nevertheless, continuous debates ranging over four decades within the psychological literature pertaining to statistical inference[52] and hypotheses testing led to the resultant Task Force on Statistical Inference (TFSI) being instituted by the Board of Scientific Affairs. This was convened under the American Psychological Association (APA) to deliver a framework offering guidelines as how best to pursue various contentious issues within statistical psychological reports (Krantz, 1999; Wilkinson & TFSI, 1999). At the outset, it must be made clear that the historical development of statistical techniques within the social sciences is not a straightforward history and that issues such as the utilisation of hypothesis and significance testing is a perennial debate. Dynamic assessment, as scientific method of change assessment is lodged between various approaches and philosophies towards such assessment. On the one hand a wholly qualitative and clinical approach towards change inducement and assessment is at odds with the more robust quantitative approaches of change assessment. Both seek to bring about measurable change within the individual in the assessment situation which can last from an hour to many years.

Measurable change is accounted for in terms of numerical differences between pretest and posttest scores and also by clinical assessments of individuals over a lengthy period of time. The role played by statistics within all of this can vary from purely descriptive statistics enabling the researcher to gauge the starting level of competence through to the level at which skills have either been improved or not; to inferential statistics which enable the researcher to infer and possibly predict outcomes of various interventions efforts and future performance or behaviour within any one domain or various domains. Typically, as with much of the literature on psychological assessment, research designs employ hypotheses as points of departure leveraging conclusions on design and method which can vary across studies. Questions which can be posed about dynamic assessment stretch far beyond what the method does and how it accomplishes what it sets out to achieve. In order to address the statistical influences on dynamic assessment as method of change assessment fundamental issues need to be addressed. Such core issues have received attention in chapters 2-4.

Accusations lodged at any particular approach of psychological assessment need to be very mindful of the development of the method within the greater realm of psychology, for it alone is not wholly responsible for what perhaps may seem to be misguided efforts on a number of fronts. Issues pertaining to how the brain and mind is viewed by psychology and other related disciplines has been discussed, along with issues involving the topic of environmental and genetic contributions towards behaviour. Core philosophies are what mould the approach. The approach then is housed within a context of scientific growth and change and is dominated by the settings of scientific discourse. Many issues within this context are yet to be resolved. These issues were highlighted in chapter 3. Dynamic assessment, as manner of change evaluation, is not alone in terms of being ineluctably drawn into and being overwhelmingly influenced by historical trends - it is by nature (as with many disciplines) a contingent sub-discipline. Issues that plague dynamic assessment are issues which plague other disciplines. There is no quick and simple answer as how best to approach such myriad problems. Scrapping statistical inferential techniques along with the traditional practice of hypotheses testing is hardly a path to be followed. The point of building a case in chapters 2-4 is to illustrate the entanglement of dynamic assessment as method of change evaluation and to highlight the need to pry apart the various layers of historical impingements. Dynamic assessment is firmly grounded in solid scientific technique and will not be easily extricated from its foundations.

The TFSI broached the debate on the applications of significance testing and stimulated discussion of the topic in their 1999 article in American Psychologist, concluding that the APA revise their statistical sections in their publication manuals (Wilkinson & TFSI, 1999). The summarised guidelines which are relevant to this section are set forth below:

- clarity of study aims is sought by refraining from "cloaking" the study in a guise that does not apply to it - are the studies hypotheses generating or hypotheses testing?
- the measurement variables described need to remain consistent throughout the study. Measurement precision is not, however, synonymous with operationalisation

[52] The forerunner of statistical inference is the probable error (PE) of a mean which was extended from the PE of test scores and was first utilised in 1910 (Huberty, 1993) but Nester (1996) maintains that it was used as early as 1840 in the area of biology. Probability per se, though was evident in work by 1713 (Wright, 1997b). Probability's role is paramount within the social sciences specifically.

- utilise effect size estimates when reporting *p* values and depending on the nature of the units of measurement, the effects sizes can be either standardised or not[53]
- interval estimates of effect sizes should be stipulated

The article, although addressing issues of concern to the psychological research community did not really offer guidelines radically different from guidelines offered in the past but concluded that "statistical methods should guide and discipline our thinking but should not determine it" (Wilkinson & TFSI, 1999, p.603). This sentiment can likewise be lauded as the prevailing sentiment regarding dynamic assessment and its acceptance within the broader assessment arena.

4.3.1.1 Historical contingency

Dynamic assessment, as with many sub-disciplines within psychological assessment and psychology as a whole, is bound by traditionally accepted modes of operation, research design and results-reporting and deviations from these norms of scientific practice may often be judged as flaws. The rigid application of NHST can be seen to be ritualistic (Huysamen, 2005; Nester, 1996) and even a bad habit (Shrout, 1997). The struggle that dynamic assessment in its clinical mode faces is one common to other areas within psychology which too are not necessarily amenable to statistical inference as this is simply not the right mode for their existence as mode of knowledge-acquisition. The prevailing context of scientific progress needs to be considered (as has been in chapter 3) as well as fundamental philosophies pervading the history and origins of certain key points of departures (as evidenced in chapter 2). Why has psychology "progressed" to this point in time in which statistical technique plays a greater role than it perhaps should? Once again, not much can be said to discredit many statistical techniques, and this has never been the aim of this defensive treatise on the subject but a common theme thus far established in this study is that the application of these tools into an area not amenable to such tools is perhaps ill-conceived.

Much of social sciences statistics emanates from issues surrounding probability and chance and until the advent of the twentieth century, there were no formally directed research endeavours to develop the methodology of significance testing beyond Laplace's 'probability of causes' which relied on Bayes' rule (Fisher, 1958; Gigerenzer, Swijtink, Porter, Daston, Beatty & Krüger, 1990; Mulaik, Raju & Harshman, 1997). Incidentally, Laplace was instrumental in rescuing the work on Bayes' inverse probability and it is due to Laplace that the theory of probability grew as it did (Boyer, 1991). The theory of errors, unlike the probability of causes was a more developed branch of statistics (Fisher, 1958) and found its utility in the estimation of errors in areas like astronomy (Quetelet, as mentioned in chapter 2 brought over ideas from astronomy into the social sciences). The determination of errors in observations in applied areas of work as well as the minimisation of these errors played a large role in the development of the statistical techniques which psychologists currently employ (Gigerenzer et al., 1990) and many statistical techniques were originally developed for deployment within the social sciences (Porter, 1997).

The general aim of evaluating hypotheses based on data is quite an old one, is associated with the name of Bayes and dates to 1710 (Huberty, 1993; Nester, 1996; Neyman & Pearson, 1933). It was epitomised in the early twentieth century by the combined efforts of Karl Pearson, Ronald Fisher and W.S. Gosset. Fisher's[54] work in the fields of agronomy,[55] physiology and medicine; Gosset's work in the Guinness brewery in Dublin and Pearson's work in biometry (and thus founding the basis of psychometrics; Porter, 1997). These efforts are telling of the need to impose order and a mechanised framework on data collected in the field (Fisher, 1958). Hypothesis testing was, however, sporadic during the nineteenth century and only came into its own after the publication of Gosset and Fisher's various works (specifically the *t* test, the Chi square test and *F* test; exact distributions which were developed) (Lehmann, 1992). Moreover, Lehmann (1992) maintains that Fisher drew attention to hypothesis testing only as incidental. Kendall (1978) dates the utilisation of statistics, as is understood today, to 1660 and thus excludes an already lengthy history of enumeration for purposes other than for estimation and prediction; the terms in which he ensconces statistics. Kendall's (1978) main emphasis shared by Lazersfeld (1977) illustrates what prompted the development of statistics concerns;

---

[53] Appendix 1 utilises standardised effect sizes as the measured variables differed in most of the studies which were meta-analysed.

[54] Fisher, like Galton was a polymath and there aren't too many subjects on which he did not lay his hands (Stigler, 1999). Fisher's driving motivation behind some of his statistical work was eugenics, which was also Galton's favourite. Egon Pearson, Karl's son worked at the Galton laboratory in London. Small world indeed. Prominent figures cannot however be divorced from the times in which they lived and worked. Mendelian inheritance, evolutionary theory, the eugenics movement, the development of intelligence assessments and utilisation of newly developed statistical techniques must have been terribly exciting in the early days. Especially when pioneers started to weave together all the threads. Are there any such pioneers today? The world has become too complicated for polymaths to flourish quite as successfully as they once did. The history of intelligence testing, the progression of psychological science and statistical technique and methodology can all be traced to a number of individuals working in-and-around similar issues of the day. The Galton-Darwin-K.Pearson-Fisher-Neyman-E.Pearson link is particularly evident within this subject area. Contrast this historical line to early works within dynamic assessment (see chapter 5). No wonder such a chasm is evident; core philosophies cannot be more widely dispersed! Moreover, the basis of mathematical statistics was closely associated with research in the areas of biological evolution, heredity and the eugenics movement (Porter, 1997).

[55] No pun intended. Just as an aside, it is quite humorous reading through Fisher's books (1956, 1958) in terms of some of the examples chosen for his explanations of statistical techniques. For instance he cites examples such as "comparison of relative growth rate of two cultures of an alga" (1958, p.140); "effect of nitrogenous fertilizers in maintaining yield"(1958, p.136) and particularly pertinent to this thesis "frequency of criminality among the twin brothers or sisters of criminals"(1958, p.94).

namely politics, which fuelled the need for arithmetic which could aid in the most judicious use of resources and as a means of controlling taxes. Huberty (1993) dates the first statistical approach to the English scholar John Arbuthnot in 1710 in which the gender ratio of births was the main concern. Stigler (1999) maintains that the history of statistics, at least since the seventeenth century, is one of a collection of ad hoc and miscellaneous mathematical tools dealing with data, the process being governed by an evolutionary pruning of sorts within the tradition of scientific enquiry.

Fisher also helped introduce the idea of the experiment which, until 1910, had not been associated with the use of statistics (Gigerenzer et al., 1990). The combination, then, of experimental design and statistical inference is a relatively new one and was of course adopted very early in the formal development of psychology as scientific discipline. As chapter 3 highlighted, the growth of psychology as science as formalised discipline is everywhere evident, none more so than in its methodology. Rigid deployment of Fisherian statistics, some might advocate, has served to hamper the development and progress of a psychology equally relevant to individuals as to groups. Early twentieth century social science statistics, it must be recalled, cantered on detection of error and probability of distributions of groups; groups involving crops, diseases, measurements, stars and stout among other areas of concern.[56] The development of positivist influenced behaviourism (the philosophies of the Vienna Circle were at their heights; Oakes, 1986) with the emphasis on sensory data in the 1940-1950's; the synonymy of randomised experimentation (Fisher) with treatment and control groups and the "institutionalization of inferential statistics in American experimental psychology" (Gigerenzer, 1991, p.255; Stigler, 1999) pushed psychological testing on a course of robust scientific technique even though significance testing as technique within psychology was critiqued as early as 1955 (Schmidt & Hunter, 1997). Prior to the 1940's very few psychological articles published statistical reports (Kline, 2004). This was also the time during which Stevens published his work on the scales of measurement which have become the mainstay of measurement scaling in psychology (Stevens, 1946, 1951b). Stevens was considered a "middleman", voicing indirectly, opinions from the logical positivists, especially those of Carnap (Michell, 2003). Pure statistical techniques were never developed for application in areas as diverse as psychology, education and economics (Porter, 1997) but have inevitably been applied and misapplied in these varied contexts (Rossi, 1997). In other words these ingredients were what made and makes psychology a science. The Fisher-Neyman-Pearson school of statistics became the orthodox approach to follow (Oakes, 1986). Unfortunately not all sub-disciplines within this subject avail themselves of such mechanisation. Michell (2001) adds that "if no attempt is made to test a hypothesis, then there is no adequate scientific reason to accepts it. Thus, mainstream psychologists adopted the rhetoric of measurement because of the political advantages to be gained for the discipline, not for any adequate scientific reason" (p.214). The author's already stated opinions on philosophical affiliations in chapter 2 highlighted the leaning towards positivist influenced methodology.

This may seem contrary to the critique to be levelled at null hypothesis significance testing. However, it is not the method per se that is being criticised but the incorrect utilisation of the method in an area of psychological research which does not necessarily avail itself of the method's techniques.[57] However, during the 1960's a noticeable decline in positivist strategising and the upsurge in Bayesian statistics heralded a new era in technique (Oakes, 1986) and has been utilised within the cognitive domain, where the need for less time-intensive methods calls for more efficient models (Neufeld, 2002). Strides in modern test theory were also being made during the 1950's in contrast to techniques utilised in traditional classical test theory, even though the requisite computational power was not yet at the disposal of statisticians (McDonald, 1999). As with statistics, being an applied branch of mathematics, test theory is also considered a branch of applied mathematics and for this reason the co-occurrence of trends within statistics and test theory is not surprising. Simultaneous developments in mathematical statistics flowed over into both realms.

---

[56] Early descriptive statistics (although not formally recognised as such) concentrated on the counting of people within countries and can be related to the need, as discussed above in the section on mathematics, to count and measure. As early as 1532 the number of deaths in London was counted on a weekly basis (Donnelly, 2004), although tallies of deaths, citizen numbers and accounts for actuarial reasons go back as far as ancient Rome (Kendall, 1978). It seems that this propensity to count, measure, statistically infer and mathematically manipulate are fundamental needs in the world in which we live. Melding the inclination to measure and count along with the rigour of science discourse makes it seem almost natural that assessment should have proceeded along the course it has; however ill-fitting this course has proven to be.

[57] Once again, the author reiterates the original stance promoted at the outset of this study. Psychology has to decide where in the spectrum of formal science it should be placed. "Soft" areas of concern are fine. "Hard" areas of concern are also fine. However, the constant overlapping of methodologies and statistical choice of data analyses is pressurising these vastly different areas of concern by moulding them into structures that are anathema to their core philosophies. Soft psychology (clinical and counselling among others) can function entirely on their own in their own methodological manner without necessitating the use of hard core psychological technique. The issues need to be separated. There is nothing wrong with such a separation but holding the whole of the discipline hostage to a specified approach is suicidal. Unification via subject matter or separatism via estrangement of method and technique are two ways in which this can be broached.

Fisher vs. Neyman-Pearson - prelude to null hypothesis significance testing[58]

Both the Fisherian and Neyman-Pearson (after Jerzy Neyman and Egon Pearson, son of Karl Pearson)[59] schools are frequentist or classical school and the Bayesian school (after the Reverend Thomas Bayes; 1702-1761) is subjectivist and is often referred to as the common sense approach to statistics, preceding work done in classical statistics[60] (Edwards, Lindman & Savage, 1963; Harlow, 1997; Kline, 2004; Mulaik, Raju & Harshman, 1997; Pruzek, 1997). However, Reichardt and Gollob (1997) are reluctant to equate Bayes theorem with the subjectivist approach as one need not be a subjectivist to utilise Bayes. Fisher's approach is synonymous with statistical testing whereas Neyman-Pearson's approach is synonymous with hypothesis testing (Reichardt & Gollob, 1997). Crudely then, the two issues have, throughout the years, become melded into one erroneous method of statistical inferential logic. The Bayesian approach has not been as influential in social science statistics as the former, even though Karl Pearson had occasionally made use of Bayesian assumptions (Gigerenzer et al., 1990). The Neyman-Pearson school has overshadowed the Bayesians even though the latter regularly rears its head (Lindley, 1992). In essence, Bayes' theorem relies on the availability of prior distributed probabilities which conventional significance testing does not (Oakes, 1986). These priors vary over a number of hypotheses which is perceived to be its biggest flaw (DuMouchel, 1992; Moore, 1997) as initial priors are estimated,[61]leading to less exact measures and based largely on belief (usually expert or well informed belief) (Kline, 2004; Neufeld, 2002; Rindskopf, 1997; Trafimow, 2003) and are thus unconditional (Killeen, 2005). The theorem is an instance of normative decision theory which necessitates base rate information which is combined in the probability of future events; failure of which leads to the use of heuristics as guiding factors[62] (Kleiter, 1994). Due to the probable nature of Bayesian statistics, the conditional priors are set to a limit of 1. This entails the setting up of a parameter which necessitates certainty on the part of the assessor. However, being certain about such a parameter also leaves room for being wrong about it resulting in what Borsboom, Mellenbergh and Van Heerden (2003, p.209) aptly state as being very difficult "to be wrong about something if one cannot be right about it". Parameter estimation[63] requires a true value, one which is not available even within prior conditionals. It is perhaps prudent to think about how dynamic assessment might avail itself to the use of Bayesian statistics in a pretest-posttest scenario in which prior probabilities of success or failure or even growth are taken into account when assessing the likelihood of improvement over mediatory interventions. Oakes (1986) points out that the historical antecedent to this personal belief in priors is in fact rational belief which is measured in an objective manner. Nevertheless, this is in contrast to conditional posterior probabilities in which unknown population parameters are considered random variables as opposed to their unknowable yet fixed status in the frequentist paradigm (Rupp, Dey & Zumbo, 2004). However, as more data is gathered the original subjectivist choice of prior probability is partialled out over successive iterations and the new data serves to update prior beliefs (Pruzek, 1997; Rupp et al., 2004) or revises the rule based on new information (Majumdar, 2004; Walliser & Zwirn, 2002). Surely this can in some manner attest to a happy medium in which dynamic assessment specifically can function? Probability can be interpreted in two main ways depending on the subscription to either frequentist or subjectivist approaches (Reichardt & Gollob, 1997). These differences are illustrated in table 11.

---

[58] It has been contended that mathematical statisticians do not see what all the fuss is about in terms of both approaches (Huberty, 1993). Mathematically speaking, the two techniques are similar and yield useful information but it is the misinterpretations stemming from misuse of the approaches that niggles some research psychologists.

[59] The relationship between Fisher and the duo Neyman and Pearson was strained to say the least. They did not personally like each other one bit (Kline, 2004).

[60] Bayesian inference is perhaps the newest technique added to its already extensive repertoire (Edwards, Lindman & Savage, 1963).

[61] Very similar to face validity actually where expert opinion is called on to make certain judgements about items in a test or questionnaire (Anastasi, 1988; Murphy & Daivdshofer, 1998; Reber & Reber, 2001; Rust & Golombok, 1992; Smit, 1996). This is the first place to start; how "objective" is that? Bayesian prior estimates are usually doing the same thing! Reliance on expert opinion and committee members' experience play a decisive role in determining what is and is not included in a prior probability distribution (Rupp et al., 2004). The functional form of the prior probability distribution is of course taken into account, i.e. binomial probabilities for success will vary between 0 and 1. However, caution is attached to these decisions especially if the sample size is small. Larger sample sizes allow the data to dominate posterior distributions and previous personal opinion does not carry as much weight as it did before (Rupp et al., 2004). Perhaps this subjectivist prior approach is appealing in areas such as intelligence assessment in which prior expertise is used to determine intuitive concepts of what intelligence as a construct purports to measure (Rupp et al., 2004). As Dawkins (2006, p.106) states "Bayes' Theorem … is a mathematical engine for combining many estimated likelihoods and coming up with a final verdict, which bears its own quantitative estimate of likelihood. But of course that final estimate can only be as good as the original numbers fed in …. the GIGO principle is applicable here…".

[62] Cognitive heuristics are themselves very powerful and perform adequately in many instances; think of stereotyping (Augoustinos & Walker, 1995). On the face of it, treating individuals as if they were the mean stereotype is clearly absurd if not unethical; but this method of survival has performed an invaluable service! However, within psychometric testing, the use of heuristics is considered untenable due to its unscientific stature as decision-maker thus the need for verifiable estimates such as offered through Bayes (Martin & VanLehn, 1995).

[63] Parameter estimation and hypothesis testing are the main goals within a statistical rendering of psychological reality and one will never escape this it seems, for even advanced forms of mathematical modelling of responses to items on tests are preoccupied with such issues (Van der Linden, 1994).

Table 11 Differences in probability definitions within the frequentist and subjectivist approaches (Reichardt & Gollob, 1997)

Definition of probability	
Frequentist / Classical	**Subjectivist / Bayesian[64] / Inverse**
• Both agree that probability be associated with mathematical axioms of probability	
• Probability of a repeatable event is asymptotically relative to the frequency of its occurrence over unlimited repeats (under identical circumstance bar random variation)	• Probability is the prior belief in the occurrence of an event in nature. The state of nature is dependent on the observer
• Probability of a finite occurring event is either 1 or 0	• No differentiation between repeatable and finite events being probable
• Probability is independent of human cognition	• Probability is dependent on human cognition and this also varies across observers
• Prior probability is determined by the population from which the sample is taken ♦ *Uniform prior* distribution yield parameters that are all equally likely (non-informative)[65] ♦ For the frequentist this would mean that the randomly chosen sample was drawn from a randomly chosen population (across populations) ♦ *Nonuniform prior* distribution (any prior distribution that is not uniform and informative; such as a *beta* prior which is suited to binomial estimates as well as the normal distribution) ♦ For the frequentist this would mean sampling from a randomly sampled population in which the population parameters themselves are normally distributed ♦ *No usable prior* distribution (when there is simply no information available or too little information is given). The frequentist would have no usable prior distribution if a random sample is drawn from a randomly drawn population from which the population parameter is unknown • Underlies the Neyman-Pearson school of statistics. Inference to the parameter is made via analogous findings from sample statistics. Decisions are made prior to data collection and inference follows automatically once the data is collected	• Prior probability is determined on subjectivist yet expert beliefs ♦ *Uniform prior* distribution yields parameters that are all equally likely (non-informative) ♦ For the subjectivist this would mean that a random sample was drawn from a population in which all possible population parameters were considered as equally likely ♦ *Nonuniform prior* distribution (any prior distribution that is not uniform and informative) ♦ For the subjectivist this would mean randomly sampling from what is thought to be a normally distributed population of population parameters ♦ *No usable prior* distribution (when there is simply no information available or too little information if given). The subjectivist would have no usable prior distribution if the uniform distribution was thought to be incorrect but the shape of the nonuniform distribution could not be completely specified • Underlies the Bayesian school of statistics. Inference to the parameter is made directly. Competing rival hypotheses are evaluated in tandem with the gathered data

The theorem is a method for taking into account the conditional probabilities when estimating the probability that an hypothesis is true (Reber & Reber, 2001) as well as determining the probability that the sample effects are worth further investigation (Harlow, 1997). In utilising prior information (unconditionals) with empirical data (data collected), posterior distributions are generated (likelihood of results) which are utilised as the basis for statistical inference (Pruzek, 1997). Bayes' theorem has been successfully used within economics, military and combat strategies as well as artificial intelligence for instance (Gigerenzer et al., 1990; Majumdar, 2004). Mathematically, the joint or combined probability of data and the hypothesis is the product of probability of the data and the conditional probability of the hypothesis, given the data (Anastasi, 1988). Reluctance to use Bayesian statistics may be partially due to its non-positivist tenets, which as has been discussed in chapter 3 does not make for good science progress, as perceived by positivist classical significance testing (Barlow, 1992; Edwards, Lindman & Savage, 1963).

[64] Recall that not all subjectivists ascribe to Bayes' theorem. The two are not synonymous but have both been included here for purposes of broad-based comparisons.
[65] Coined "the principle of insufficient reason" by Laplace and "the principle of indifference" by Keynes (Oakes, 1986). If there is no information as to the likelihood of an event given a spread of probabilities for events, one assigns equal probabilities to all such events.

Expressed in more compact fashion after Kline, 2004:

$p(d \wedge h) = p(d)\, p(h \mid d) = p(h)\, p(d \mid h)$[66] (i)

which if solved for the conditional probability allows for the expression to be written out as Bayes' theorem:

$p(h \mid d) = p(h)\, p(d \mid h) \,/\, p(d)$[67] (ii)

Equation (i):

- $p(d \wedge h)$ is the conjunction of the data and the hypothesis which is the product of
- $p(d)$ which is the probability of the data and
- $p(h \mid d)$ which is the likelihood[68] or posterior probability (and hence conditional)

Equation (ii):

- $p(h \mid d)$ is the posterior probability of the hypothesis given the data which is determined by the
- $p(h)$ and $p(d)$ which are the prior (unconditional) probabilities of the data (regardless of the truth of h) and the hypothesis (its probability before the data are collected) and
- $p(d \mid h)$ which is the likelihood or conditional probability of the data given the hypothesis (similar to the p value in NHST)

Fisher contended that knowledge of prior probabilities was rarely the case and argued against Bayesian probability (Geisser, 1992; Krueger, 2001; Mulaik, Raju & Harshman, 1997; Oakes, 1986) and maintained "that the theory of inverse probability is founded upon an error, and must be wholly rejected" (Fisher, 1958, p.9). As a frequentist, he preferred sampling from known existing populations and not relying on probabilities of these populations parameters, in other words he preferred certainties and subsequently Bayesian inference was largely ignored within early psychological research (Good, 1992). It resurfaced in 1937 in a paper by De Finetti and has since been intermittent in its influence within the social and behavioural sciences as well as statistics (Barlow, 1992). Fisher is known for his views on sufficient statistics which is the notion that a quantity x depends on an observable random variable x but not on an unobservable parameter θ. It is sufficient if that statistic captures all of the information in x that is relevant to the estimation of θ (http://en.wikipedia.org/wiki/Sufficiency_%28statistics%29). In other words, a statistic exhaustive of information pertaining to its modeled parameter (Wright, 1997b).

One can see the logical reasoning of Fisher given his advocacy of experimentation as research technique. Neyman and Pearson agreed with Fisher regarding the use of Bayesian assumptions but the former also believed that Fisher's work was not entirely deducible from first principles and was also not logically coherent (Oakes, 1986).[69] Neyman and Pearson added the alternative hypothesis to their theory incorporating the probability of accepting an hypothesis when it is false thus attempting to make more complete, practically applicable and consistent Fisher's original test of significance (Gigerenzer, Swijtink, Porter, Daston, Beatty & Krüger, 1990; Harlow, 1997; Howell, 1987; Huberty, 1993; Kline, 2004) but in so doing they were vehemently opposed by Fisher (Cohen, 1990). The Neyman-Pearson approach formalised the accept-reject theory of hypothesis testing (Fraser, 1992) which is currently the more generally accepted form (Nester, 1996). Optimum decision rules were to be enforced so as to allow for acceptance or rejection of hypotheses, which was contrary to Fisherian and Bayesian ideas of testing with the aim of informing belief (Oakes, 1986). To determine if there is a difference between the dynamic assessment performance of an experimental group who has received mediation and a control group who has received no mediation the null hypothesis is in fact tested, i.e. assuming that both groups are equal in terms of having gained something which might on the surface seem odd seeing as that is not what one is testing but rather the alternative hypothesis (Cohen, 1990, 1994). Moreover, significance testing is only inferential in as far as the associated probability is concerned and is decision-theoretic as far as its significance levels go (Oakes, 1986).

---

[66] For $p(h \mid d)$, read: " the probability of the hypothesis given the data".

[67] Making use of two random events A and B, the equation is exactly the same: $p(A \mid B) = p(A \text{ and } B) \,/\, p(d) = p(B \mid A)\, p(A) \,/\, p(B)$ (Rupp et al., 2004).

[68] Likelihood is proportional to probability. In other words if $p(A \mid B) = p(B \mid A)\, p(A) / p(B)$; then $p(B \mid A) / p(B)$ are the likelihood functions of $p(A \mid B)$. The likelihood of hypothesis A given the data B is proportional to the probability of the data B given hypothesis A. Hence $L(A \mid B) \propto P(B \mid A)$ assuming $\alpha > 0$ (Edwards, Lindman & Savage, 1963; http://en.wikipedia.org/wiki/Likelihood; Oakes, 1986). $L(A \mid B) = p(B \mid A)$. But $p(A \mid B) \neq p(B \mid A)$. Likelihood cannot be interpreted in a probable manner. Miller (2004) states succinctly: probability = knowing parameters and hence predicting the outcome; likelihood = observation of data and hence estimating the parameters. Hence, "if the probability of an event X dependent on model parameters p is written $p(x \mid B)$ then we would talk about the likelihood $L(p \mid x)$" (p.26).

[69] An echo perhaps of the axiomatisation of mathematics during the formalist period? It is interesting how similar the paths taken by different subject areas really are.

What researchers want to know is the probability that the null hypothesis is true given the observed data. NHST focuses on the given of a null hypothesis and then tries to determine the probability of obtaining the observed data, i.e. p (d ¦ H_0) (Cohen, 1994; Huysamen, 2005). Fisher maintained that H_0 was to be nullified and the alternative hypothesis to be debunked (Hunter, 1997; Porter, 1997) or at the very least to have judgement suspended (Howell, 1987) if there is insufficient evidence to reject the hypothesis. This is a scenario often playing out within the natural sciences (Rossi, 1997). Suspending judgement about the alternative hypothesis (which Fisher never employed in any event) does not mean that the null is true. Fisher chose, rather, to replicate his experiments (Gauch, 2006); an aspect neglected in the social sciences. The greater the number of replications the greater the likelihood of accuracy in findings because there is a concomitant decrease in chance variance, akin to power. Current erroneous logic has since led to thinking that H_0 is in fact zero; hence the "nil null hypothesis" (Cohen, 1994; Huysamen, 2005; Nickerson, 2000) even though it is merely nonsignificant (Schmidt & Hunter, 1997). Compounded by the tendency of researchers to equate observed differences to true differences (Nester, 1996) the tradition of NHST has been seriously questioned and its use as a tool of science has also been scrutinised.

Theoretically then, Fisher would suspend judgement based on an experiment which yielded a non significant result but Neyman-Pearson would make a definitive decision as to the outcome or conclusion (Howell, 1987) a situation which has persisted in psychological practice and is quite pervasive in introductory psychological statistics textbooks (Huberty, 1993). Fisher (1958) even states that some researchers and academics may have misread him in prior publications on this issue and reiterates his stance on the process of nullifying the hypothesis saying that this process is "part of the hypothesis to be tested" (p.125). This brings into conflict the whole idea of scientific progression based on refutation of the alternative or research hypothesis as advocated by Popper and the idea propounded by Fisher; that of inductive inference via rejection of the null hypothesis (Cohen, 1990; Meehl, 1978). Recall Meehl's preoccupation with the progress of psychology as a scientific endeavour by seeking to value the affirmation of theoretically predicted research results from practice as opposed to merely stating groups' statistical significant differences at a prior determined probability level (Ray, 2006). Citing Huberty (1993, p.318) the following step-wise discussion (table 12) highlights in brief the difference between the Fisher and Neyman-Pearson approaches to significance and hypothesis testing:

Table 12 Main difference between Fisher's P-value based approach and Neyman-Pearson's fixed alpha approach (Huberty, 1993, p.318; Kline, 2004)

Fisher significance testing or P-value based approach	Neyman-Pearson hypothesis testing or fixed-alpha approach
1. State H_O	1. State H_O and H_A
2. Specify test statistic (T) and referent distribution	2. Specify test statistic (T) and referent distribution
3. Collect data and calculate value of T	3. Specify α value and determine rejection region (R)
4. Determine P value	4. Collect data and calculate value of T
5. Reject H_O if P value is small; otherwise retain H_O (but suspend judgement on H_A; Howell, 1987)	5. Reject H_O in favour of H_A (in other words accept H_A; Howell, 1987) if T value is in the rejection region; otherwise retain H_O

The arbitrary choice of zero as a prior probability for H_O cannot account for cumulated knowledge in the discipline (Killeen, 2005). How does Popper's rationalisation of the attempts to falsify theory work within NHST? The predictive variable (one which upholds the theory) is set up as H_0 and attempts at rejection of H_0 leading to a challenge of the predictive theory. This is referred to as the "strong" form of testing (Cohen, 1994; Meehl, 1997) or the "acceptance-support" form (Kline, 2004; Nickerson, 2000). It is Popper's philosophy in NHST action (Meehl, 1967) and what Chow (1998b) refers to as repeated attempts at falsification which results in a convergence of operations over a number of such attempts. The alternative hypothesis is what the researcher wants to try and disprove (Kline, 2004). Its opposite incantation is, however, decried as "weak" when theories are confirmed by the rejection of H_0 or the "rejection-support" contention (Huysamen, 2005; Meehl, 1997; Steiger & Fouladi, 1997)[70] where H_A is the researcher's alternative hypothesis illustrating the chosen theory (Kline, 2004). The strong/weak issue is more an epistemological and logical issue as opposed to a statistical one (Meehl, 1997). This is what Trafimow (2003) refers to as researchers setting up obvious hypotheses to be rejected when in fact they should concentrate on testing nonobvious hypotheses (although this makes life at present difficult for those seeking article publication, as H_O acceptance is not "good science" as far as current ideas on psychological science go; Hunter, 1997. See chapter 2 section 2.4.5.3.6 where this very issue is cited as militating against the growth and progress of psychology as a formal discipline). Meehl (1989) begrudgingly refers to this process as hypothesis concoction seeking "to preserve a theory from falsification" (p.36) and the "weakness of null hypothesis *refutation* as a means of corroborating psychological theories" (original emphasis) (Meehl, 1998, p.5, 1990). See Table 13 for a clearer exposition of the weak and strong renditions of NHST.

[70] Thompson (1988) in Nickerson (2000) asks the question of how many software packages in fact test for the alternative hypothesis as most concentrate on the "no difference" relationship.

Table 13 Alternative forms of NHST

NHST weak and strong forms (research designs predicated on both deduction and induction)	
Weak	**Strong**
▪ Reject-support contention ▪ H_O to be rejected ▪ H_A is the researcher's theory ▪ Researchers take care to make α as low as possible due to the greater likelihood of falsely rejecting H_O. Thus emphasising Type I errors and increasing the power of the test ▪ Sample size is important ▪ Over-reliance on point nil null estimates hypotheses. Unable in most cases to generate point hypotheses due to the weak form of ad-hoc theories ▪ Makes the goal easy to obtain ▪ Typifies the social sciences via a process of verficationism ▪ Seemingly objective veneer of science practice ▪ Sets the scene for rejecting an already likely null hypothesis thus offering nothing new in the way of information ▪ Uninformed social science researchers are either unwilling or unable to change this scenario for fear of non-publication and/or due to ignorance of the philosophy behind the weak form of NHST ▪ Should start to present findings which are non-significant. Meta-analysis goes some way in rectifying this situation by including the "file drawer" calculation	▪ Accept-support contention ▪ H_O to be accepted ▪ H_A is not the researcher's theory ▪ Researchers should concentrate more on Type II errors as there is a greater likelihood of falsely accepting H_O. Hence α should not necessarily be set too low. Power is inextricably linked to NHST. Increasing power would necessitate increasing sample sizes but to obtain prior levels of power sample sizes become increasingly difficult to obtain ▪ Sample size can work against the research hypothesis ▪ Does not necessarily rely on point nil null estimates hypotheses but rather a range or composite; in many instances relying on point estimates thus making the job of acceptance-rejection more clear cut. This is possible only due to the theory's already strong status ▪ Makes the goal difficult to obtain ▪ Typifies the natural science approach towards disconfirming the research hypothesis; a trend of falsificationism. Note that natural sciences thrive despite the non use of NHST ▪ A more objective form of science practice than is offered by the weak form ▪ Sets the scene for a stringent application of nullifying the research hypothesis ▪ Informed researchers as well as editorial boards see the light in terms of assessing the progress of a discipline by means of setting up more stringent hurdles. Publication of negative results needs to be taken more seriously ▪ Should encourage further publication of negative or non significant results

Physics, for instance, proceeds by setting itself ever-increasing standards of severity (invoking the strong form of testing) whereas the behavioural and social sciences seek the way forward by relying on less severe although scientifically feasible standards (and hence invoking the weak form of testing) (Meehl, 1967). Recall in chapter 3 though that Meehl (1997) has stated that psychology is far from the stage of Popperian falsificationsim in terms of its mechanism of growth – it is too immature to proceed on the basis of refutation alone. It seems Meehl has, since 1967, tempered his opinions regarding the utilisation of this approach; in fact he as much admits his staunch Popperian views which have since 1967 mellowed (Meehl, 1998). Likewise falsificationist NHST (attempting to falsify H_A as opposed to H_O) is thus not quite yet viable nor feasible at this stage (Krueger, 2001). In addition, Huysamen (2005) mentions the ill-preparedness of statistical courses (specifically in South Africa) in dealing with a change in NHST and other related areas such as introducing Bayesian statistics (Moore, 1997), due to the unfamiliarity of the technique within the social sciences even though it was introduced into psychology in the 1960's (Kline, 2004). The author is not convinced of this retort as it does not even attempt to rectify the situation. It is logical that people are unfamiliar with some aspects, but most people are unfamiliar about most things until they learn them!

The p value is an exact level of significance statistic from Fisher (but recall that he only utilised H_O); fixed levels of α across studies is the Neyman-Pearson approach which represents values for the probability of a Type I error and β the probability of making a Type II error (Kline, 2004; Oakes, 1986). This hybridised approach towards statistics utilising Fisherian statistics (historically preceding Neyman-Pearson by approximately ten years; Huberty, 1993) along with the Neyman-Pearson school of statistics is interesting as they are both quite different in approach (Gigerenzer et al., 1990; Huysamen, 2005; Kline, 2004; Loftus, 1996; Mulaik, Raju & Harshman, 1997). The resulting "mishmash" of the hybridised approach has had the unfortunate consequence of being incorrectly perceived as statistical methodology within many statistics courses (Cohen, 1994). Neither frequentist school would have looked with favour upon the "forced marriage" of the approaches (Chow, 1998a; Gigerenzer et al., 1990, p.106, 1991). Are assessment specialists even aware of the controversy surrounding the historical origins and misapplied statistical techniques which are used everyday? Do practitioners ever question the fundamental philosophies behind their

reasoning? Is this one reason why dynamic assessment fails at times to deliver the necessary goods in terms of robust science? Dynamic assessment predicates are already tuned towards qualitative methodologies which do not necessitate the use of NHST in order to ensure its progress within assessment. The above-mentioned issues which plague mainstream intelligence assessment issues need not bear upon dynamic assessment in the first place. Of note are the models assessed in chapter 5 which evidence strong testability according to Madsen's HQ ratio but do not, as a rule, necessitate NHST in advocating their science. Even though this type of statistic is used to "prove" the veracity of the method it is not employed in the original derivation of hypothesised novel constructs (meaning-making constructs which will be discussed in chapter 5).

The afore-mentioned progenitor's statistical work as utilised in psychology today are sometimes at odds with what occurs in practice. Practical instances include:

- disregard for other indicators of potential leanings towards significance such as confidence intervals utilised in classical frequentist statistics, Bayesian inference[71] (two techniques which have been in existence far longer than significance tests; Schmidt & Hunter, 1997), the plotting of data, the likelihood function
- the ensuing miscegenation of two approaches towards chance and certainty
- the simultaneous following of divergent recommendations such as the need to establish a prior level of significance before an experiment is conducted, thus following the dictates of Neyman-Pearson only to be followed by the suggestion of not commenting on non-significant results and thus following Fisher
- Neyman's behaviourist slant on data interpretation has been all but ignored
- Type I and Type II errors are regarded as philosophical issues
- which leads to statements extolling the rigour and accuracies with which null hypotheses have been rejected

Appendix 1's meta-analysis takes into account the number of studies needed to make insignificant a significant cumulated effect size. Although as Huysamen (2005) points out, a single study with a very large sample size could theoretically have the same NHST result as a meta-analysis utilising the same sample size even though it emanated from various smaller studies. However the NHST logic would be applied to the individual studies at the level of participants whereas the meta-analysis treats study as sample. All non-significant findings which have been filed away in drawers never to see the light of day are hence given weight. Will data manipulation, design and interpretation ever work back onto the reasoning behind certain psychological assumptions? Perhaps if we start to look at how we use and interpret data we will start to ask different questions about phenomena and not first work towards a research design. "Phenomenon - then - research design" and not "research design which is set in stone - then - phenomenon" could reflect a credo of sorts. It is perhaps timely to once again interject an authorial note here and state that however unequivocally the ideal of science is supported, as stated and maintained throughout chapters 2 and 3, the need to recognise the futility of some perennial efforts within a social science discipline is emphasised. It is not the question of maths, science, statistics, methodology, and overarching framework of ensconced theory growth that is being questioned but their subsequent ill-fitting use within psychology. Meehl (1978) critically appraises the resultant lack of integration of ideas behind statistical significance and a progression of science built on refutation and laments the unfortunate path that has been followed in the "soft sciences" headed by significance testing. It is a thesis that many psychologists trained today are simply not aware of original problems and are perhaps not even aware of prevalent issues surrounding nagging issues. This is evident in their current existence. NHST smacks of logic and mathematical certainty; after all the foundation of statistics is mathematics and logic and these have already been highlighted as being fraught with uncertainty and foundational cracks.

4.3.1.2 The hegemony of null-hypothesis significance testing

As with any contentious debate there are always two-sides to each story and the rendering of an objective picture surrounding this issue will result in a resolution of sorts in a shorter period of time. Many dynamic assessment research designs (specifically the South African literature; Murphy, 2002; Murphy & Maree, 2006) are specifically deployed to ensure that results point to either one of two scenarios: either an intervention is significant in terms of producing measurable change or it is not significant in bringing about change. The result cannot lurk about in no-man's land as this is simply not good science as it is not objective and does not adhere to positivist tenets of "progression". Statistical decision theory as exemplified by null hypothesis significance testing within psychology (Boniface, 1995) has served the positivist agenda well in terms of following its most basic mandates. However, as to whether it follows the development of science as exemplified by a number of theories within the philosophy of science is questionable (Brown, Hendrix & Williams, 2003). The usefulness of this type of accept-reject research design within psychology has been raised as a philosophical query (Krantz, 1999) befitting the style of this study. Krantz (1999) ties two issues pertinent to chapters 3 and 4 together in his question of whether the logic of null hypothesis significance testing matches the logic of scientific inferential practice specifically in psychology and dynamic assessment even though NHST is not all there is to scientific inference. This leans heavily on the question of how a science progresses from naïve ideas concerning human

[71] Bayesian "confidence intervals" are referred to as credible intervals (Oakes, 1986; Rindskopf, 1997).

behaviour, intelligence and learning potential towards the science of human behaviour and intelligence. NHST is often employed to bring the veneer of objectivity to psychology (Loftus, 1996). Often the decisions made on the basis of NHST findings are logical derivations but are not necessarily reflective of insight gained into any particular psychological phenomena. The "dubious epistemological claims" (Brown, Hendrix & Williams, 2003, p.78) of NHST has thus come under fire from a number of fronts as will be illustrated below. Recall that great names in psychology have founded schools based not on NHST but on sound observation, theoretical development and experimental endeavours and include Piaget, Skinner, Pavlov and Bartlett (Barrett, 2005).[72]

A criticism often levelled at psychology in general is that is fails to replicate, the "field often spends a lot of time spinning its wheels without really making much progress" (Edwards, Lindman & Savage, 1963; Kline, 2004; Loftus, 1996, p.161; Lykken, 1991; McDonald, 1997), although it is conceded that replication is more difficult within the behavioural sciences as opposed to the natural sciences (Kline, 2004). What is it about NHST that is so contentious to some and not to others? One such issue is NHST's reliance on inductive inference which can only propel a field forward if studies are replicated (Krueger, 2001) which psychology often does not do (Huysamen, 2005). As highlighted above, the link between scientific progress in keeping with the tenets of positivist leanings towards the growth and development of a subject considered scientific and the methods utilised in order to achieve and maintain such standings within the framework are not always manifest or clear. Perhaps they are simply taken for granted. In the hope of maintaining psychology's reputation (such that it is) within the broader practise of science, research designs which look objective, formal and scientific are employed within varied domains and none has been carried out with such wide-spread enthusiasm as NHST (Nester, 1996; Nickerson, 2000). Ironically, natural science disciplines such as physics and chemistry perceive reliance on significance testing as unscientific (Schmidt & Hunter, 1997)! Note the emphasis, once again, on the ideal of formalism just as was discussed in the section on mathematical formalism. Loftus (1996) highlights the logic of NHST and why it is thus a universally employed technique within the social and behavioural sciences:

i. the research question is suited to the design which is one way of trying to show that an independent variable will have an effect on a dependent variable (dynamic assessment interventions of varied sorts will effect posttest scores; especially on those whose resident potential has been severely underestimated by static or conventional assessments)
ii. two or more groups are compared for differences which may or may not have been caused by the intervening variable (control and experimental groups' scores are compared for differences after dynamic assessment interventions have taken place). The major task is to determine whether any differences are due to the intervention or if they are randomly occurring measurement errors
iii. in order to accomplish this, a probability estimate (p) is necessitated which will show the probability of observing differences as great as those in fact observed. It is usually the case that the 'no-difference between groups hypothesis' (H_0) needs to be rejected (the probability of evidencing as large a difference as was evidenced is greater than chance alone; the control and experimental groups really did differ on posttest scores).
iv. the researcher makes a binary decision based on the level chosen for acceptance of the probability. This level is known as alpha (α). If the p value is less than α then the null hypothesis is accepted; or as Loftus (1996) states, a strong decision is taken to reject H_0. If the p value is greater than α then H_0 is accepted or a weak decision is taken to reject H_0. As this is all a game of chance, it is possible to incorrectly reject or accept H_0. If a true H_0 is incorrectly rejected then the probability of making this error is equal to α, referred to as Type I error. This is known. However, if the researcher incorrectly fails to reject a false H_0 then the probability of doing so is known as β which is not usually known as there is usually no information pertaining to populations means assuming that H_0 is false. Power is usually also not known because power is equal to 1-β. Referred to as Type II error, but were it known, the power of the test then too would be conditional and not exact (Chow, 1991, 2002), an aspect highlighted by Chow as perhaps having escaped attention or being misunderstood by those touting the utilisation of power versus NHST as indicator of an experimental effect
v. based on the decisions taken at each step, the researcher then has to cope, with at times, very complex data sets

Now that the scene is set for NHST interpretation, Loftus (1996) summarises six questionable aspects regarding the use of NHST within psychological settings and dynamic assessment is one such domain:

- H_0 is rarely true to begin with (Cohen, 1994; Harris, 1997; Hunter, 1997; Killeen, 2005; Krueger, 2001; McDonald, 1997; Meehl, 1978, 1989, 1990; Rindskopf, 1997; Schmidt & Hunter, 1997). It is usually the "default hypothesis" (Kline, 2004); the "silly null hypothesis" (Nester, 1996); the "notoriously non-powerful statistic" (Rust & Golombok, 1992) or "the usual dismal prediction" (Meehl, 1990) that is rejected at a later stage and can be dismissed without any data collection at all (Rindskopf, 1997). In this vein the two hypotheses are treated asymmetrically and very rarely is a prior probability

[72] Of course it can be stated that these researchers' work was far more amenable to these strategies than perhaps it is for researchers working in fields where constructs are not so easily defined.

(Bayesian) associated with H_0 other than zero[73] (Oakes, 1986). If H_0 is usually false, sufficient sample sizes will effectively deal with their consequent rejection (Kline, 2004) and of course this is not stringent science. As a point hypothesis[74] this makes sense as no two populations means' will ever be precisely the same. In other words there will always be a difference somewhere lurking in the means, the task is to determine the feasibility of stating how different the two means in fact are. Yet, classical significance testing relies on a point null hypothesis and a distributed or diffuse alternative hypothesis (Edwards, Lindman & Savage, 1963) and in a way, stacks the odds in favour of rejecting the null. Three groups receiving qualitatively different mediatory exposures will obviously differ on mean performance as they all receive differing amounts of mediation. H_0 is employed only as a means of illustrating the implausibility of a result being obtained given the data and as such "a significance test does not permit one to assign any specific degree of probability to the hypothesis" (Gigerenzer, Swijtink, Porter, Daston, Beatty & Krüger, 1990, p.93). Nester (1996) agrees with this sentiment as he states that even if treatments yielded similar effects the chances of this happening are zero! Significance testing was never meant to be universally applicable in the first place (Hunter, 1997). Logically, one cannot defend the reasoning behind why H_0 should be a hypothesis of no difference regardless of how large the sample size is and subsequent acceptance of H_0 due to an insignificant result cannot be justified (Krueger, 2001). If H_0 is almost always false then logically the rate of Type I errors is 0% and not 5% which results in Type II errors being made (Cohen, 1994). Testing for differences between all three groups' means, results in no new information as there will be differences regardless. "Rejecting a typical null hypothesis is like rejecting the proposition that the moon is made of green cheese. The appropriate response would be 'well, yes, okay but so what?'" (Loftus, 1996, p.163). Recall that a null hypothesis is really just a hypothetical model (Gigerenzer et al, 1990). *P* is not the probability that the null hypothesis is true and rejecting the null hypothesis does not necessarily lead to the vindication of the alternative hypothesis (Cohen, 1994; Nickerson, 2000); i.e. $p \neq$ value of H_0 and 1-$p \neq$ the value of H_A

- attaining a level of statistical significance only indicates that H_0 is false and conveys no useful information as to the underlying pattern of the populations means' which is, after all, what researchers are after. Employing post-hoc tests are often riddled with their own problems of similar false rejections of H_0 when errors of the likes of number iv above, are made. Planned comparisons are able to overcome this problem but are rarely used in practice
- NHST rarely pays attention to power primarily because it cannot be accurately computed. Type II errors and hence power cannot usually be computed because there is simply no information available detailing quantitative hypotheses. With high power the researcher can conclude that there are small differences between group means and accepting H_0 may be justified. Low power implies that there may be large undetected differences between means. The use of confidence intervals (allowing a range of data including the parameter whose probability is known; Krishnamurty, Kasovia-Schmitt & Ostroff, 1995) is an alternative to the use of power, a recommendation shared by the TFSI (Nickerson, 2000; Wilkinson & TFSI, 1999). "A lack of power analyses often stems from the lack of quantifiable alternative hypotheses that characterizes the social sciences in general, and psychology in particular" (Loftus, 1996, p.163). Within some natural science disciplines, confidence intervals are not interpreted as significance tests (Schmidt & Hunter, 1997). Section 4.4.2.3 below includes an IRT model specifically developed to encompass change across testing situations which utilises confidence intervals as measure of estimated ability and modifiability estimates
- the mechanical, "formulaic mode of inquiry" (Robinson 2000 in Brown, Hendrix & Williams, 2003, p.78) and automatic need to dichotomise the two-valued decision of acceptance or rejection of H_0 is lamentable as very few social issues ever present in which decisions of the categorical yes/no variety obtain (Abelson, 1997a; Cohen, 1994; Edwards, Lindman & Savage, 1963; Gigerenzer, 1991; Huysamen, 2005; Kline, 2004; Krueger, 2001; Lehmann, 1992; Rossi, 1997; Schmidt & Hunter, 1997). To cite Cohen (1990) "there is no ontological basis for dichotomous decision making in psychological inquiry" (p.1311). Most judgements within the social sciences especially need to be mindful and this is very difficult to apply when the decision is based on a dichotomy (Harlow, 1997). Statistical inference especially within psychology should only allow for more or less support to be added to a conjecture and should not be utilised for the sole purposes decision-theoretic making based on arbitrary figures (Oakes, 1986). This is more an outgrowth of the Neyman-Pearson "decision-theoretic significance testing" approach as opposed to the Fisherian approach (Mulaik, Raju & Harshman, 1997, p.106; Oakes, 1986). Harris (1997) makes a convincing claim for the utilisation of the three-valued alternative as

---

[73] Recall Fisher's objection to this idea of subjective priors. But also bear in mind that his area of concern (agronomy) was better suited to nullification of H_0.

[74] This is not to say that NHST as is cannot be functionally utilisable as stringent procedure of acceptance or rejection within certain set-ups (Kline, 2004; Meehl, 1990; Oakes, 1986). For instance, if one is determining the point at which a metal will start to collapse structurally given sufficient weight it is clearly best to signify a definite point at which this will happen. However, to play devil's advocate, the same philosophy should undergird health issues such as medical treatments for instance. At which point is it decided that a certain medication is indeed safe? Does one disregard outcomes evidencing *p* values over 0.05? What about trials evidencing *p* values of 0.06; should the medication not be given the go ahead merely because of the arbitrary choice of a *p* value? What about individual study results touting efficacy of certain treatments only to find that once cumulated in a meta-analysis such significance is partialled out and the subsequent effect sizes decrease? A call for Bayesian inference has often been made but due to the hesitancy of researchers to subjectively assign prior probabilities this method has been viewed with suspicion as it does not adhere to NHST's "objective" veneer (http://www.cs.ucsd.edu/users/goguen/courses/275f00/stat.html) (Nester, 1996). Bayesian inference typically utilises a range of values as opposed to point estimates (Rupp et al., 2004).

opposed to this traditional two-valued approach to NHST which in essence is the "simultaneous use of two directional alternative hypotheses one for and one against the research hypothesis" (Kline, 2004, p.84). Users of such an approach will be less likely to equate the nonrejection of H_0 with the acceptance of H_0, a major point in the confusion in the interpretation of nonrejected hypotheses for the last sixty years (Howell, 1987). Conflation of the substantive theory (clinical or practical theory/hypothesis or causal theory depending on the context; Meehl, 1990; Nickerson, 2000 or practical reality; Danziger, 1990) with the statistical hypothesis (derived from the substantive; Oakes, 1986 or the constructed statistical reality; Danziger, 1990; Meehl, 1990) is often the reason that a number is the determining factor behind the acceptance or rejection of a hypothesis which does not necessarily lead to the acceptance or rejection of the theory behind the substantive hypothesis (Brown, Hendrix & Williams, 2003; Kline, 2004; Meehl, 1967; Nickerson, 2000; Wallis, 2004); i.e. rejection of the null hypothesis does not mean it has been disproved (Krueger, 2001). Meehl (1990, 1998) makes the valid point that the nullification of H_O as an import from Fisherian statistics can hardly be blamed on Fisher but rather on those misapplying the rationale behind NHST. After all, agronomy experiments rarely presented with major differences between the theoretical and statistical hypotheses; the work involved in agronomy and the technique imported into psychological research are conceptually distant from each other. One cannot blame Fisher; he worked with crop yields and his methods aided him in so doing. One can however look upon social scientists with some disdain for having imported his techniques along with a mix of Neyman-Pearson into an area almost entirely ill-equipped to utilise those very techniques. To highlight the, at times, confusing role played by substantive theory (practical importance of theory) and statistical hypothesis Table 14 is illustrated below:

Table 14 Relation between substantive and statistical hypotheses and how sample size can impact on conclusions (Johnson, 1999; Nester, 1996)

Sates of mind of a null hypothesis tester and the interpretation of sample size as related to results of a statistical significance test		
Practical importance of observed difference	*Statistical* significance of difference and sample size status	
	Not significant	Significant
Not important	Happy / *n* okay	Annoyed / *n* too big**
Important	Very sad / *n* too small	Elated / *n* okay

** Related to Meehl's (1990, 1997)[75] "crud factor" (too much noise in the data) as at some level everything is related to everything (McDonald, 1997; cf. Nunnaly; 1978, who delineates seven ways to fool oneself with factor analysis in similar vein). This is where expertise, experience and sound understanding of a field can bring into line the findings of NHST. NHST per se as statistical method is not flawed in terms of its method (Kline, 2004) but the unfortunate historical combination of hypothesis testing on the one hand and significance testing on the other within psychology is perhaps the cause of NHST's hegemony over the years. Researchers' blind reliance on statistical significance and the move away from the research underpinning it had already been critiqued by Boring in the 1920's (Danziger, 1990). This is in keeping with Brown et al's., (2003) contention that psychologists have confused numerical inference and conceptual inference. Once again the issue of a one-to-one mapping, an isomorphic relation between two conceptually different aspects is brought to the fore.

[75] Meehl (1998) in fact attributed this "crud factor" term to Lykken.

Chapter 2 discussed, among other things, the isomorphic mapping of neural-mind correlates. In this particular instance such a mapping is feasible as the one concept is reducible to the other. A similar mapping of number to reality however is not as tenable (Brown et al., 2003).

- However, in fairness to proponents of significance testing, Mulaik, Raju and Harshman (1997) highlight the fact that significance testing does not intend to disprove the substantive theory but only provide evidence for the validity of the statistical hypothesis, i.e. it is not the method which is to blame but the misinterpretation of its utility by researchers (Abelson, 1997b). Scientific significance and statistical significance are not synonymous. Recall that almost no single experiment within the behavioural sciences[76] is anywhere near conclusive even though the statistical hypothesis is quite clear on the issue. The leap from statistical and computational conclusions to inferential conclusions is an area fraught with problematic issues (Krueger, 2001). In fact Meehl (1967) points out with startling consequences that:

    - if there was increased precision of instrumentation and
    - even more sound logic flowing through experimental design and
    - increased sample sizes; one would in effect

        - increase the probability of corroborating the substantive theory via NHST "*even if the theory is totally without merit*" (original emphasis) (Meehl, 1967, p.111)
        - this manner of statistically navigating via means of significance testing might hold for statistical results but should not hold for substantive theory rejection or acceptance. This is, as already mentioned, the weak form of theory testing which allows for the further development of a theory when in fact is should not proceed in this fashion

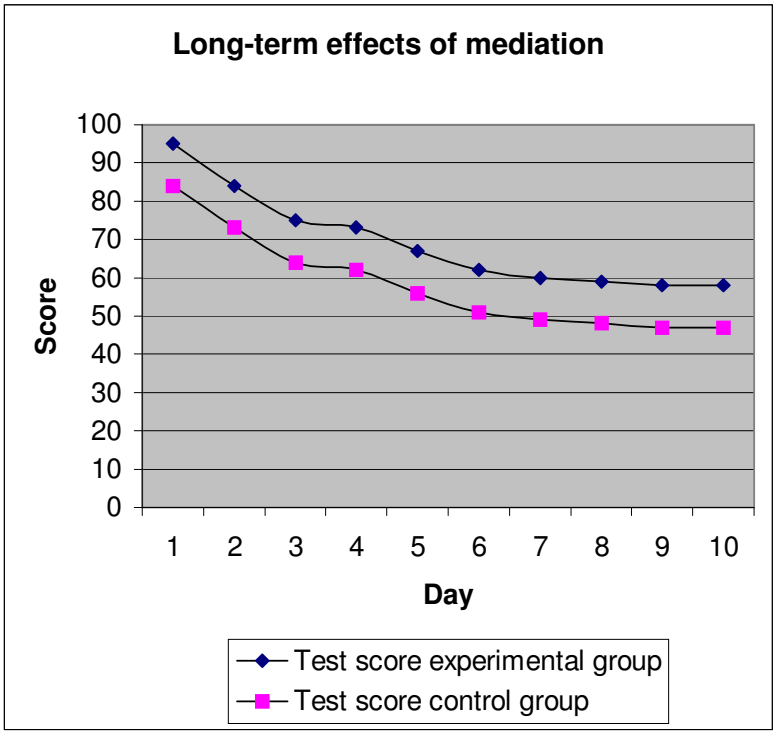
- It is little wonder that confusion abounds when one considers the immense and vast array of incredibly accurate reportings of results in many areas within psychology and dynamic assessment but when taken as a whole no single substantive law-like conclusion is obtainable. Thousands of studies grace journals yielding impressive results but very few are able to make a substantive claim. Something is clearly wrong somewhere and the thread of this story began with chapter 2 and basic philosophical core issues followed by chapter 3 and basic understandings of the scientific method culminating in the discussion on prime considerations here. Nevertheless to continue with the present concern, Huysamen (2005) makes a case for the need that does perhaps arise where decisions of this nature need to be taken. In other words researchers are not necessarily interested in the size of or the degree to which something is or is not acceptable but are really interested in just knowing ordinal values (the answer is greater than zero as opposed to knowing by how much greater than zero it is). This human obsession with dichotomy has been addressed in chapters 2 and 3 as playing a decisive role within the scientific framework and has been mentioned in conjunction with the discussion on mathematics. It rears its head once again within the realm of statistical inference. NHST circles round the critical norm of accepting or rejecting hypotheses based on the 0.05 or 0.01 α level. This is an arbitrary number and not a proven fact chosen for the acceptance or rejection of an hypothesis originally deployed by Fisher (Gigerenzer et al., 1990; Huysamen, 2005; Kline, 2004) but not supported as sole determiner of the conclusion of an experiment (Huberty, 1993). An argument in support of this number is that a line has to be drawn somewhere and this just happens to the chosen point and in addition it does provide standardisation (Chow, 1998a; Huysamen, 2005). It also adds pragmatism to the method. The same information, namely, the probability of the data given the null hypothesis, will be extracted from the same data (Krueger, 2001). Granted this may be true but how valid is the application of this rule of thumb within many social science and behaviour contexts? Can dynamic assessment efficacy be reduced to an arbitrary figure? The assessment worked or it didn't according to the α level obtained. The acceptance of an hypothesis based on its 0.05 level or its rejection based on its 0.06 level is odd to say the least (Huysamen, 2005). Surely there is more to this method than this? Appendix 1's meta-analysis results does not make use of NHST in calculating the cumulated effect size of dynamic assessment studies (Chow, 1998a)

---

[76] "*Almost* no" as the author simply does not know. It is very doubtful if any experiment within any "soft" area ever comes close to approximating law-like results evidenced in some areas of the natural sciences. All the more reason to locate for the discipline an area amenable to either "non-science" and by this is meant a referral to science in terms of its concept discussed in chapter 3; or an area amenable to scientific methodology. The discipline as is cannot continue its existence somewhere in between, which is precisely where dynamic assessment is located. The author cannot help but insert this quote from Meehl (1967) regarding the practice of psychology; "… a zealous and clever investigator can slowly wend his way through a tenuous nomological network, performing long series of related experiments which appear to the uncritical reader as a fine example of 'an integrated research programe' , *without ever once refuting or corroborating so much as a single stand of the network*. Some of the more horrible examples of this process would require the combined analytic and reconstructive efforts of Carnap, Hempel, and Popper to unscramble the logical relationships of theories and hypotheses to evidence. Meanwhile our eager-beaver researcher, undismayed by logic-of-science considerations and relying blissfully on the 'exactitude' of modern statistical hypothesis-testing, has produced a long publication list and been promoted to a full professorship. In terms of his contribution to the enduring body of psychological knowledge, he has done hardly anything" (original emphasis) (p.114). Traifmow (2003) is in agreement with this sentiment and it is also highlighted by Krueger (2001).

- Circular reasoning pervades theory and NHST. NHST posits that error is distributed in Gaussian fashion; the distributions are equally distributed over various conditions and adding up effects and error sources in a linear manner results in numerical indices. These assumptions are built into theory which biases against other types of reasoning within theories. Loftus (1996) employs a real-life example which will be attenuated to suit the following context-relevant example:
    - two groups, an experimental and control group are administered interventions. The control group is shown a video about how to study and the experimental group is given intensive mediatory assistance regarding study methods and techniques
    - over a period of time, the two groups are monitored for long-term changes in study behaviour. Unsurprisingly, the experimental group shows a higher level of acuity regarding study methods and does better in tests and exams than does the control group
    - the researcher concludes that a significant effect in test results was due to the intervention and illustrates this with the following figure

Figure 60 The theoretical significant difference between two groups either receiving or not receiving mediatory interventions couched within NHST technique (attenuated after Loftus, 1996, p.165)



    - standard data-analysis will conclude that the experimental group significantly outperformed the control group thus evidencing the superiority of the intervention. This, however, masks a true pattern which can be seen by the similar trends evidenced from both groups. They both decrease over a number of days with equal rates. The equality of rates of reduction, then, are testimony to the negligent effects of the intervention. The counter-argument would be that the experimental group nevertheless did significantly outperform the control group. The main aim though of the research was to document the long-term effects of mediation and based on the evidence did not seem to make any difference at all. NHST will miss this

- lastly but quite importantly is the logical error made when assuming that
    - H_0 should be rejected when

(i) p (observed data | null hypothesis) < 0.05

yet, rejecting H_0 implies that

(ii) p (null hypothesis | observed data) is small; which is similar to employing the logical argument that "P then Q" = "Q then P", which is obviously false (Nickerson, 2000)

(iii) *but* $p(d \mid H_0) \neq p(H_0 \mid d)$[77] which is what NHST logic in fact states (Cohen, 1994; Killeen, 2005; Mulaik, Rju & Harshman, 1997; Nickerson, 2000; Oakes, 1986; Trafimow, 2003). Probability and its inverse is not equal, to cite Carver (1978) in Suen (1990) "the probability that a person is dead given that the person was hanged is high. However, given that a person is dead, the probability that the person was hanged is quite low. In other words p (dead ¦ hanged) ≠ p(hanged ¦ dead) (p.21). Bayesian statistics, however, does make use of inverse probabilities

- the second statement does not necessarily logically follow on from the first statement due mainly to the fact that the probability of the null hypothesis obtaining given the observed data could be anything at all (unless we have access to other information) (Chow, 2002). "When H_0 is rejected, it can be because of the falsity of any of the auxiliary theories about instrumentation or the nature of the psyche and not of the substantive theory that precipitated the research" (Cohen, 1994, p.999). Prior data on H_0 being true would be most useful which is precisely what a Bayesian technique would allow - it works with and re-integrates prior probability distributions in determining posterior probability distributions (Reber & Reber, 2001). In sum, the arbitrary level of 0.05 is just as meaningful as it is arbitrary! See chapter 2 section 2.4.5.3.6 for a reference to this in the context of the progression of science and instance confirmation which is based on conditional probabilities
- Huysamen (2005) adds that rejection of H_0 does not logically imply that H_A is supported nor does it yield information about the degree to which the rejected hypothesis is false and lastly NHST is silent on its reliance on sample size

NHST yields $p(d \mid H_0)$ and does not yield $p(H_0 \mid d)$, $p(H_A \mid d)$, p(replication ¦ d) or $p(H_0 \mid$ reject $H_0)$ which some researchers think it does (Kline, 2004). To link back to the ideas discussed in chapter 3 on the logical positivists and the ensuing development of science within such a framework, Cohen (1994) details the misinterpretation surrounding H_0. If the logic of H_0 was phrased in syllogistic Aristotelian fashion the following would result:

- If H_0 is correct, then D (data) cannot occur (note the IF-THEN pattern)
- D occurs
- Therefore H_0 is incorrect

However, NHST (induction) is not couched in syllogistic reasoning (deduction) and the above *modus tollens* is not what usually results in NHST (Krueger, 2001; Nickerson, 2000) which is couched in probabilities such that you find the following:

A

- If H_0 is correct, then D is very unlikely
- D occurs
- Therefore H_0 is very unlikely

And by instituting probabilistic thinking the argument becomes invalid. In stating a false premise but making it appear palatable by dressing it in probabilistic clothing one is able to formally deduce incorrect conclusions. For instance:

B

- If the organism has legs it is a mammal
- the dolphin has no legs
- therefore it is not a mammal

The first premise is of course incorrect (lizards have legs for instance) however the syllogistic reasoning is correct leading to a correctly deduced conclusion (a Type III error, resulting from having asked the wrong question or in this instance, having posited the wrong statement; Harris, 1997; Killeen, 2005; Sehlapelo & Terre Blanche, 1996). If the statements were couched in probabilistic terms it would run as follows:

C

- If the organism has legs it is probably a mammal
- the dolphin has no legs

[77] $p(H_0 \mid d)$ is an inverse probability (Oakes, 1986). Neyman and Pearson maintained that the data might or might not support H_0 but given the data one cannot assume H_0 is upheld or not (Oakes, 1986).

- therefore it probably is not a mammal [78]

The premise is probable, not absolute, thus making the premise more acceptable which is precisely the reasoning behind NHST. Following on from **C**, NHST would conclude the following:

D

- If H_0 is false, then the result (of statistical significance) would probably occur
- The result does not occur
- Therefore H_0 is probably true and hence valid

Gigerenzer (1991) refers to the tools of the trade within a science and distinguishes tools for data processing, namely, statistics; tools for nonempirical hypothesis testing such as logical consistency and tools for the original measurement for justifying what is being researched. The logic behind mathematics (chapter 4), the reasoning behind theory growth (chapter 3) and the need for quantification (chapter 4) tie in with this discussion on the logic of hypothesis testing along with the need for experimentation within psychology. Is it possible that, through the use of such tools, new theories can emerge and attest to Gigerenzer's (1991) "tools-to-theories" heuristic? This question is answered in the affirmative as often "the apparatus of statistics may be so closely tied to a regulatory function that the quantifiers help to generate or reshape the phenomena which they set out to describe" (Porter, 1997, p.102). Perhaps we should assess chapters 2- 4 in the light of tools having created for researchers new theories hitherto unavailable? Gigerenzer (1991) does make a compelling argument, one which needs to be taken seriously considering the leanings of this study which have questioned the techniques utilised for the study of psychology and assessment.

Oakes (1986), although now quite dated, obtained results from a variety of people (scholars, researchers and lecturers) on a number of questions concerning the proper interpretation of significance testing usage. The hypothetical set-up entailed a treatment given to an experimental group after which both the control and experimental groups' means were considered (20 in each sample). An independent means *t* test is administered and the statistic yielded is 2.7, *df* 18 at $p = 0.01$. The following table 15 details the most common misconceptions relating to NHST and the answer to each is an unequivocal 'no'!

Table 15 Common misconceptions pertaining to NHST

	Usual interpretation of NHST
i.	The null hypothesis is absolutely disproved
ii.	The probability of the null hypothesis has been found
iii.	The experimental hypothesis is absolutely proved
iv.	The probability of the experimental hypothesis can be deduced
v.	The probability that the decision taken is wrong is known
vi.	A replication has a 0.99 probability of being significant
vii.	The probability of the data given the null hypothesis is known

[78] Cohen (1994) includes syllogistic arguments which are inverted to the ones employed here. He makes use of the premise "If a person is an American, then he is probably not a member of Congress; the person is a member of Congress, therefore he is probably not an American". Cohen's example is more fitting because it denies the validity of H_0 and thus illustrates the whole logic much better than does the syllogism used by the author. What holds the one way will nevertheless hold vice versa.

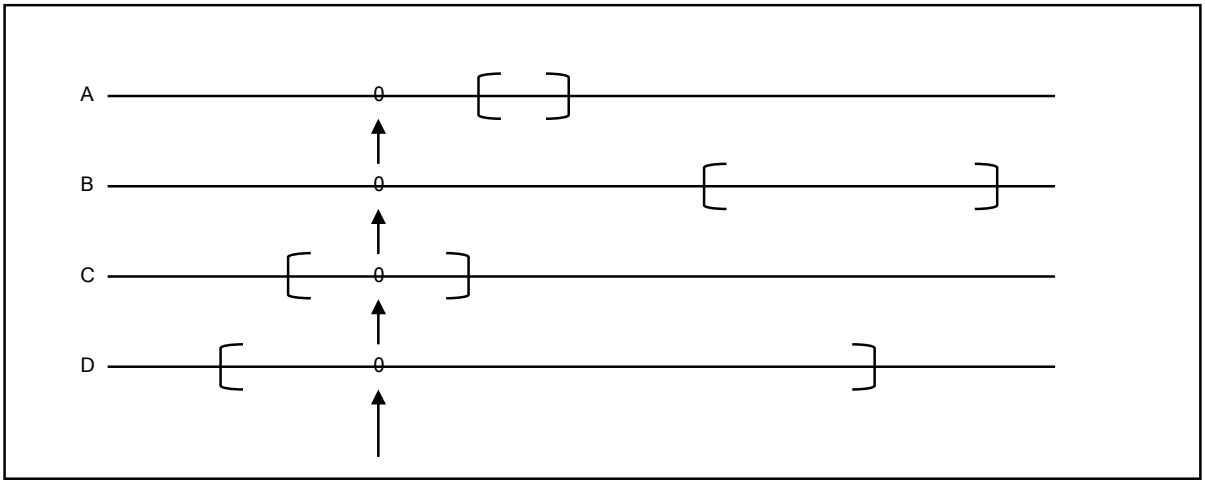
Table 16 Justification of the reasoning pertaining to NHST table 15

Reasoning
i. All that has been shown is the probability of equivalent means between the two groups. But then again, H_0 is rarely true to begin with! This is in and of itself no great feat. This same rationale underpins number iii.
ii. As with i, we only know that the probability of the H_0 is not due to chance.
iii. All that one can conclude is that there is a 99% probability that the results are *not* due to chance alone. This says absolutely nothing about the veracity of the alternative hypothesis' claim. We are no wiser about the substantive hypothesis than we were before the start of the experiment. A Neyman-Pearson confidence interval is not reflective of a parameter but an interval. This same rationale underpins number i.
iv. As with iii, we only know that the probability of the H_A is not due to chance.
v. The probability that the decision was based on chance is known and nothing more. Similar rationale underlies ii.
vi. Nothing can be said about the replicability of this experiment. The only way to further substantiate H_A is to conduct the study again. Unfortunately, replication in the social sciences is rare. The rule of thumb to follow if a claim is to be further supported would not be to increase power or sample size but to replicate studies more often. Associated probability is not power.
vii. Practically, this is the accepted interpretation of NHST; i.e. $p(d \mid H_0)$. But it has been argued that this can never be! You only have one known aspect and that is the data with which you work. You cannot know absolutely what H_0 is otherwise you wouldn't need to test it in the first place. You do not know H_0. You know the data. The data is given. Hence the probability of H_0 given the data is what you want to know. And as already pointed out, $p(d \mid H_0)$ is not equivalent to $p(H_0 \mid d)$.

Loftus (1996) suggests alternatives to NHST which, he states, are neither fancy nor esoteric. He suggests (as others have likewise suggested; Cohen, 1994; Hunter, 1997; Huysamen, 2005; Krueger, 2001; Meehl, 1997, 1998; Moore, 1997; Mulaik, Raju & Harshman, 1997; Nester, 1996; Nickerson, 2000; Steiger & Fouladi, 1997) that researchers provide confidence intervals which allow for the degree of statistical power to become evident; the range of probability estimates renders a much more plausible scenario than point hypotheses can (Reichardt & Gollob, 1997; Rindskopf, 1997). A significance test may well indicate significance or lack thereof, but yields no information as to the sizes of these differences between sample and population parameters. H_0 is rejected for instance and that is all that is known. Confidence intervals, however, are able to illustrate by how much the values were divergent from zero. Figure 61 illustrates this point. All four confidence intervals yield information beyond the confines of "H_0 is rejected".

- Line A contains an interval with a different range as well as a different length from zero to that of line B although statistically the two are equivalent in terms of H_0 being rejected. Line A's confidence interval is closer to zero than is line B which means that Line A's parameter values are closer to zero than Line B. This difference in size is an unknown in point estimate tests even though both are significant
- Lines C and D are both non significant. However, the parameter estimates are much closer to zero in Line C than they are in line D
- Statistically this makes no difference at all as NHST either accepts or rejects H_0. However, the practical significance of this extra information makes interpretation of data much more amenable to discussion especially in an area fraught with difficulties in inferential conclusion. This reiterates the difference between the statistical and substantial hypotheses

Figure 61 Information provided by confidence intervals which are not provided by point estimate statistical tests (Reichardt & Gollob, 1997, p.273)



Steiger and Fouladi (1997) state four main reasons as to the reluctance of psychologists to employ confidence intervals with which Reichardt and Gollob (1997) are in agreement. Traditionally NHST is favoured due to the pragmatic concerns which override its utility. For instance, detailing narrow confidence intervals might have the resultant effect of supporting highly statistically significant findings but which are in fact practically trivial. Larger confidence intervals have the opposite effect of seemingly less accuracy. Most psychologists are unaware of confidence interval procedures and lastly, a few such innovative procedures are not included in major statistical packages. Relying on either one of the two methods of reporting significance would be less beneficial than perhaps reporting both (Reichardt & Gollob, 1997), as it has been noted that, at times, it is necessary to report results utilising stringent point estimates.

Loftus (1996) also maintains that due to some natural science disciplines' strong form of theory testing, attempting to falsify H_A is usually beneficial and easier to do due to point hypothesis values that are assigned to the hypothesis. This is as a result of the stronger form of theory within the natural sciences (Meehl, 1998). Also, researchers should plot data as opposed to providing the information in the form of tables with *F* and *p* values which are more cumbersome to read. Huysamen (2005) does, however, present a cogent argument for the equally cautious use of confidence intervals and notes that although much is said against the designated arbitrary level of α not much is said about the equally arbitrary level of confidence intervals which are typically set at 95% and 99%. The reason for highlighting the NHST debate is not to castigate the method's utility value (especially in certain instances where its value is noted) but to highlight issues which have permeated through to areas such as dynamic assessment (among many other areas of social science research endeavour). Graphically displayed confidence intervals will better illustrate how closely the observed means of the sample data reflect the pattern of population means and hence confidence intervals serve as guides to the degree of acceptance of null hypotheses. Graphical illustrations of data in general is emphasised as an idea worth pursuing, prompted by the inadequacies of statistical decision theory (significance testing) to more accurately evidence what the data is saying (Brown, Hendrix & Williams, 2003). Confidence intervals attest to the chance of locating the true population means usually to within a 95% accuracy, so the larger the interval the less accurate the data and vice versa. NHST has a habit of imposing the "illusion of certainty on a domain that is inherently ambiguous" (Loftus, 1996, p.168) and by illustrating the data with accompanying confidence intervals one is better equipped to determine the degree of rigid acceptance or rejection of observed data. Harlow (1997), Kline (2004), Loftus (1996), Rossi (1997) as well as Schmidt and Hunter (1997) moreover advocate the use of meta-analysis as technique for eschewing traditional NHST over multiple studies (Kline, 2004), although Oakes (1986) has serious misgivings about the use of meta-analysis in this regard, maintaining that it perpetuates certain incorrectly held NHST notions.

Abelson (1997a, 1997b), Harlow (1997), Huysamen (2005), Nickerson (2000) as well as Steiger and Fouladi (1997) mention a few more options open to those wishing to peruse data from other alternatives available, such as by employing a model-fitting approach to data interpretation in which the goodness-of-fit is tested via the model of *data = model + residual (error)*. Errors are more easily picked up in this method as opposed to NHST. This mode of inference adopts both the falsificationist (strong form of theory testing; i.e. creating defeatable hypotheses; Harlow, 1997) option of NHST in addition to confidence intervals' parameter estimation, although this method is highly influenced by sample size. Bayesian statistics, as another complementary statistical technique is offered as means of data interpretation. As has been mentioned, the crux of the Bayesian method involves assigning a prior probability to the null hypothesis as being true and by adding the research data to the prior probability, the posterior probability of the null hypothesis being true is obtained. Immediately though, one can critically state that the choice of prior probability assignment is open to debate which is where the standardisation of NHST comes to the fore in its advantageous nature. Nevertheless, once the posterior probability of the null hypothesis is ascertained the result can be utilised as a prior probability in another study and over time and over a number of studies the procedure is self-correcting and is able to negate any "damage" incurred by arbitrary human choice of prior probabilities (which is the one main feature of Bayesian statistics against which criticisms have been lodged; Huber, 2002).

Huysamen (2005) not only offers alternatives to NHST but also offers recommendations as to complementing the technique as opposed to doing away with it entirely:

- Replace the nil null hypothesis with the non-nil hypothesis (a directional null hypothesis; thus including a range). In doing so, more information than just a rejection of a value of zero will be available, although if used in a sound manner, the point nil null at least brings a definitive probable point of rejection or acceptance (McDonald, 1997). The nil null hypothesis is often set up as a straw-man argument within the social sciences making the conclusion all the more scientifically useless (Kline, 2004; Oakes, 1986)
- Sample size determination via the analysis of power
    - such that effect sizes considered significant are detectable
    - yet to limit the amount of power which would detect effect sizes that are negligible
        - however, a prior decision needs to be made as to the size of α, the chosen level of power and the size of effects that will be detected by NHST. Tables stating these criteria are available

- At the very least if all else proves impossible, detail the effect sizes as well as NHST findings

Oakes (1986) summarises the fundamental characteristics of the various schools of statistical inferences and an attenuated table is reproduced in table 17 below.

Table 17 Comparison of schools of statistical inference (Oakes, 1986)

A comparison of schools of statistical inference							
School	**Interpretation of probability**	**Emphasis on decision or inference**	**Emphasis on testing or estimation**	**Use of prior information**	**Initial or final precision**	**Location and nature of inference**	**Sensitivity to sampling procedure**
Neyman-Pearson	Frequentist	Decision	Both	Informal	Initial	Probability statement on sample domain	Sensitive
Fisher	Fiducial[79]	Inference	Both	Informal	Final	Probability statement on parameter domain	Variable
Bayesian	Subjective	Inference	Estimation	Formal	Final	Probability statement on parameter domain	Insensitive

4.3.2 Summary

Psychology's over-reliance on statistical modelling and corrective techniques has long been considered by critics as testament to its perceived unsatisfactory status as science. This is erroneous thinking and has led to the, at times, unfortunate misuse of well thought out and mathematically sound statistical tools. Parallel to the developments within mathematics as formal abstract science arose statistical modelling which can be dated to early seventeenth century use. Utilisation of statistical techniques has co-evolved within psychological measurement and it is due, today, to the psychometric enterprise that many statistical models are used in other fields outside the domain of psychological testing. This is one of the very few occasions in psychology's life-span during which the discipline was able to give something to outside domains as opposed to continually taking from others. Statistical modelling has become suffused with the notions of probability and certainty, two fundamental criteria which it most certainly cannot avail of itself. This is an error and it has yet to be corrected although a substantial minority of advocates posit forth measures to correct for this error. Dynamic assessment is not blameless in its wholesale use of current statistical techniques which it uses to bolster its findings. It has been argued, however, that in order to remain relevant to the field and in order to help ensure its existence, it has had to resort to mainstream options and statistical utility is one such way. The most significant component of the discipline of statistics which grips psychological assessment is that of inference and the degree to which results are applicable to populations and how results are due to chance effects. This obsession with chance and population parameters is, interestingly enough, not a concern within the natural sciences.

To further the cause of statistical modelling and the requisite need to ensure correct levels of chance findings and its applicability to the broader population, hypothesis testing has grown and become almost the defining criteria for the existence of a statistical psychology. The discussion on statistical and mathematical underpinnings thus leveraging measurement in psychology cannot be divorced from reigning trends of science and philosophies of science. The use of statistics does not occur in a vacuum, and these impinging issues were highlighted in chapter 3. Although major bodies have looked into the question of ill-fitting statistics within the domain of psychology and other critics doing so for dynamic assessment, the full effects of these messages have yet to seep through and as with any discipline, psychological assessment is contingent upon historical contexts which in this case is peppered with a striving towards natural science models culminating in the use of null hypothesis significance testing (NHST). Originating in far flung fields such as agriculture and breweries, statistics was imported into domains and territories which, at the time, seemed to warrant their introduction. Psychology was a growing formal discipline barely twenty years old when great developments were being made within the field of statistics, so it is hardly surprising that the new techniques were set upon and factored into research designs.

[79] A measure of rational belief. Fisher does not outright reject or accept H_0 based on the data but suspends judgement.

Ronald Fisher, Egon Pearson and Jerzy Neyman are the names which punctuate the story of social science statistics particularly the concern with hypothesis testing. Due to historical events and the erroneous intertwinement of Fisherian experimentation and Neyman-Pearson hypothesis testing, much of psychology's later research and ensuing directionality was falsely skewed, mostly because of negligence of later psychologists' unwillingness or lack of knowledge in correcting the status of the discipline as it stood regarding hypothesis testing in the early 1940-1950's. Due to this negligence, a continuation of ill-fitting methods is still utilised, in areas of concern to this study, namely dynamic assessment. Alternatives to these trajectories have been explored, namely Bayesian inference, but little has come of it in mainstream assessment. The issue of probability and resultant inference is treated effectively yet differently within Bayesian inference and there is a call to utilise the method as opposed to the treatment of probability within the more classical approach as it has been argued, Bayesian inference is more in keeping with real-world psychological issues. The role of historical contingency plays out in the area of classical statistic vs. subjectivist statistics, after all Bayes wrote his works in the eighteenth century and thus could not champion his stance as could Fisher, at a time which was ripe for interference of this nature. In keeping with the tenets of natural science philosophy, both forms of NHST are assessed in terms of their weakness and strengths in determining how results are utilised within psychology, namely the weak and strong forms of NHST. Currently, the weak form is holding fast in a domain where it should relinquish its outmoded and long overdue hold on psychology, but this can only be done if researchers become aware of events. In furthering its own agenda dynamic assessment should move towards alternative measures of research design so that it no longer has to be accountable to a method which is seriously and philosophically flawed and largely misunderstood. It would in fact gain and not lose out by shifting its focus in this manner. Creating a false sense of certainty via statistical utilisation of flawed techniques is due to psychology's over-emphasis on striving towards a natural science orientation. But this can be corrected in ways which nevertheless remain robust yet different.

4.4 Measurement foundation

Prelude

Psychological measurement begins with theory extending from data and ends with inferences of stimuli and people from the data. Assessment has grown from measurement and the two are not synonymous (Meier, 1994). Ellis (1966) maintains that measurement is the link between mathematics and science, two aspects which have thus far been discussed. At least this is the formal understanding as given by Coombs (1967). If only the situation was that straightforward. Before measurement is even promulgated at the level of scaling, such activities are already a consequence of behavioural theory (Coombs, 1967). The fact that we assume constructs to be measurable is telling of two things: that behaviour is in some sense measurable (is it?) and /or the need to fulfil the accepted notion of science progress which inevitably entails measurement of some kind. Dynamic assessment's change-based philosophy of assessment strikes a disconsolate chord for those practitioners more inclined to view stable prediction as the pinnacle of test evolution (Ghiselli, Campbell, Zedeck, 1981). It is here, at the interface of educational assessment and prediction that the two aspects are conceived of as opposing rather than complementary forces (Bereiter, 1962; Biesheuvel, 1972). Changeability is the sought after behavioural aspect within dynamic assessment philosophy which, although at odds with mainstream edumetric assessment, seeks to make valid the idea of change within a robust psychometric framework. Mainstream intelligence assessment is informed from a framework encompassing a myriad of impinging variables or as Ceci, Rosenblum and Kumpf (1998, p.299) state "a galaxy of factors". Variable notions of environmental, biogenetic, ecological and behavioural factors are considered as contributory aspects constituting the holistic concept of what is popularly understood to be intelligence. One influential model of intelligence which emphasises a general underlying structure is of particular importance in this discussion on the measurement foundations upon which rest dynamic assessment encompassed within the broader arena of intelligence assessment. The notion of a general intelligence factor, *g*, underscores the stability model of unchanging scores in tests of mental ability.

Dynamic assessment's change based outlook is contrary to this very model of classical stability. To further deliberate on this issue a brief digression into the validity of practice effects is warranted. Although practice effects can hardly be equated with directed mediatory interventions of the sort espoused within dynamic assessment, the degree to which test scores can be altered due to practice effects may be of concern primarily because the underlying purported unchangeable *g* is not variable (Reeve & Lam, 2005). Practice effects seem to superficially alter scores within repeated measures designs, such as evidenced by Reeve and Lam (2005). Employing a host of statistical techniques including multi-group confirmatory factor analysis to test for measurement invariance as well as scalar invariance and item uniqueness across testing situations, the authors illustrated that *g*-based variances did not vary across tests. Reeve and Lam (2005) convincingly argued that the non-cognitive aspects which were being enhanced via practice sessions were not related to the general factor intelligence. Also no psychometric property of the instrument was being changed in any way due to these practice effects. What would be useful in this author's opinion would be to conduct this study utilising dynamic assessment as intervention variable (akin to the practice effect) and to subsequently analyse the output in the manner described by Reeve and Lam (2005). How would researchers argue lack of *g* change? Or would *g* change? If *g* changed this would entail psychometric property variance which would result in an overhaul of basic classical test theory upon which many such tests reside (Reeve & Lam, 2005). The issue of classical test theory and it's more modern counterpart, modern test theory, will be discussed below in conjunction with dynamic assessment's placement within the

two approaches. Cronbach and Furby's (1970) difference score can perhaps be equated with the Reeve and Lam's (2005) practice effect as can the former's change score be related to true growth and maturation over tests. Although this still brings into question the notion of the variability of *g* as it has been statistically shown that test-retest score changes are not *g*-loaded (Coyle, 2006) which may not be such good news for dynamic assessment initiatives. Such entanglements are indeed unresolved knotty problems which need to be taken apart and studied from as many angles as possible, hence the necessity of including chapter 2's discussion on the physiological contributions to the study of *g* and related issues.

Making manifest latent traits is the rationale underlying the need for measurement (Ghiselli et al., 1981) but the determination of what exactly is meant by latent trait very much hinges on the core philosophy attending such conceptualisation (Borsboom, Mellenbergh & Van Heerden, 2003). These authors contend that latent traits can only be considered within a realist framework as this interpretation is the only one that will suffice in terms of accounting for a causal fit between formal-theoretical and operational-empirical concepts of latent constructs. This issue is highlighted in the ACFS battery for children (Lidz, 2000b) in section 5.2.10. Moreover, Borsboom, Mellenbergh and Van Heerden (2004) are not in agreement with Cronbach and Meehl's (1955) conceptualisation of construct validity as being dependent on the nomological network in which it supposedly occurs. It either exists or it does not. Validity is more often than not assumed to refer to what the test measures but is in fact a reflection of the test scores themselves or the subsequent test score interpretations; constructs are not representative of test scores, rather the construct becomes manifest through interpretation of test scores and then perhaps only partially so (Borsboom, Mellenbergh & Van Heerden, 2004; Borsboom, Van Heerden & Mellenbergh, 2003; Suen, 1990). As there are a variety of interpretations surrounding test theory, there is also no consensus on what validity is. Semantically, validity will mean different things within different test theories. IQ results are the process of summing scores on multidimensional scales which are erroneously carried out on unidimensional scales. This results in a misunderstanding of what is meant by IQ. Moreover, whether or not IQ is related to these measures is also questioned. The tools of the trade are perhaps not to blame when assessments fail to accurately predict scores or resultant behaviour. As has been discussed above in the sections dealing with the mathematical and statistical foundations, the tools themselves cannot necessarily always be held accountable for errors that may at times be present within conclusions. Mathematical modelling of so-called lawful phenomena within psychometrics is an erroneous position from which to build more foundation. There is nothing suspect about developing mathematical models to aid in the understanding of data (Coombs, Dawes & Tversky, 1970) but the issues of lawfulness is in question. The concern running throughout chapter 4 is the original rationale for performing tests in the first place. Deriving statistical and mathematical models from first principles and providing what can at times be considered as proofs is not to be shunned or looked upon as trifling contributions to the science of psychological assessment and progress of the discipline. The question is the very need to do so in the first place. Are the very psychological latent constructs themselves amenable to quantification at all (Borsboom & Mellenbergh, 2004; Michell, 2001, 2004)?

Whether or not the construct "intelligence" exists is in any event not a matter for psychometric modelling or technique but is a question for substantive theory and will not be solved by psychometrics alone (Borsboom, Mellenbergh & Van Heerden, 2004). Theories of mental testing accommodate developments within cognitive and learning theory (Dillon, 1997) as is evidenced with newer IRT models encompassing change in their structure. Operationalising concepts is hardly a means of magically transforming latent traits into manifest quantity. The measurement foundation section will look at basic history and philosophy of measurement as it pertains to the psychosocial sphere even though it has been almost wholly informed from natural science rigour (Savage & Ehrlich, 1992). It will highlight the plight of dynamic assessment as method of assessment which, due to historical contingency, had to devise and uphold a measurement strategy in keeping with mainstream requirements in terms of reliability, validity and change score stability. Clearly proved and impeccably well thought out mathematical models of various test theories along with statistical techniques with which to manipulate, constrain and free the data can unfortunately do nothing to first principles which state that aspects are quantifiable when there is clearly no proof that this is the case (Michell, 2001).

Dynamic assessment currently dances to the tune of mainstream intelligence assessment models and current psychometric theory. Mainstream perception also dictates that the only future in which it can adequately serve a function, is one in which progress will be made along a continuum which it is currently following. Why does dynamic assessment have to fit in with mainstream psychometric theory in the first place and secondly why does it have to adhere to mainstream intelligence assessment models and theory? Why is it necessary that it should envision for itself a place within a hierarchy on intelligence measurement at all? Can it not do the following and still maintain its scientific credibility, after all, science is not measurement! Measurement is one of many facets endearing the method of science to its followers. Measurement does not make or break scientific method and hence neither should it do so for psychology, which unfortunately is precisely what is has been doing since ifs formal inception into the domain of science:

- Dynamic assessment already possesses its own unique theories and models of intelligence
- It should refrain from selling itself short as relevant model of change-assessment model
- Current intelligence research cannot even avail of its own track record a definition which is usable from one model to the next much less adapt from one testing situation to the next
- The logic is:

  - If intelligence assessment and measurement is well nigh hopeless in its current state, then
  - Why should dynamic assessment which has its own philosophy and history seek to follow in the wake of intelligence measurement and assessment as currently practiced?
- Dynamic assessment should carve out for itself a new vision and path to follow, one which encompasses what intelligence assessment has been unable (or unwilling) to do:
  - Follow on from its own repertoire of knowledge gathered since its inception
  - Predicate its trajectory on change-based assessment
  - Cease to work with the outmoded and weakly defined intelligence measurement procedures
  - Employ non-quantitative measures which nevertheless adhere to science as commonly understood by and within the community of philosophers and practitioners of science. Merely construing a construct as measurable does not necessarily mean that the construct is being measured, assuming that it exists as construed[80]
  - Employ techniques such as conjoint measurement in such a manner as to allow for the utility of intensive measures which parallel the extensive measures utilised within the natural sciences
  - Divorce itself from current hagglings in intelligence assessment in terms of construct validity, issues which, as history has so eloquently illustrated, is no nearer resolution that it was over a hundred years ago[81]
  - Predicating constructs on supposed correlated constructs is circular reasoning which dynamic assessment should not seek to replicate (added to this malaise is the very real concern of knowing very little of the original construct (intelligence) in the first place; Oberhauer, 2005)[82]
  - The time has come to forge a new path and to leave the one well-trodden in its wake

The following will substantiate and motivate the impassioned reasoning in the above bulleted concerns.

4.4.1 Elements of measurement

Figures 34 and 54 in sections 3.4.2.1 and 4.1.1 respectively are instances where rules of translation are necessitated if transformations from one realm are to be made into another associated realm. Figure 34 utilises what was referred to as correspondence rules or bridge principles which served as translation-transformation functions allowing for information at one level (a concept of learning potential for instance) to be encoded at another level (conductivity of neural speed). Likewise, figure 54 illustrated the concept of isomorphism for a two-layered system of physical attribute-to-calculus transformation via a process of coordinating rules and rules of translation. Similarly, measures of psychological attributes are designated rules of transformation via scales of sorts like those epitomised by Stevens (1946) which is essentially a representationalist theory of test measures (Borsboom, 2005). Steven's scales of measurement were not arbitrary in the sense of lack of insight into the characteristics inherent in the scales. It is the seeming lack of critical forethought into the issue of quantifiable construct which is made to map onto his scales that is arbitrary. As discussed in chapter 3, psychology's foundation as formal science has ineluctably fostered a rigorous natural science framework for its future development and has engendered a philosophy of quantification as core to this enterprise. Stevens wrote much of his work in the 1930's - 1960's and one cannot hold him responsible for being entrenched in a time and place where psychology's future as "scientific" was an almost given; a time which was, before Luce and Tukey's 1964 article, dominated by the conformity of numbers to scales which itself was representative of supposed relations between observables and numbers (Cliff, 1992). However, fifty years later, the need to re-look philosophical issues of prime consideration is necessary. It is ironic to note that measurement theories, competing with the received view, were developing in parallel and that work was published in various areas in an effort to give to psychometrics what was lacking in more traditional views of measurement.

There are a number of reasons as to why these parallel developments did not take hold as firmly as had the traditional views which by the early 1940's were rooted in mainstream discourse. Among other reasons cited are lack of computational power with which to carry out large scale and power-hungry sub-routines; the somewhat bare and abstract nature of axiomatic measurement theory; the lack of a cadre of psychologists who were mathematically and statistically able to follow the logic of what was being propounded in various avenues of measurement theory (Kline, 1998; McDonald, 1999; Schönemann, 1994)

---

[80] The more one ponders this situation the more strikingly absurd and silly it becomes. Can it be that the whole enterprise is based on such silly notions and ill-conceived logic? Upon this, we build ever more grandiose and sophisticated statistical models which we use to help ourselves out of theoretically questionable findings. Do we blame the statisticians or the psychologists who continue blindly with such methods? We use the tools "because they work"; but never seem to question their foundations.

[81] This is not to say that there is no consensus at all when it comes to defining constructs such as intelligence (Owen, 1998). But the leap from informal agreement to formally posited unanimous agreement has yet to be made.

[82] "The only way out is going through this circle again and again, each time refining the measurement instruments in light of more precise theoretical formulations and refining the theory in light of experience with current measurement instruments" (Oberhauer, 2005, p.393). Here it is patently evident that practice informs theory which informs practice, a theme resonating from chapter 3. Yet the limitations are pervasive and keep us within the system of known theory and measurement instruments. It is a necessary task to move away from these constraints into new territory but this is far easier said than done.

especially mathematically more complex item response theory (Sijtsma, 1993a) and the concomitant lack of appreciation for other means of obtaining information in a scientific manner yet remaining true to lived experiences (i.e. quantifiable yet non-numeric). This notion of quantifiable non-numeric measure can be seen in many dynamic assessment models where an attempt is often made to secure quantifiable scores for qualitative behaviour. Ultimately, such scores are numericised according to the strictures of normative testing, but there is an uneasy feeling surrounding the need to quantify for the sake of allowing the model a right to exist within mainstream assessment.

Assigning numbers or numerals or at least quantifying objects, events or notions by way of numerics presupposes that such objects, events and notions are quantifiable in the first place. Once such objects, events and notions are numerically quantifiable the deployment of various scales of measurement is, at their own rules of engagement, not an illogical path to follow. However, such numerical assignations and the operations carried out are bound by various rules expanded upon at great length in many texts dealing with measurement theory. A number of axioms are upheld when seeking to manipulate quantified objects, which, if adhered to, allow for mathematically acceptable notions of quantification and subsequent manipulation of these notions. Measurement theory as propounded and followed in the natural sciences is coherently defined and utilised, but its use within a domain which clearly does not stand in the same realm as that of the natural sciences needs to be questioned and critically assessed as its veracity as tool of correspondence may be significantly different. Rules of measurement ensure that such measures are adequately entitled to serve as measurement representations. Coombs, Dawes & Tversky (1970) and Pawlowski (1980) discuss four problems which present within measurement theory which take to task key assumptions within such a theory, namely:

- The representation problem
    - The questions
        - What can be measured? It is indeed pointed irony that intelligence is being measured without a universally defined understanding of what " it" is
        - What conditions will suffice for measurement to take place?
        - What rules are employed to ensure consistency throughout measures?
    - The discussion
        - Extensive attributes are measurable and possess additive structure. Intensive non-quantifiable structures present problems as they cannot be concatenated
        - A formal system in which correspondence rules allow for assumptions to be made regarding the assignation of numbers to events can be logically derived and deductively deduced. This set-up will follow a relational structure in which formal properties determine the type of relation occurring between the elements of a set. "The process of modeling and measurement are described as representations of empirical systems by formal ones" (Coombs, Dawes & Tversky, 1970, p.11). The authors continue to state the relationship between two systems as such:

            ⇒ A system α = <*A, R*> can be represented by another system ß = <*B, S*> if there exists a function *f* from *A* into *B* (which assigns to each *x* in *A* a unique *f*(*x*) in *B*) such that for all *x*, *y* in *A*:

            $$x\,R\,y \text{ implies } f(x)\,S\,f(y)$$

            ⇒ In essence, the relation that holds between *x* and *y* via *R* can be mapped onto the relation that holds between *f*(*x*) and *f*(y) via *S*. If the two systems are representative of each other such that α and ß map onto one another then the two systems are said to be isomorphic (this is delineated below in section 4.4.2)

            ⇒ If the model imposed is numerical then the process can be considered as measurement but numerical assignment to objects does not necessarily mean that the system is measurable. One needs to be able to prove that such representation follows rules in accordance with the above. For instance assigning numbers arbitrarily to people based on how early they arrive at a film cannot be included as measures

            ⇒ Measures have to be transitive in order to be concatenated and not all measures (such as attitude or preference) are transitive. A person prefers *x* to *y* and *y* to *z* and *z* to *x*. Clearly the transitivity of this relation has been violated in the strict sense of number-to-object relation

            ⇒ The interaction between formal and empirical analysis is the hallmark of measurement in science

            ⇒ The question is whether psychological constructs are isomorphic to systems purporting to represent them

- The uniqueness problem
    - The questions

- How much freedom is there when assigning numbers to entities (assuming of course these entities can be measured)?
- Is the choice of number assignation arbitrary or is it dictated by the measurement process itself?

- The discussion
  - Three types of uniqueness problems present: the mapping of empirical structure into a unique numerical structure; how suitable various numerical models are for such mapping of the same empirical structure and lastly how to estimate response probabilities and the concomitant issue of how the unique estimation is dependent on a finite response set (Irtel, 1994)
  - In assigning numbers to objects utilising various scales, certain aspects need to be kept in mind
  - Ordinal measures are order preserving and the assignment of numbers is arbitrary in the sense that as long as the order-preserving function is maintained the scale is usable (this is a mapping from one empirical reality to another representational reality; Maxwell & Delaney, 1985; Narens & Luce, 1986). Scales are monotonically transformed if this order is preserved and any two scales which are so preserved can be said to be related. For instance:
    - ⇒ If a person prefers x to y and y to z and x to z then any number can be assigned to the values of preferences provided they preserve the ordering (note that these preferences are transitive and are thus amenable to this type of numerical assignment). We can, to all intents and purposes, assign values of 7.5 to x; 0.0054 to y and 0.0023 to z or 1365 to x, 154 to y and 0.25 to z. Both scenarios preserve the order of $x > y > z$
  - The above example cannot hold for intervals measures however, as this scale concerns itself with the added feature of preserving the interval between successive numbers. Such scales are possible up to positively linear transformations. For instance:
    - ⇒ If preference order as well as interval magnitude is to be retained in transformations then, using the above example for $x > y > z$, it should be shown that $x - y = y - z$ assuming that the intervals are equidistant. Also, $2y = x + z$ should hold. Transforming to interval scale the notation becomes $u(x) > u(y) > u(z)$, it should be shown that $u(x) - u(y) = u(y) - u(z)$ assuming that the intervals are equidistant. Also, $2u(y) = u(x) + u(z)$ should hold. Transforming into another scale maintaining these two considerations, $u(x) = a$ and $u(y) = b$ for instance, hence $u(z) = 2b - a$
  - Absolute scales do not allow for transformations and an instance of this would be counting when viewed as measurement. For instance $3 \neq 4$ and cannot be transformed in any way without violating its inherent meaning
  - The scale type is determined by the admissible transformations. Note that latent trait theory does not avail of additive representation in the strict sense of representing response probabilities directly. Latent trait theory utilises mapping functions which restrict the range to an interval of (0,1) an example being Rasch's logistic function (Irtel, 1994)

- The meaningfulness problem
  - The questions
    - What are the different inferences that can be made when utilising different scales of measures?
    - What decisions can be made following on from the measures?
  - The discussion
    - Inferences based on scales should be invariant across admissible transformations. If worthwhile information is to be sought from scales, limits inherent in scale properties have to be taken into account. The truth or falsity embedded within measured events need to remain invariant across transformations otherwise these values change, which is pointless. Events amenable to transformations yet preservative of their truth status necessitates that the event and scale are congruent. For instance, ordering fruit according to the time at which they are placed on the table is an arbitrary number-assigning process. Each piece of fruit is assigned a number based on its order of placement. It is nonsensical to state that the "pear is twice the apple" because the pear happened to be assigned the number 4 while the apple was assigned the number 2. The interval between them is not transitive
    - As Coombs, Dawes and Tversky (1970, p.17) state "a more difficult problem arises with respect to a statement involving numerical values for which no explicit measurement model exists. The measurement of intelligence is a case in point". The question then posed is: how can measurement be possible if number assignment does not follow any of the rules described above? The answer to this is three-fold and entails the following:

⇒ Prediction - in which the elusive dependent variable cannot be identified unless it is predicted for by concurrent measures yielding statistical (and hence substantive[83]) correlations. It is partialled or factored out as existing

⇒ Description - such statistics are only informative so far as they elucidate or infer the underlying construct which for obvious reasons is very limited in descriptive statistics

⇒ Direct assignment - most often utilised in psychological research are scales where direct assignment of numbers occurs. Rating, category and magnitude discrimination are such scales which make no use of any rules as described above as there are no measurement models as such which have been developed (however see section 4.4.2.3 on IRT test theory). They are usually treated as nominal and ordinal scales

- The scaling problem
  - The questions
    - What is involved in developing a scale?
    - How many types of scales are there?
    - How is error dealt within in these various scales?
  - The discussion
    - Campbell's (1920) treatise on measurement stipulated the lack of progress of psychological measurement due to the fact that only extensive measures are amenable to mapping onto interval scales and since only intensive measures exist for psychological constructs, no psychological constructs are interval scalable. However measurement in the form of derived additive conjoint measurement makes it possible (Brogeden, 1977; Luce & Tukey, 1964; Michell, 2003a; Perline, Wright & Wainer, 1979) to measure psychological constructs derived from axiomatic principles (Velleman & Wilkinson, 1994) (see section on conjoint measurement below). Scales, other than those advocated and discussed by Stevens (1946) exist (Velleman & Wilkinson, 1994). Scaling techniques and the manner in which error is treated will be discussed separately in the measurement models below

4.4.1.1 Extensive and intensive measures

The natural sciences pride their disciplinary success on primitive or extensive measures which are amenable to concatenation (they posses additive structure); hence quantification is paramount (Kline, 1998). Michell (2001) and Tyron (1996) echo Lazersfeld (1977) in their opinions regarding the necessity and utility value of quantifying units of "things" or "stuff". The usual idea of measurement progression stems from the seemingly intuitive idea of the flow of number to quantity to measurement (Michell, 2001). Basic philosophy underpinning the growth of a psychological science meant that in order for a secure place to be found for psychology as robust discipline nothing short of physical attainment of concepts would be accepted. Lykkens' (1991) citing of Richard Feynman's "cargo cult"[84] syndrome comes to the fore in this particular instance. Psychology, as formal discipline, might look right, might have everything in place and be set to go. Alas, the various endeavours never seem to get off the ground at the conceptualisation stage. Attempts to salvage for the discipline some respectability and thereby engaging in some serious damage control in terms of its image results in natural science methodology being employed (Berka, 1992). This turn of events has subsequently only served to further entrench the discipline into a quagmire of unease. However unpalatable an equating of natural science concepts to social concepts is, a concerted effort in this regard is proffered by Tyron (1996) for instance who envisions a future of numerically assigned units of measures worthy of natural science respect.

Let us, for the moment regard just such a future in which concepts, as borrowed from the natural sciences are indeed implemented in the social science sphere. Taking the lead from primitive extensives (fundamental) from which all other ratios are derived, Tyron (1996) seeks to build for psychology a knowledge hierarchy similar to that which is in place for the natural sciences. Two fundamental extensive theoretical quantities in the natural sciences upon which most other quantities are derived are length and time (Barrett, 1998; Ellis, 1992) defined exclusively in terms of the number of wavelengths emitted from krypton-86 (defining the metre) and the number of transitions between two energy levels of cesium-133 atom (time). Units are assigned to these measures and find unanimous agreement throughout the physics world. This is not to say that length and time could not be defined in other terms and had physics developed in some other manner this might well have been the case (Michell, 2001). Derived from these two extensive measures are intensive measures or ratios expressing these primitives in terms of other measures such as mass, density, volume and area which are expressed as the products of two or more extensives (Domoter,

[83] Recall that the statistical and substantive hypothesis or construct is not necessarily one and the same thing!
[84] Feynman's "cargo cult" syndrome is an apt description of the mammoth task psychology faces in progressing beyond its current situation. Briefly, the South Seas cargo cult peoples during the second world war witnessed an unprecedented number of airplane landings which delivered favoured goods. In an attempt to bring the scenario to life once the war had ended they reconstructed the exact series of events such as would allow for the planes to land again. But to no avail, no matter how perfectly the scene had been laid, no planes ever landed. Psychology, it seems, is suffering the very same delusions. Everything is in place. So why is nothing happening?

1992). Derived concepts are predicated on primitives which have been clearly defined and as psychology has no purported extensive measures, perhaps as a formal discipline it can predicate itself on derived measures (or so the logic proceeds) (Schönemann, 1994).

Length and time are definable due to explicit agreement on objective criteria of measurement. The fact that these concepts can be measured is indeed perhaps the most noticeable aspect of the measurement enterprise, at least within the world of physics but the same cannot necessarily be said of psychology for which measured constructs do not always exist (Maraun, 1998). Numerical assignation does not entail that measurement cannot take place however and it also does not preclude quantitative measurement (nonnumerical quantitative objects or events), as Euclid's geometric axioms clearly show; these are utilised in nonumerical ratios (Burgess, 1992; Koslow, 1992; Savage & Ehrlich, 1992). The assignment of numbers to events is only one such manner of measurement and is unfortunately the one dominant trend within measurement due almost entirely to the persuasive arguments detailed by Stevens (see below) (Cliff, 1992). Originating with physicist Fechner's (1860) publication on the elements of psychophysics and the research into reaction times, ca. 1862; Ebbinghaus' (1885) work on learning as well as Galton's work in individual differences; Helmholtz's (1887) treatise on counting and measuring followed by Hölder's (1901) treatise on axiomatic quantity and mass and Campbell's (1920) work on fundamental extensive measurement; Bridgman's 1931 discourse on dimensional analysis as well as Steven's (1946) treatment of the scales of measurement it was only in the 1960's that other views espousing measurement made their presence felt. For instance Luce and Tukey (1964) (Boring, 1961; Ellis, 1966; Krantz, 1964; Luce & Tukey, 1964; McGrath, 2005; Michell, 1999; Narens & Luce, 1986). The latter's approach towards measurement as well as the under appreciated works of Luce, Krantz, Suppes and Tversky (1971, 1990) has not taken as firm a hold on the measurement community as has the formers' representative works (Balzer, 1992; Cliff, 1992; Kyburg, 1992; Savage & Ehrlich, 1992; Schönemann, 1994).

Tyron's (1996) clearly physically inclined psychological knowledge hierarchy (which he refers to as behavioural physics) encompasses units of measures such as latency measured in time; duration measured in time units; countability measured in cycles; frequency measured in cycles per unit time; celeration[85] measured in cycles per unit time per unit time (an interesting squared notion) and inter-response times measured in time per cycle. Such a knowledge hierarchy could perhaps be instituted for a quantified realm such as that offered in figure 51 below. Unfortunately there is as yet no agreed upon unit of measurement resulting in numeric ratios of magnitudes on various scales of measurement being in error to an unknown degree (Barrett, 2000). The fundamentals discussed in chapter 3 are pivotal to the discussions taking place in this chapter. Psychology as formal discipline should engage in either a natural science flavoured approach to the study of aspects which are amenable to such renderings or it should not. Construing a nonquantifiable enterprise as pre-scientific (Michell, 2001) is naïve but in order for this view to be bolstered it will have to define smartly what it means by various nonquantifiable aspects. This can be done and moreover can be conducted in an enlightened scientific manner (Michell, 1999). To return to the notion of predicating psychometrics on derived measures (which strictly speaking is not the natural science model at all), the advent of conjoint measurement was brought in to "plug the hole" (Schönemann, 1994). Conjoint measurement would replace the need to derive measures, as it would itself become a fundamental measure. Before turning attention towards conjoint measurement, a brief tour of concatenation is necessitated.

4.4.1.2 Rules of concatenation – the necessary additive structure

Proposing to measure an attribute assumes that the attribute can be assigned a quantity, usually via a number. Defining quantity would then be a logical step to take in allowing the magnitude of quantities to be represented numerically. This allows for both a definition of the kinds of attributes which are quantifiable as well as what measurement is (Michell, 1999). Thirdly, hypothesising an attribute's quantitative nature via observational means enlightens the process of how to quantify. The proven track record of quantification within the natural sciences attests to the necessary concatenation structures inherent within their quantitative constructs but whether this feature of additive constructs can be said to include constructs such as intelligence is questionable (Schönemann, 1994). This discussion on quantification, as explicated by Michell (1999) is indebted to the writings of the German Otto Hölder who set forth his axioms of quantity and theory of mass (Coombs, Dawes & Tversky, 1970). Michell (1999) firmly states that regardless of the mathematics involved in the axioms, the logic of quantification is very much a branch of philosophy and not mathematics and must be treated accordingly. Quantifiable attributes, for instance, may posses length. The fact that an object can be said to be x metres long assumes a numerical relation; one that can stand in comparison to another object of y metres long. In order for this statement to ring true, numerical relations must exist. They do not exist for nonquantifiable objects (and thus, according to Michell (1999) a quantifiable psychology in terms of numerical relations does not exist. He does not state that the science of psychology cannot exist, only that as understood by measurement theory, it is nonquantifiable in most instances). The following exposition is taken from Michell (1999) who has relied on Hölder's axioms for measurement. Hölder's (1901) work was, apart form mathematicians' interest, effectively ignored for over fifty years and was resuscitated by the

[85] The word used in Tyron's (1996) article is "celeration". The author assumes that he meant something akin to "acceleration" but is not entirely sure of this.

subsequent works of Suppes (1951) and Nagel (1931) (Michell, 1999). Stevens's work, however was not ignored even though he had not explicitly set forth any guidance as to what to look for in measurement (Michell, 2002). Michell (1999, 2002) employs length as an example but any attribute *Q* can be substituted (Ross, 1964 uses a balance pan to illustrate his examples pertaining to the axioms of additivity). Numerical relations require additivity or an additive structure. Two objects of *x* and *y* metres stand in an additive relation to each other provided that:

1. for any lengths, *x* and *y*, one and only one of the following is true;
   a. $x = y$
   b. there exists *z* such that $x = y + z$
   c. there exists *z* such that $y = x + z$
2. for any lengths *z* and y , z + y > z
3. for any lengths *z* and *y*, $z + y = y + z$
4. for any lengths *x*, *y*, and *z*, $x + (y + z) = (x + y) + z$

---

5. for any length *x*, there is another *y*, such that $y < x$
6. for any pair of lengths, *x* and *y*, there is another *z*, such that $z = x + y$
7. for every non-empty class of lengths having an upper bound, there is a least upper bound

The additive relation is a permanent property of these lengths and is independent of what is done to any quantified object. The first condition stipulates that both lengths are identical and, if not, that the difference between them is made up of another length. The second condition stipulates that for two lengths when added together will always result in a summation which is larger than either of the two separate lengths. The third condition stipulates the irrelevance of order of additivity. Likewise, the fourth condition stipulates the irrelevance of the order of the compound additivity. If all lengths have the structure imposed by conditions 1 – 4, then the lengths are additive. However, can all possible measurable lengths be considered? In other words, can other lengths that are not necessarily a part of those admitting to the first four conditions be included? If conditions 5 –7 are upheld then the answer is an affirmative one. The fifth condition stipulates that there is no smallest length as smaller lengths merely keep getting smaller. However condition two stipulates a lower bounded level which is not zero or smaller. Likewise, condition six stipulates that there is no upper bound, as lengths merely increase. Unlike condition five though, where there is a lower boundary, there is no upper boundary for lengths. This is tempered by condition seven which stipulates that all possible lengths are continuous.[86]This concludes the necessary structure if lengths are to be considered measurable; but Borsboom (2005) gets to the heart of the matter and states: "additivity is often desirable because it is simple, but it is only desirable if substantive theory suggests that additivity should hold. Substantive theory *may* prescribe an additive model, but I do not see why it *must* be so" (original emphasis p.116).

Wholesale import of physical notions of measurement might not hold in the psychological realm. The need for additivity in psychometric construct delineation is expressed most cogently in latent trait models of change across measures and the criterion of necessary additivity is illustrated best in the discussion on multidimensional Rasch models for learning and change in section 4.4.2.3 below. Comparing relations between magnitudes, such that $x > y$ for instance, does not yield information directly bearing on the objects per se, but is an arrangement bearing on the relationship of the magnitudes of the two objects. This is perhaps the key point when detailing what can and cannot be identified as quantifiable measure. If a psychological construct such as intelligence is assigned magnitude (and this is not an agreed upon scenario) then a score $x >$ score *y* only holds in so far as the relation between the magnitudes holds true. This has to be translated back into the empirical realm where $x' > y'$ may not hold as an isomorphic relation. It has commonly and lamentably been accepted that the transformation from empirical notion (x' or y') is isomorphic to measured *x* and *y*. This is a gross and unproven assumption. Moreover, these conditions are upheld as magnitudes which need not necessarily lead onto quantification. Frege's conceptualisation of natural numbers as classes of similar classes and of real numbers as ratios of magnitudes is telling of the early understanding of considering ratios of magnitude independent of the quantity or magnitude assigned to individual objects. Whitehead and Russell also followed Frege in this regard. Recall Frege's logicist foundations and his use of letters of the alphabet to define number (see section 4.2.1 above). Narens and Luce (1986) provide three very short yet comprehensive expositions on structure preserving concepts, conjoint structures and concatenation structure (pp.179-180).

4.4.1.3 Conjoint measurement as fundamental extensive

Psychological traits were originally considered immeasurable (cf. Campbell, 1920) and in answer to this criticism conjoint measurement, as nonextensive structure interval-scalable method was offered as partial solution (Narens & Luce, 1986; Perline, Wright & Wainer, 1979) and has been referred to as deep measurement (Narens & Luce, 1986). In order for psychological traits

---

[86] The elaboration on the necessity of condition 7 is given in Michell (1999, p.51). See also Ross (1964, chapter 2) for more detailed descriptions of the elements of a philosophy of physical measurement and specifically pp.52-62 for elaborations on the theories of the additive type.

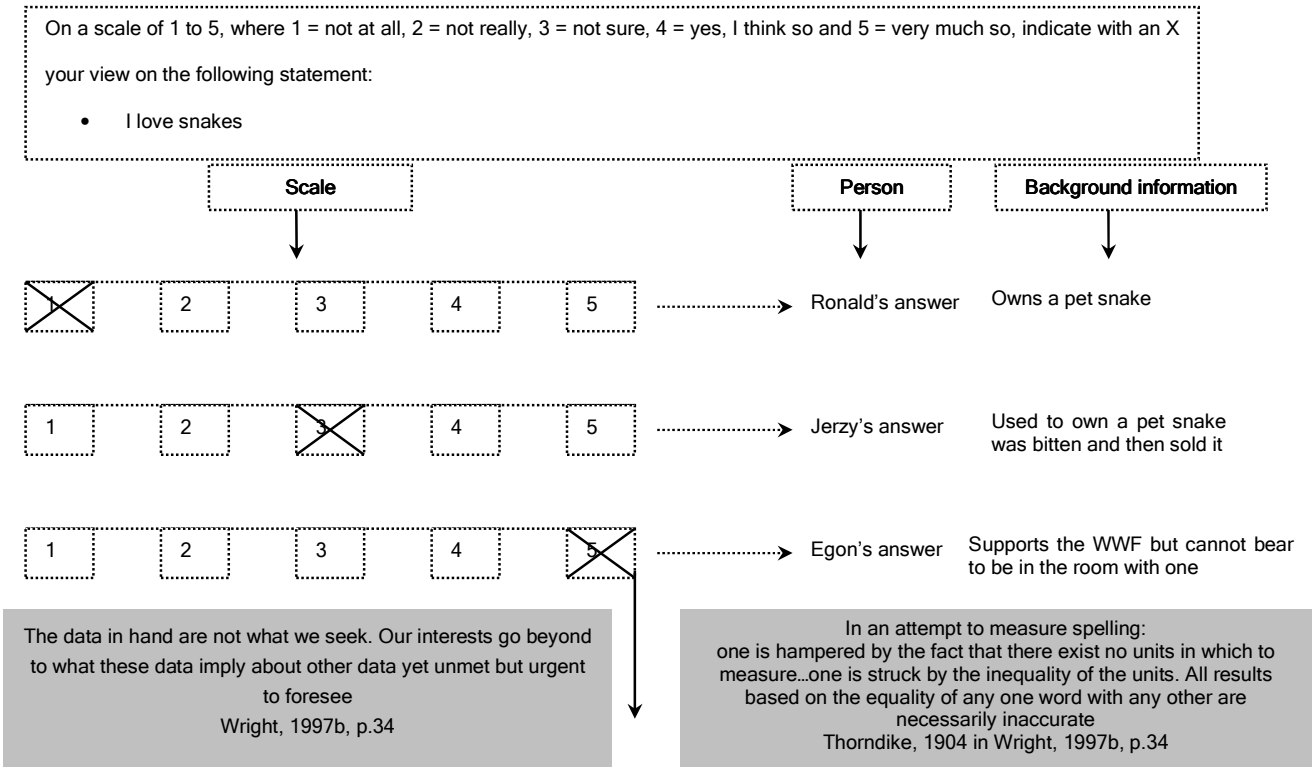
to be measures they would need to avail themselves of more than just ordinal measurement (Michell, 2003a) a scale on which quantitative measurement is suspect. Ordinance denotes order not magnitude and as such the two are hardly commensurate with each other. Recall the above discussion on numerical relations which require additivity or an additive structure. For instance, Thurstone's "law of comparative judgement" in which it is stipulated that an attitude being endorsed more so in one person than in another is evidence of a quantifiable measure is incorrect (Michell, 2003). There is simply no self-evident proof here evidencing that this is the case. That there may be a geometric distance between two points on a line is not contested, but making the leap between different emphasis of endorsing statements and equating those to points on a line is not only unscientific, but simply not thought through.[87]Paul Levy's (1937) proof of indivisibility, which is logarithmically equivalent to conjoint additivity illustrated that stable laws could be constructed even when the decision as to what to count is an arbitrary one, but this requires infinitely divisible parameters (in Wright, 1997b). Rasch (1980) was later to apply the divisibility requirement for stability. An attribute is a simple order only if the following three manifest in terms of the levels' ordering (assuming a continuous variable) (Michell, 2003a):

8. transitive IFF for any a, b, and c, if $a \geq b$ and $b \geq c$, then $a \geq c$
9. antisymmetric IFF for any levels, a and b, if $a \geq b$ and $b \geq a$, then $a = b$ and
10. strongly connected IFF for any levels, a and b, either $a \geq b$ or $b \geq a$

Any continuous quantity is a simple order but a simply ordered attribute is only a quantity if the relation $a+b=c$ exists and satisfies the above-stipulated conditions 1-5. An attribute can fulfil conditions 8-10 but this does not suffice as additive, hence, an ordered attribute is not necessarily an additive one. Psychological measures (ordinal measures) are therefore not quantitative (Michell, 2003a) and without interval scales meaningful statements about differential and developmental psychology will be difficult to make (Jensen, 2005). Thorndike, as far back as 1904, had already recognised this fact in addition to the non-linear nature of raw scores (Wright, 1997b). Figure 62 attempts to illustrate this.

[87] Making available a line segment with markings of equal intervals in a questionnaire does not solve the problem either! Who is to say that the arbitrarily determined intervals should be so designated?

Figure 62 Attitude measures are not geometrical measures and even if they were they are not quantitative by default



- Clearly, it can be seen that Egon favours snakes to a much greater extent than Ronald. But Ronald owns a pet snake. This already indicates that Ronald's idea of "not at all" is not quite the same as Egon's. That's the first problem: raw scores are not measures (Wright, 1997b). Ordinal measures cannot assume nominally equal intervals on the scale (Cliff, 1991b)
- Egon cannot therefore be said to like snakes five times as much as Ronald purely by looking at the geometric display or units on a line. That's the second problem: additivity as understood from raw scores cannot be
- Jerzy seems to be pretty neutral but he in fact once owned a snake which would make one think that he actually likes them a lot more than he is showing
- Egon in fact is a nature supporter and thus endorses the "highest" level but as for snakes in particular he cannot stand them. He should really have endorsed number 1 or 2 had there been a common metric or scale and that's the third problem
- In essence: comparisons cannot be made because no common metric has been defined and assuming we relax this assumption for the time being what can be said of the additivity of the endorsements? They might be ordinal (but here they are not even that!) but there is nothing to suggest that the results are additive

A related aside - worthy of some consideration……

Michell's (2003) insights are particularly illuminating. Paraphrasing him, but taking almost directly from his message the following is asserted:

- The distance bewteen two points in multidimensional space is a function of the differences between the points on each dimension
- The simplest scenario would involve two dimensions. In order to ascertain the difference between the two measures, one needs to calculate the difference between the two points on each dimension, square it, add the two together and utilise the positive square root of the sum
- As the number of dimensions increase so too do the number of squared components
- This process is a Euclidian one (see the discussion above on the importance of non-Euclidian space and what it meant to the philosophers of mathematics in terms of the importance and solidity of presumed timeless axioms)
- Here is the crux: Euclidian dimensions work very well indeed for physical objects in such space but there is nothing to substantiate a wholesale import of this methodology into the psychological realm. Who is to say that psychological space is equivalent to Euclidian space? Scientists (mostly mathematical psychologists) seemed to have jumped a chasm without building a bridge to serve as a foundation for allowing them to do so in the first place ……→

- Euclidian distance belongs to a family of distance functions known as Minkowski metrics (1864-1909 and who was highly esteemed by Hilbert; Hey & Walters, 1997)
- Any distance function which takes all the distance components and raises then to the same power, say *r*, ($r \geq 1$ and is a real number) and takes the r^{th} root of the sum is known as a Minkowski metric
- Psychologists who study attitudes have available two parameters that can be varied; the number of dimensions they think underlie the variable (note that they do not know, they are guessing, albeit an educated guess) and the value of the Minkowski constant, *r*
- These two values can be adjusted to fit the data at hand
- Due to the 'ease' with which data can be utilised within such a framework, psychologists do not hesitate in doing so and hence feel that abandoning this quantitative imperative would be a step backwards perhaps (and hence not in keeping with a more progressive stance; at least this may be the misperception)
- Psychologists are not, it seems, bothered by the number of questions that can be raised against the use of such multidimensional quantitative hypotheses and as such
- "will remain locked inside their own closed system of quantitative thought" (Michell, 2003, p.24)
- The place for a qualitative psychology is maintained and the overpowering need to remain transfixed within a quantitative system (which is clearly at odds with a number of research directions in psychology, among them, dynamic assessment) is bordering on delusion. Dynamic assessment, as posed throughout this study, should align with a more qualitative approach or if it is to continue on a path of quantification, should pursue measurement tools utilising conjoint measures. To employ "quantitative techniques" not befitting the research questions asked is counter-intuitive

However, the place for conjoint measurement is found in this predicament and interestingly enough, was already in existence as early as 1901 with Hölder's work. This predicament was overcome by the simultaneous derivation of a measure without the need for its intensive measure to be concatenated, a necessary condition for additivity within the physical sciences. Adams and Fagot (1959) (in Schönemann, 1994) first proposed a scaling method of measurement by the simultaneous scaling of two measures; hence co-joint or conjoint measurement. Additive conjoint measurement is the most popular of this type of scaling technique (Schönemann, 1994). This was deemed a suitable technique (albeit an indirect route for identifying additive structure; Michell, 2002) to replace the void left by the lack of concatenation constructs in psychology. What makes this a fundamental measure though is questionable as the logic behind this proceeds on the basis of two measures which are not themselves fundamental and with the joining of the two a fundamental construct arises. However, it is due to the joint distribution of item responses that latent trait models are testable in the first place (Borsboom, 2005). Latent trait modelling cannot be tested directly for any particular item because the underlying independent trait is latent and therefore not known (whereas if CTT posits known *a priori* ability distributions it is unable to tell whether endorsements are due to ability or item functioning), but it can be tested indirectly through the joint probability distribution for the items responses but the model is refuted by one single instance of axiom violation (if double cancellation does not hold for instance; see below) (Borsboom, 2005; Ellis, 1990). In other words, it specifies the conditions under which the structure of the correlation between factors provides information about the underlying attributes (Michell, 2002).

In order to obtain a measurable variable which can be concatenated, two or more joint factors apply to some event which is made manifest via these joint independent variables. For instance, the amount of progress made between pre and posttests co-varies with concomitant drops in the rate of questions asked during mediation. Potential, as dependent variable is evident via (i) the amount of progress (answers successfully completed after training) and (ii) the drop in the rate of questions asked or aid sought. Two independent conjoint factors which make manifest this potential might include, as an example, the type of mediation utilised (i) long-term or short-term and (ii) intensive small-scale qualitative or cursory large-scale interventions. "Potential", like its "intelligence" counterpart is a construct which is not isomorphic to quantitative measurement which need not mean that they are not amenable to non-quantifiable measurement (recall Michell, 1999). Coombs, Dawes and Tversky (1970) state that

- by simultaneously measuring both the dependent and independent variables (Anastasi, 1988) and
- assuming that the empirical system is sufficiently rich in these types of measures (which dynamic assessment has yet to make obvious) and
- by axiomatising the conjoined ordinal measures into interval scale measures
- an additive construct is obtained for psychological measures (potential / intelligence) where there was none before. The notion of additive representation takes the place of interactions within analysis of variance
    - The difference between analysis of variance and the additive model is that the former seeks to determine whether the cell means are descriptive of additive combination of their column and row components whereas the latter model seeks to monotonically transform scale values in such in a way that the requirement of additivity is adhered to via the transformed cell values. Such transformation will exist dependent on three axioms being satisfied, namely the double cancellation and solvability as well as the

Archimedean conditions[88] (Kline, 1998), of which only the first will be briefly sketched (note that conditions 1-10 above in 4.4.1.2 and 4.4.1.3 will need to be taken into consideration when interpreting the following):

Double Cancellation

Double cancellation is a consequence of additivity (Borsboom, 2005). As is now known, additivity necessitates that objects (*a, b*) can be represented as $f(a) + g(b)$. Assume the following matrix consisting of Factors A and B.

		Factor A		
		1	2	3
Factor B	1	(a_1, b_1)	(a_2, b_1)	(a_3, b_1)
	2	(a_1, b_2)	(a_2, b_2)	(a_3, b_2)
	3	(a_1, b_3)	(a_2, b_3)	(a_3, b_3)

If certain pairs of values of A are ordered by ≥ then other particular pairs of values will also be ordered. As with transitivity, ≥ is a simple order (Michell, 1990). Within conjoint measurement, the transitivity of ≥ on A is a special case of cancellation (Borsboom, 2005):

- (a_2, b_1) has to be ≥ for instance (a_1, b_2) IFF
- $f(a_2) + g(b_1) \geq f(a_1) + g(b_2)$
- If this is so, then $(a_3, b_2) \geq (a_2, b_3)$ and $f(a_3) + g(b_2) \geq f(a_2) + g(b_3)$
- Thus, $f(a_2) + g(b_1) + f(a_3) + g(b_2) \geq f(a_1) + g(b_2) + f(a_2) + g(b_3)$
- which is simplified to $f(a_3) + g(b_1) \geq f(a_1) + g(b_3)$.

Additivity implies that $(a_2, b_1) \geq (a_1, b_2)$ and $(a_3, b_2) \geq (a_2, b_3)$ then $(a_3, b_1) \geq (a_1, b_3)$ which can be proven and is known as the condition of double cancellation. This is represented in the following two matrices. Double cancellation (the second conjoint matrix) is a consequence of the first conjoint matrix. Double cancellation thus provides indirect evidence for quantitative structure and conjoint measurement allows for the determination of attributes' additive structure as opposed to their merely being ordinal (Michell, 2003).

- The Rasch model is a particular example of such conjoint measurement with an underlying stochastic structure (Brogden, 1977; Embretson, 1996; Embretson & McCollam, 2004; Michell, 2002, 2003; Perline, Wright, 1999; Wright, & Wainer, 1979; Rasch, 1980). In other words it is a probabilistic model which models the probability of a response and does not model the actual response (Brogden, 1977) (see section 4.4.1.3 below for IRT model information) or in Rasch's (1980) words "a means of describing a series of events which cannot be predicted to occur as definite points of time, but to which probabilities of occurrence may be ascribed" (p.36). Part of converting qualitative ordinal level data into interval data would be to apply such a stochastic measurement model (Wright, 1997a). Rasch utilised Poisson's distribution of exponential additivity as it enabled the equation of two tests (rather than items) to be independent of a distribution (Jansen, 1994; Rasch, 1980; Wright, 1999). The importance of probability as sub-discipline within mathematics, statistics and measurement can be clearly traced from the early writings of Bernoulli, Poisson and Bayes.[89] They have played forth in behavioural statistics as well as measurement and IRT as can be seen in the Rasch model and even though Rasch started from a probability angle, the resultant IRT curve was a logistic model (Baker, 2001. "Rasch models construct conjoint additivity by applying inverse probability to empirical data and then testing these data for their goodness of fit to this construction" (Wright, 1999, p.80).
- This discussion on conjoint measurement as illustrated by the Rasch model is pre-empting the brief introduction to IRT below but it is necessary at this juncture to qualify why the Rasch model is one of conjoint measurement. The following is taken exclusively from Borsboom (2005). Additive versions of latent

[88] See Borsboom (2005, p.98); Coombs, Dawes and Tversky (1970, pp.26-29); Michell (1990, p.72) who present 36 substitution instances of double cancellations; Luce and Tukey (1964, p.3) and Perline, Wright and Wainer (1979) for more extensive explanations.
[89] It is perhaps necessary at this juncture to point out some of the advantages as well as disadvantages of utilising Bayesian statistics within a psychological assessment situation. For instance (Kingsbury & Houser, 1999) utilised Bayesian priors for a student scoring procedure which proved useful in an IRT adaptive test set-up.Two students evidencing the same true score and same test performance will receive different Bayesian level estimates if their priors are similarly different. Although true scores are unobservable, test scores are not and if they evidence equal results it will be hard to understand why different scores are eventually allocated to these two students. Cognisance must be taken of such situations.

trait models, such as the Rasch model hypothesises expected item responses to be logistic functions of the latent variables; this function, which encapsulates the subject *i*'s response to an item, *j*, is as follows:

$$P(Uij) = \frac{e^{\theta i + \beta j}}{1 + e^{\theta i + \beta j}}$$

- $P(U_{ij})$ is the probability of a correct response; β_j is the location of item *j* on the θ scale where the probability of an endorsement would be 0.5
- Item response probabilities are then monotonically transformed and evidence simple additive representation and the above model is then rewritten as follows:

$$\ln \left[\frac{P(U_{ii})}{1 - P(U_{ij})} \right] = \theta_i + \beta_i$$

- ln is the natural logarithm.
- The axioms of conjoint measurement are applicable to the model in its stochastic form if both the following hold
  - the probability, $P(U_{ij})$ is transitive;
    - i.e. if $P(U_{ij}) \geq P(U_{kl})$, and
    - $P(U_{kl}) \geq P(U_{mn})$, then
    - $P(U_{ij}) \geq P(U_{nm})$
  - and if it is connected
    - Either $P(U_{ij}) \geq P(U_{kl})$ or
    - $P(U_{kl}) \geq P(U_{ij})$ or both
  - because probabilities are numerical and numbers are ordered which are a result of Kolmogorov's probability axioms (see Kolmogorov's role earlier in this chapter)
  - independence is still upheld as both item difficulty and person ability are independent variables
  - the Rasch model will uphold double cancellation (assuming it to be true) because as was shown above,
    - If $\theta_2 + \beta_1 \geq \theta_1 + \beta_2$ and
    - $\theta_3 + \beta_2 \geq \theta_2 + \beta_3$ then
    - $\theta_2 + \beta_1 + \theta_3 + \beta_2 \geq \theta_1 + \beta_2 + \theta_2 + \beta_3$ resulting in
    - $\theta_3 + \beta_1 \geq \theta_1 + \beta_3$ which upholds the double cancellation

The solvability condition implies that either the values of *a* and *b* are equidistant or that they are rational numbers (Krantz, 1964).; i.e. personal ability and item difficulty are continuous (Borsboom, 2005). The Archimedean condition limits the degree to which differences can be infinitely larger than any other difference within the conjoint matrix and this is independent of the column one wishes to inspect (Michell, 1990; Narens & Luce, 1986) resulting in person ability and item difficulty being unbounded (Borsboom, 2005). In effect, conjoint measurement takes its lead from analysis of variance models in which dependent variables vary alongside the joint effect of at least two variables. A crossed factorial design tests for the manner in which the dependent variable can be represented, either as a sum of the rows or columns and thus illustrates its similarities with this design (Perline, Wright & Wainer, 1979). Since the Rasch model is additive it is considered a form of fundamental measurement procedure utilised within psychometrics but as to whether it is truly scientific is questionable, according to Borsboom (2005) and Kline (1998).[90] Item response theory will be discussed below but it was necessary to introduce certain concepts in the conjoint discussion in order to elucidate the notion via such as model. Also, it was necessary to highlight the role played by conjoint measurement as fundamental measurement within such a model. This is because the Rasch model is used in newer models within change assessment research designs, such as Embretson's multidimensional Rasch model for learning and change. In order to maintain consistency in argument it needs to be illustrated that the Rasch model, as one instance of IRT, test theory is used in newer models and in particular can be utilised within a dynamic assessment framework thus upholding a scientific approach to the subject matter yet edging closer to fundamental measurement within a qualitative domain.

[90] It is interesting that Borsboom (2005) states that Kline (1998) has adopted the view that psychology can only be scientific if it is measurable. Yet Kline (1998) specifically states that what there is to fundamental measurement as viewed through the Rasch model (as substantiating fundamental measurement) is questionable in terms of its being scientific!

4.4.2 Test theory

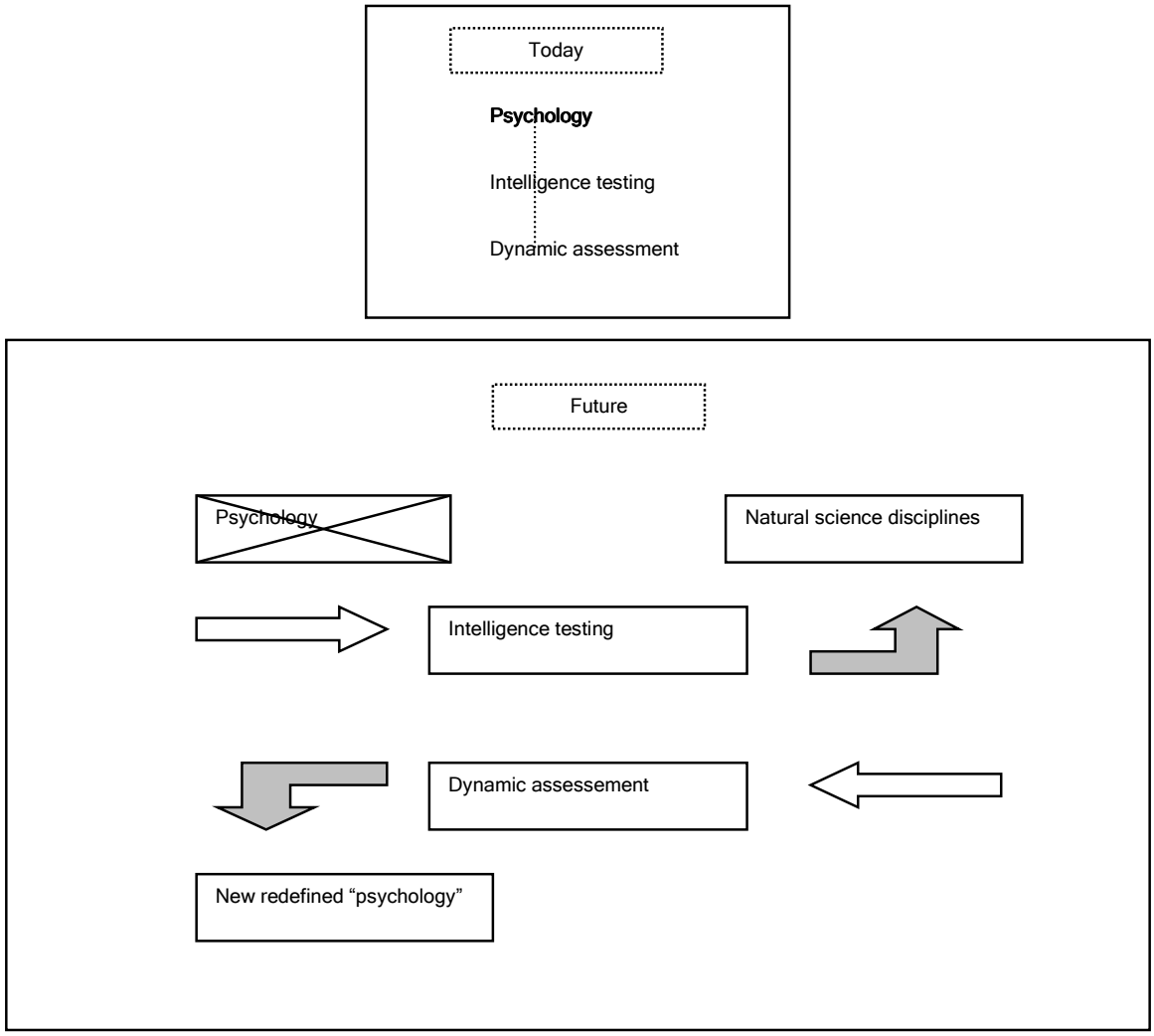
Blinkhorn (1997) humbly yet purposefully states that test theory is an undefined area of study which does not define what in fact takes place during test-taking but is a framework from which to view "categories of methods, models and techniques associated with the construction and evaluation of tests it is about attempting to fit tractable statistical models to test score data" (p.176). Test theory, he adds, is a collection of computation techniques with statistical manipulations attempting to give the enterprise credence. Reflecting concerns that Borsboom and Mellenbergh (2002) raise about the lack of concern regarding construct identification and the lack of clarity surrounding what test theory propounds as viable constructs, Blinkhorn (1997) admits of psychologists' unwavering regard for the supposed accuracy and meaningfulness of what test theory has to offer by way of tests. McDonald (1999) maintains that test theory is nothing more than theory underlying psychological tests or measurement which is in turn derived from psychometric theory. But most importantly is his regard for the mathematical and statistical undergirding of psychometric theory "it is necessary to recognise from the beginning that test theory is essentially applied mathematics, overlapping with statistics" (p.3) this underscoring the need to re-look mathematics and statistics within the social sciences. Embretson (1987) states that psychometric test theory is being increasingly considered as a science, hence the need to assess what exactly is meant by social science (see chapter 3). Does trait underlie behaviour? Should behaviour (including intelligence and potential) even be considered as traits or should such constructs be considered as states as Feuerstein advocates? Lord and Novick contend that it does (1968). Psychological processes, whatever they happen to be, physiological, behavioural, bioecological and so forth is not consistent with a theory of underlying traits and this is simply not reflected in test theory (Blinkhorn, 1997). State and trait models cohere with test theory models but process and mechanism are left unaccounted for. Trait theories conveniently dismiss situation (which dynamic assessment as manner of assessment does not). The trade-off between relevance in a modern scientific age is to allow for as parsimonious a model as possible and such models are often found in test theory (Meier, 1994).

Test theory's roots are located in individual difference research but has long since become reminiscent of statistical sophistication which has everything to do with large samples and population parameters – there is an irony in this state of affairs. The argument is such that no amount of statistical analysis and reconceptualisation can change what is necessitated, which is a measure of a psychological behaviour. Chapter 3 looked at the realm of psychological behaviour within science. No amount of rescue attempts at salvaging what is purported to exist within an ethereal realm can be consistent with a scientific approach and statistical test theory can perhaps be regarded as just one such attempt. If psychological theory cannot satisfy for itself a coherent definition of what it seeks to understand it seems pointless to bring in the services of other sciences in an attempt to rectify the situation. Psychology needs a theory consistent unto itself and cannot rely on methods and techniques of which it largely does not consist. Once again, psychological theory and statistical theory are investigating two very different areas of concern and through the interface of test theory it is assumed that contact is being made. Mainstream assessment purports to locate differences between individuals on test scores which is obviously what one is going to find in the manner in which it is secured. Psychometric theory's historical context is couched in intelligence assessment hence the emphasis on intellectual factors in the growth and development of various test theories, although self-report questionnaires were being drafted in the 1920's (McDonald, 1999). Dynamic assessment is predicated on another philosophy altogether and possibly should not (and in some instances does not) equate itself with individual difference research. Descriptive research is very powerful and often underestimated and one cannot deny glaring differences between various groups on various tests of intelligence for instance. The tests are so designed.

Dynamic assessment is not so designed and hence does not align itself with mainstream assessment. If mainstream assessment is so bent on statistically attuning itself towards finer discrimination between and within tests resulting in less rather than further progress, dynamic assessment should steer clear of yet even greater reform in the manner of assessment and test theory. Its place was never there to begin with and should subsequently not be there now. There exist alternative beginnings, different philosophies disparate tools and different methodology. Why the need to emulate mainstream trends which have in any event not proven terribly helpful or progressive? It is necessary to once again punctuate the story with yet another reiteration of the same sentiment which is being echoed throughout this thesis: psychology as a discipline needs to realign itself within a sphere which it can call its own and remain progressive. Currently, it is spread into domains with which it is clearly at odds and is struggling to maintain its existence. Dynamic assessment should be placed within an entirely different realm in comparison to current mainstream testing which itself should be placed within physiological and natural science methodology. There is nothing right or wrong about techniques utilised within such separate realms but the various movements within the discipline cannot co-exist in their current form, as the field will continue to be riddled with misgivings such as with which this thesis has been concerned. Figure 63 illustrates the reshuffling that needs to take place if dynamic assessment is to thrive in a progressive environment where it is unencumbered by the reigning tradition of psychology. Psychology should seek to dismantle its supposed unity as it is currently defined. The discipline already crosses so many boundaries that its existence is defined by nebulosity and uncontrolled spread. It should redefine itself according to its own progressive dictates and subsume dynamic assessment as method of assessment. Current intelligence assessment should shift over to natural science methodology where it can proceed unhindered by ill-defined behavioural constructs which are not and never have been defined in psychology.

Nevertheless, test theory's main preoccupation throughout its nascent years was the concern with reliability; reliability of the test as ascertained from parallel forms of the same test purporting to measure the same construct, or the same test delivered at different intervals over time (Spearman) which lead to work within classical test theory. Guttman's[91] work situated reliability in terms of items and further elaboration of this work led to the development of generalizability theory as an extension of classical test theory. The Guttman structure can be seen within IRT models which were to come later (McDonald, 1999) and was able to deal more effectively with binary response options with which classical test theory could not. However, Guttman's scales were only ordinal and unidimensional (akin to the common factor in factor analytical approaches; Bond, 2000) (Kline, 1998). Modern test theory then, is really an amalgamation of continuous development within psychometric test theory and each theory exists because of the need to further refine the theory. There is thus a common thread running through these, at times, parallel and sequential developments. Due to time and space limitations within this thesis the discussion detailing classical test theory and item response theory will assume a background knowledge pertaining to these theories. A brief introduction to CTT and IRT as it pertains to dynamic assessment is given in Murphy (2002) as well as Murphy and Maree (2006). This discussion will be detailed from the outset.

Figure 63 The discipline today and in the future: a possible scenario for maintaining both mainstream and dynamic assessment methodologies



[91] A sociologist. When one thinks about who contributed what to the discipline of psychology and why it was they did so, a clearer picture starts to emerge as to why they conceived of the methods they did.

4.4.2.1 Representational measurement: the legacy of Stevens' arbitrary powerful scales of measurement

Representational measurement is an apt description of what it entails; a representation of reality to numericism. Stevens (1946) eschewed the notion of additivity by stating simply that an attribute, represented numerically on a scale, would suffice as measurement (Kline, 1998).[92] Borsboom (2005) maintains that the representationalist view on measurement is in fact constructivist and not operationalist even though Stevens' scales are considered as operational. Operationalist and representationalist views are often considered synonymous (Poortinga, 1991). It is constructivist due to its scales being constructed representations of data within a highly restricted representational framework of what constitutes a measurement scale. By assigning numbers to observables via rules, he avoided the necessary aspect of concatenation. Four conditions need to be upheld if such representation can conceivably take place and is taken largely from the originators of representational measurement foundations, namely Scott and Suppes (1958) as well as Suppes and Zinnes (1963) (in Narens & Luce, 1986). Measurement obtains when the following is made manifest:

- An ordered relational and operational structure provides the bedrock of empirical reality such that
  - $\chi = < X, \geq S_1 \ldots\ldots\ldots\ldots S_n >$ where $\geq S_1 \ldots\ldots\ldots\ldots S_n$ are primitives of the structure. These primitives are the empirical relations on X.
- Axioms restrict the structure which reflects this truth about the empirical reality (recall axioms' functions as possibly underivable within its own system; that is, they are a given which needs to be accepted for theory or structure in this case, to be at all meaningful). Hence, Narens and Luce (1986) state that these axioms are putative empirical laws. This of course is the crux of the issue in which it has been stated over and over that the degree to which psychological attributes can be axiomatised within statistical and mathematical as well as measurement theorems remains questionable
- A numerically based relational structure is comparable to the above mentioned empirical structure. Namely,
  - $\Re = < R, \geq R_1 \ldots\ldots\ldots\ldots R_2 >$. R is the subset of real numbers and R_i represents the relations and operations comparable to the empirical relations and operations given above
- Lastly, proof of mapping is needed which illustrates that the ordering between empirical and relational has been preserved from χ to $\Re$. Structural preservation from one system to anther is conducted homomorphically in such a manner that the structure is now related as a scale; enter Stevens (1946)

Writing in 1946, Stevens nowhere mentions in his trend-setting article, the controversy surrounding the assignment of numerals to characteristics of behaviour that are not necessarily amenable to such treatment; in fact it is almost taken as a given that psychological constructs are amenable to such tactics. Taking his lead from N.R. Campbell's classical treatise on measurement (Kyburg, 1992; Ross, 1964) he states that measurement is largely a semantic exercise. This is easily surmounted when devising suitable scales for such measurement and his conception of measurement is essentially a representationalist one (Michell, 2002) which is really just a way of stating a rule for the assignment of numbers to concepts (Ellis, 1966; Ryan, Lopez & Sumerall, 2001). Stevens was either unaware of the potency of the counter argument or had chosen to sweep it under the rug of uncomfortable questions. Steven's operational scales of measurement have been with psychometrists ever since (Michell, 1997, 1999; Ross, 1964) upholding a firm positivist assumption of operationalisation equal to that of observed relations. This in no way diminishes his contribution to the clear understanding of the roles played by each scale but eschewing the issue of paramount importance as to whether one can truly measure something psychological in the first place is not good science. Having been influenced by Fechner's psychophysics and Spearman's quantitative science, the scene was thus clearly set for Stevens' scales of measurement to make its presence felt within a positivist mode and framework (Michell, 2002). He synthesised various emphasised aspects within works by Russell (1903), Johnson (1936) and Birkhoff on whom he leaned and utilised the former's numerical representation of order (Narens & Luce, 1986); Johnson's dual consideration of classification and ordering and the latter's theory of numerical transformations (Michell, 2002).

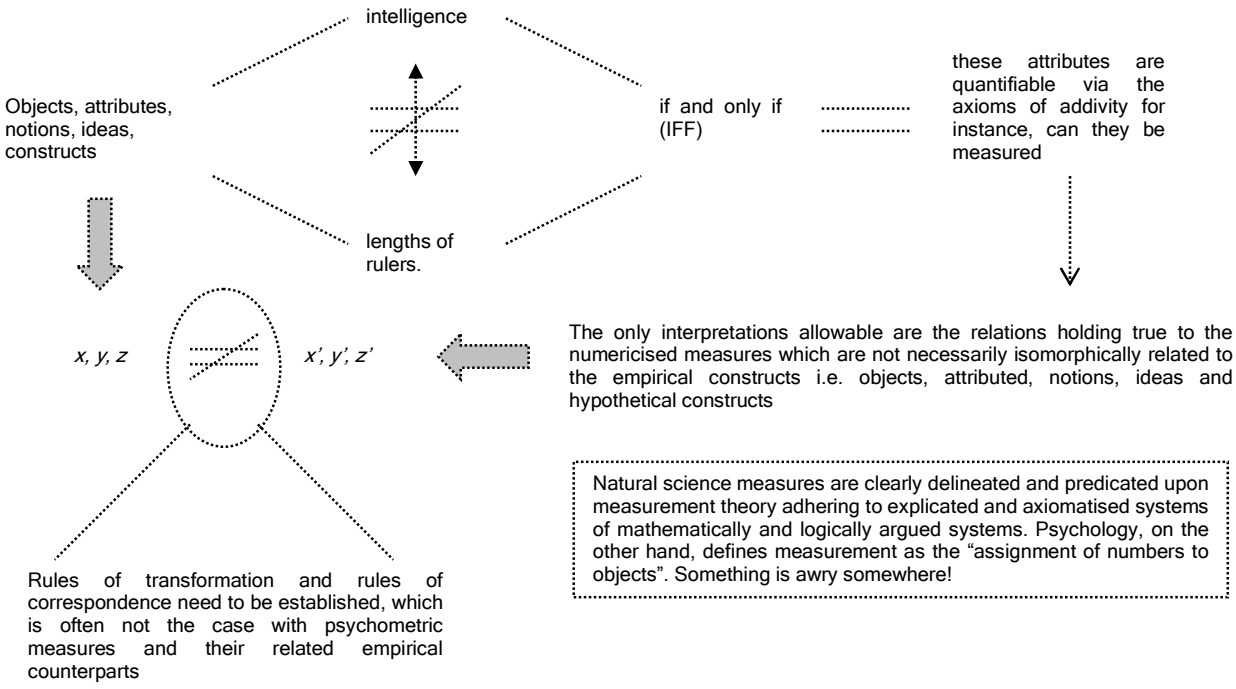
Appeasing the 1940 Ferguson Committee of the British Association for the Advancement of Science which for ten years deliberated the matter of measurement and setting about ascertaining the veracity of quantitative sensory measurement, Stevens managed to provide for his scales of measurement an operational veneer. He succeeded in circumventing the issue of quantifiable constructs in psychological measurement by assigning numerals to these constructs (Stevens, 1946) and described statistical procedures for the scales for which they were "permissible" (Velleman & Wilkinson, 1994). The committee's results were evenly split between those advocating that measurement in psychology was impossible versus those in favour of it

[92] This is reminiscent of the underlying rationale before Stevens; "if relations of qualitative increase or decrease are understood by analogy with quantitative change, and the mathematics of quantity is viewed as a mere formalism as suitable for applications in all disciplines, then psychological attributes that seem to be ordinal can be hypothesised to be quantitative (Michell, 2003, p.10).

(Borsboom, 2005). Steven's notions of measurement were, however, further refined by future generations of researchers who stipulated that it was the properties of psychological constructs rather than the constructs themselves that were being measured (Crocker & Algina, 1986). The committee did not really come to a decisive conclusion (Luce, 2000). The assignment of numerals or numbers to empirical events, was in Stevens' mind an isomorphic occurrence which is similar to the depictions illustrated in figures 34 and 54 as stated above (Ross, 1964). The scale transformations between ordinal and ratio for instance is an arbitrary one, as Ross (1964) highlights the very important fact that such transformations are applicable to the scale properties themselves but have no bearing on what is being scaled in the first place, a view that Michell (2002) shares. Hence, there was no explicitly defined philosophy behind the purported isomorphism between the formal aspects of the scale and the empirical domain to which these formalities applied. This lack of acknowledgment is a two-tiered fault (Michell, 1999, 2000) and is tantamount to the burying of heads in sand. The first error was in presuming answers to questions not empirically investigated (are psychological properties measurable in the sense of assigning numbers to them?) and the second higher-order error emanating from this error was the rank acceptance of a faulty definition by either ignoring it or glossing over its illogicalities (Michell, 1999, 2000). Figure 64 depicts the typical mode of operation of measurement. Note, however, that this illustration rules out any necessary equality between supposed quantified psychometric structures. The error continuously practised by psychometrists is in taking from a physicalist notion of measurement ideas which are absorbed in that system and transplanting these notions in another system altogether. As Ross (1964, pp.96-97) states:

> the kind of psychological measurement theory expounded here appears to differ from the corresponding physical measurement theory as the descriptive semantical rules of this kind of physical measurement theory all referred to empirical terms. This difference stems from the fact that our exposition of the physical measurement theories is based on the assumptions of *absolute reproducibility* of physical experiments, an assumption which is not fulfilled in practical work (own emphasis).

Figure 64 Rationale behind measurement: ill-conceived notions of quantifiable constructs



4.4.2.2 Classical test theory – a goal for inference[93]

The development of true score theory in classical test theory may be tentatively traced to the nineteenth century theory of errors (Stigler, 1999) during which time early statistical methods of treating astronomical data in the latter half of the eighteenth century were to meet later with the mathematical theory of probability early in the nineteenth century. The pivotal issues surrounding accurate data in terms of the relativism of observations in astronomy drove scientists to speculate on the relativism of 'correct' observations. Thus emerged the idea of focal points of observations being equal to 'truth' in addition to error (random and/or

[93] Regarding the early role of statistics in both natural and social sciences, Stigler states "what Newton had given astronomers and experimental design [had] given the psychologists: a goal for inference" (1999, p199).

systematic; Crocker and Algina, 1986) (Stigler, 1999, p.190). This is akin to the notions of later psychometric concerns with observed scores equalling true scores in addition to error: **X** (observed score) **= T** (true score) **+ E** (error score) (Borsboom & Mellenbergh, 2004; Crocker and Algina, 1986; Daniel, 1999; Embretson, 1197b, 1999; Kline, 2005; Marcoulides, 1999; Meier, 1994; Rust & Golombok, 1992; Scheuneman, 1991; Smit, 1996; Suen, 1990). The reliability (or 'attenuation' as Spearman first made use of this concept; Du Bois, 1970) resulted in a mathematical model originally developed by him in 1904 for the application in the area of intelligence measurement within the social sciences (Crocker & Algina, 1986; Du Bois, 1970; Murphy & Davidshofer, 1998).[94] "Despite the unity of statistics [the role played by statistics in the natural sciences vs. the role played in the social sciences], there are fundamental differences, and these have played a role in the historical development of all these fields" (Stigler, 1999, p.199). Classical test theory (CTT) can claim to be the earliest theory of measurement and is synonymous with classical reliability theory, true score theory, true score model and random sampling theory as its main aim is to estimate the strength of the relationship between observed and true scores (Suen, 1990). Common linkages between CTT and psychometrics are often made along with the assumption that the former is in fact the latter in totality (Gipps, 1994) and forms of assessment such as dynamic assessment have been lodged under the rubric of educational measurement. Due to the variability of the error score, possible changes within test scores can be ascribed to change or modifiability but because it is random it is without diagnostic value (Wiedl, 2002). The true score model, according to Borsboom (2005), is operationalist because the true score is defined in terms of the test score; in other words, its operationalises the notion of true score. In fact CTT and generalizability theory are considered as two approaches falling under the rubric of random sampling theory (Marcoulides, 1999). The underlying philosophy assumes statistical models which cater for an infinite number of testings thus allowing for a truer picture of the ability being assessed, or as Kerlinger (1981) puts it only an omniscient being would really know the true score. The mean score in CTT is the mean score of an infinitely long test (McDonald, 1999) (CTT's test-dependence; McCollam, 1998). This mean varies depending on the properties of the population being sampled thus according prime status to population parameters (CTT's group-dependence; Kline, 1998; McCollam, 1998), an aspect abolished in probabilistic item response models (Kline, 1998; Rasch, 1980). The deployment of various statistical techniques to achieve just this is the main rational behind CTT.

Reliability estimation, error variance and standard error of measurement lead to the estimation of a true score even though the only manifest score is the observed score. CTT only allows for ordinal scaling interpretation for the raw scores whereas item response models, via logistic functions, allow for interval level scaling (ability is estimated as a probability with comparable ability estimates for any calibrated item) (Embretson, 1983; McCollam, 1998). The total test is important to CTT whereas modern test theory emphasises the item, although there are latent trait models which are developed at the level of the test (Maxwell & Delaney, 1985). CTT principles are in fact derivable from special cases of IRT but the reverse does not hold (Embretson & Hershberger, 1999). Although it may seem that the death knell has struck for CTT, the future advancement and development of generalizability theory which encompasses CTT principles as well as IRT principles indicate that doing away with CTT would be rather drastic, notwithstanding the considerable number of test batteries still in vogue which are built on CTT principles. The situation of CTT and IRT is reminiscent of the parallel situation of dynamic assessment and static assessment; both models should be retained for what they can advance as they differ in their information yielding functions. Perhaps the most important unobtainable aspect within CTT is the theory's inability to consistently work out a reliable gain score from a pretest to posttest scenario. Utilising algebraic conversions, Pearson's correlations, variance of true scores and adhering to the central tenet of observed score equalling true score and error score, the reliability of the difference score is high only if the correlation between the two scores is low (Suen, 1990). Perhaps the most controversial aspect of CTT is the nature of the as yet improvable or at least the indefinable relationship between the true score and how it supposedly represents a construct. Once again, dynamic assessment needs not confine itself to measurement theories which have unresolved issues of their own. It does not necessarily have to align itself with any current theory of measurement in fact. CTT is underscored by the notion of a true score being the result of observations dependent on error distributions. Change can yet be accommodated within CTT as Wiedl (2002, 2003) illustrates in figure 65 below but there is a need to determine and identify the correct model within different dynamic assessment set-ups; change within IRT will be assessed below. Agreement of true scores with constructs is referred to as the fallacy of Platonic true scores which, if recalled, made its first appearance in the discussion on the Platonic realm of mathematical entities. The Platonic true score is not consistent with CTT (Borsboom & Mellenbergh, 2002). This line of argument takes us back to the whole notion of the representationalist view of measurement. CTT as incorrectly assumed by some psychometrists, is thought to equate the realm of true scores (the syntactic) with the construct (the semantic). Figure 66 illustrates this erroneous equating of true score to construct.

[94] What sent the social sciences down a path discernibly tangential to the natural sciences in terms of statistical manipulations of data was the idea (predating even Fischer's work on agricultural experimentation) of randomised experiments (Stigler, 1999); as exemplified in the work of Charles S. Peirce who made use of a version of an experiment already developed by Fechner (an experiment on sensation and stimulation; Sahakian, 1981) and in so doing provided an even more robust result due to his use of a blind randomised experiment (Stigler, 1999).

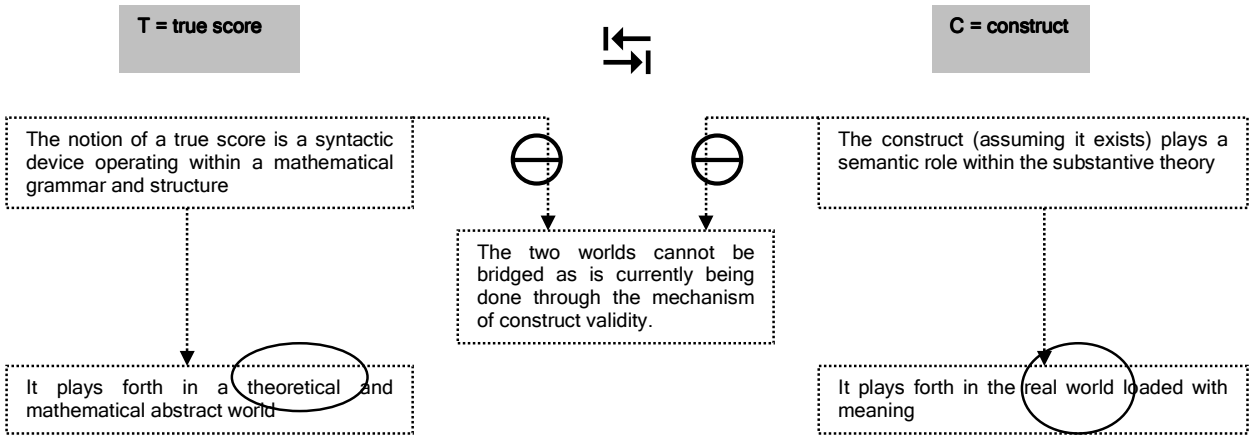
Figure 65 Change of performance in terms of CTT (Wiedl, 2002, 2003)

1. $X = T + E$
2. $X = T_S + T_v + E$
3. $X = T_S + E_{RAN} + E_{SYST}$
4. $X = T_S + T_V + E_{RAN} + E_{SYST}$

X: performance on test
T: true value
E: error term
s: stable
v: variable
ran: random
syst: systematic

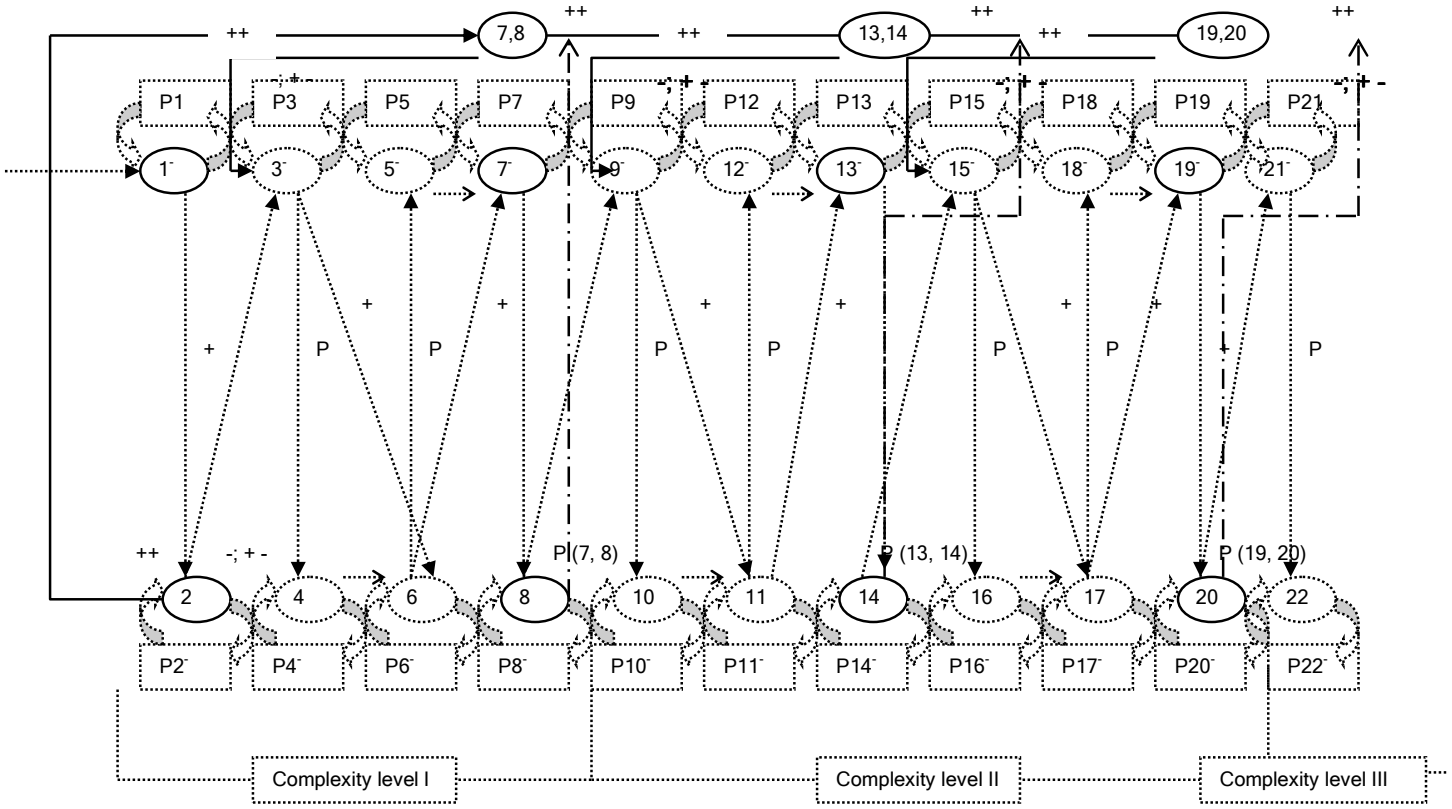
Possible ways of how change can be accommodated while still remaining within CTT tenets. The problem is to determine which of the models is the most accurate given the circumstances. Wiedl (2003) is very cautious however when he likewise cautions the reader to the as yet ill-defined construct being assessed; "no assumptions are made with regard to what is measured and the implications of this. The definitions thus cover different theoretical concepts" and refer, in turn to, the "'zone of the proximal development', 'learning ability', 'cognitive modifiability', 'baseline reserve capacity', 'developmental reserve capacity', 'responsivity to intervention' and 'rehabilitation potential'" (p.95).

Figure 66 Erroneous equating of true score and construct



Change can be captured within CTT and depending on what the goal of the assessment/intervention is, success need not be the outcome of IRT change-based models (or process models; Guthke, 1982) alone. The context, as with most things psychological, is prevalent and largely determines the what/when and how of assessments, so if the nature of the situation calls for CTT-based models then there are at least such dynamic assessment models which attune themselves to this framework. Wiedl (2003) mentions two such dynamic assessment paradigms located within CTT; the test-training-test paradigm and the paradigm of continuous integration of interventions into the test, the latter will now be briefly discussed at it pertains to dealing with the construct of change within CTT (compare to IRT). A partial glimpse of Guthke's Diagnostic Progam for the Assessment of Reasoning abilities (ADAFI) is illustrated below in figure 67. The construct of change is discussed further in section 4.4.2.2.1 but is brought in here due to its applicability within CTT.

Figure 67 The integration of dynamic assessment into the testing process: Guthke's ADAFI (Wiedl, 2003, p.96)



The ADAFI is still housed under CTT and thus conforms to its leading tenets so espoused and endorsed by the popular rendering of what it means for psychological measurement to be objective, valid, verifiable and in general robust. Change is accounted for via a process of correct and incorrect answers to items and is mediated during promptings. This model is cognisant of the task characteristics, the testee's ability level and how these parameters vary throughout the testing process (Wiedl, 2003). However, change as understood within change-based IRT models as another construct is not the same here - "the focus is not on change, but on the level of performance that can be achieved and the amount and quality of hints or time that the testee needs" (Wiedl, p.96). Nevertheless this manner of assessment is still dynamic. The point now would be to compare the IRT change-based models for dynamic assessment to ones such as this, offered by Guthke and seek the more attractive alternative which is dependent on the critical notion of context.

follows on next page ...

Continued from previous page …

The algorithm is easy to follow and follows a logic which encompasses both item and person information. For instance, in keeping with the above model the following algorithms can be specified for the following items or questions (Q):

- If Q1 is correct move to Q2; if Q2 is incorrect prompt and try again; if Q2 is correct on first try move directly to Q7 and Q8; if Q7 and Q8 incorrect move back to Q3 and if correct move to Q6 (Q5 is presumably too easy at this stage for the person following this pattern); if Q7 and Q8 correct on first try move directly to Q13 and Q14 and likewise carry on in this fashion
- If Q2 is incorrect prompt and try again and move onto Q3; If Q3 is incorrect prompt and try again; if Q3 is still incorrect prompt before moving on to Q4; if Q3 is correct in first attempt move to Q6 (note that this person follows the pattern for the person above but only after having reached Q6)
- Q6 is the first location where an item is skipped presumably to determine the level of functioning; this is most likely done purposely. Respondents answering from the Q3 and Q4 are both directed to Q6 thus skipping Q5, however, should Q6 pose problems, prompts will direct respondents back to Q5 or else they proceed to Q7 directly
- The algorithm proceeds in this fashion for the rest of the items
- Item difficulty is known as well as person performance and in this manner it is very similar to the notions underpinning IRT change-based models discussed below

The construct of intelligence as presumably measured by various intelligence tests cannot be equated as they are expressed as functions of operations which differ between such tests. “Intelligence may be a psychological construct, but its explication requires that we don the hats of other kinds of investigators and try to integrate their varying perspectives” (Gardner, Kornhaber & Wake, 1996, p.134). Also “I do not believe that psychology can be reduced to the brain; but I do believe it can be put into correspondence with it and this should happen, because study of the brain can help to clarify the postulates that our psychological theories must have” (Pascual-Leone, 1997, p.80). A question which is implicitly implied but one which is not often stated directly, is the usefulness of intelligence, which, would seem at first glance to be quite obvious, intelligence aids our survival (in whichever milieu this survival happens to express itself), which is echoed in what Johnston, Partridge and Lopez state; “the value of intelligence lies in its ability to allow *rapid adaptation* to occur within the life-span of the individual organism” (own emphasis) (1985, p.487). Thus in extending this argument further, is ‘rapid adaptation’ nothing more than learning to adapt, thus learning to train oneself; learning to ‘learn the situation’ so to speak in as quick a time-span as possible? Is intelligence, by extrapolation then, nothing more than learning to change? If this is true, then assessing for change should be a main concern within intelligence testing.

This argument, of course, hinges on the above definition of what intelligence is. Intelligence assessment is almost entirely dependent on how one chooses to define it. If the true score can not be equated with the construct score as argued above, then it stands to reason that each construct score represents something else and does so via different operationalisations. In essence, an intelligence score on one test cannot be equated with an intelligence score on another test which leaves one with the uncomfortable question of what the tests are measuring. CTT expresses a construct only in so far as it pertains to specifics which is tantamount to saying that depending on which method you use to derive the gravitational constant, the result will not remain constant but will need to be re-examined and considered from a specific operationalist point of view. Such variation cannot be good for any discipline calling itself a science. “CTT-derived scores predict performance only at the point represented by the obtained test score. Rasch-based interpretation predicts performance on any task given its scaled distance from the individual’s measured ability” (Woodcock, 1999, p.126). The need to trace an underlying variable which can be equated across tests is found in latent trait models. Parametric modern test theory models are instances of generalised linear item response theory of which the Rasch model is one such example (Borsboom & Mellenbergh, 2002). CTT true score resembles the IRT latent trait but whereas CTT is an unrestricted model IRT is not in terms of placing restrictions on the observed data. CTT posits hypotheses of true scores based on an infinite number of attempts to observe it, whereas IRT does not. However, Blinkhorn (1997) is not convinced of the veracity of the Rasch model in terms of what it can offer over and above CTT and he maintains IRT’s stance as redolent of psychophysics. He refers to its supposed benefits as a mirage due to the model’s inability to fit all types of data. There is thus disagreement within the literature over the various CTT and IRT models. This is precisely the point made in chapter 3 in which higher order problems are being felt in the chain in the lower areas within the discipline. The errors need to be rectified at the top. CTT and IRT models are robust in many instances, are mathematically sound (most of the time) and statistically amenable to manipulation but no amount of bolstering of techniques can do much to solve a largely philosophical dilemma emanating from the top most level of concern – the discipline itself.

4.4.2.2.1 Malleable stability - the gain score issue

"Th[e] gain score information has a controversial history in the psychometric world because of the unreliability of the scores and the dependence of the posttest score on the pretest" (Lidz, 2002, p.123). The paradox of the gain score can be summed up as follows: if a test-intervention-retest research design based on changeability via dynamic assessment theory yields low psychometric reliability then this in an indication of poor test construction. Score changes thus threaten psychometric properties of tests under the classical test mode (Hamers, Hessels & Pennings, 1996). However, pretest and posttest scores within a mediatory context are supposed to evidence poor reliability for the sole reason that the ability level undergoes changes due to inherent potential manifestation (via practitioner action; Elliott, 2003) which would otherwise never become evident (Embretson, 1992, 2000). The use of IQ tests within educational settings is its greatest liability (Ramey & MacPhee, 1981), which is, frankly, bordering on the absurd! Stability and modifiability models therefore preclude each other and in this respect are in stark opposition (change brought about by an intervention during a test is considered a threat to psychometric validity; Hamers & Sijtsma, 1995; Sternberg & Grigorenko, 2002) but can also be viewed as complementary depending on the nature of the measurement context (Jensen, 1992). Nevertheless it can be unequivocally stated that "A primary goal of LP (learning potential) assessments is to defeat the very predictions made by traditional IQ scores" (Glutting & McDermott, 1990, p.398). Recall that psychometrics developed specifically to test for a unitary concept of underlying intellectual functioning (Davis & Anderson, 1999). Dynamic assessment's posttest scores are often enhanced predictors of future static and dynamic assessment intellectual achievements which results in asking the question of how valid the original static (pre-test) scores are in the first place (Tzuriel, 2000a). Traditional learning theory only accounts for changes in simplistic posttest – pretest score scenarios and it often happens that initial task scores are more variable than posttest score results which evidence less covariation over time and practice periods (Ackerman, 2005). CTT's preoccupation with stable individual differences is particularly evident within psychometrics (Jensen, 1992). Wilder's Law has been recognised for some time, which attests to the fact that a gain made over a testing situation is a function of his initial ability (Guthke, 1993b). Vernon (1952, 1954 in Lidz, 1987a) had conducted research into the gain score issue as early as the 1950's detailing research dated to 1920's. Mainstream psychometric theory by and large assumes a stable trait underlying ability (Irvine & Dann, 1991; Jensen, 1992) which should not change and classical test theory's *g* undergirds this philosophy. Dynamic assessment's predicates are however entirely different and the possibility of the construct changing exists in a framework where such change is hoped for (Sijtsma, 1993b). Recall, however, "the fallacy of inferring fixity in ability from constancy in status scores" (Lohman, 1997b) which argues for malleable ability even though scores may indicate otherwise. Modern test theory, through model development which encompasses changes underlying ability as latent trait, is able to cope with this very situation (Embretson, 1987) and is expanded on later. Three main issues within the gain score debate can be summarised as follows (Embretson, 2000):

i. The paradox of reliability which is highlighted in the absurd situation of decreasing reliability of supposedly stable traits from pre to posttest situation; the so-called "unreliability-invalidity dilemma" (Bereiter, 1967). CTT requires stable reliability from one situation to the next. Dynamic assessment does not fit the CTT model well and as such evidences decreased reliability from pre to posttest situation. What is needed is a psychometric model which accounts for change as reliable. In this regard Bereiter (1962, p.7) sums up by stating that "the functions of prediction and of education are opposed rather than complementary" where dynamic assessment can be regarded as educational and mainstream psychometrics the predictive area of concern. CTT's source of error is dynamic assessment's bedrock of method! Bereiter wrote this in 1962, and IRT response models were being propounded in various forms as early as 1952 - 1960 (Embretson, 1997b) so it is a curious fact that dynamic assessment has had to wait so long for awareness into change properties to surface. Bereiter (1967) sites the situation as being the cause of deficient statistical methodology, which within CTT it is, but not within IRT as expanded on below
ii. The change evidences a negative false correlation with initial status (Bereiter, 1967; Embretson, 1991b, 2000), the so-called "over-correction-under-correction dilemma" (Bereiter, 1967). McDonald (1999) states that if the underlying trait between a pretest and posttest score is similar the error variances are additive but it is contested whether the same trait is being measured
iii. Scale units do not have a consistent meaning; there is no common metric (Cronbach & Furby, 1970). Large changes from an initial low ability as opposed to small changes from an initial high ability are not comparable which makes meaningful conclusions irrelevant. There is a need to ensure that a change of "one unit" at a low initial ability is equivalent to a change of "one unit" at a high initial ability level. Bereiter (1967) refers to this as the "physicalism-subjectivism dilemma". In other words should the scale units remain the same or should they be arbitrarily changed to accommodate changes (physicalism) or should scale units change along with changes in the construct (subjectivism)? It is noteworthy at this juncture to point out Bereiter's early 1967 claim that psychologists have tended to avoid working in areas such as this due mainly to the perceived insurmountable problems evidenced from the lack of available statistical technique, which was already being developed at that time in any event. Nevertheless, as Michell (1999, 2000) was to echo three decades later in his trenchant critique of psychologists simply ignoring these issues, it seems that such ignorance is rife through time. Differences in error measurement exists between initial and posttestings which nullifies any attempt to equate the two scores (Embretson, 1987)

Recall that most scales utilised within psychological research designs only ever reach ordinal scaled measurement as designated by Stevens (1946). Interval level scales are necessitated to make manifest and meaningful the differences in change scores. Lastly, measurement error as calculated within CTT framework posits a population parameter error source from variance ratios and hence standard error of measurement applies to all scores whereas in IRT the standard error of measurement, although generalising across populations, differs across response patterns (Embretson, 1997b). Most of the South African research in dynamic assessment surveyed in 2002 treated gain scores in the manner befitting CTT and not IRT (Murphy, 2002; Murphy & Maree, 2006) and at times treated gain scores in the fallacious manner of simply subtracting the pretest score from the posttest score and measuring change as simple gain (Cronbach & Furby, 1970; Embretson, 1987) or realising "potential" as initial scores subtracted from gain scores (Swanson, 1995). One notable exception was the study by De Beer (2000) which employed IRT via computerised adaptive testing to construct and validate a dynamic assessment instrument. The three issues mentioned above are framed within CTT framework which is precisely where the problem lies. Lidz (2002) states that one way in which the gain score issue can be approached is to ascertain statistical information regarding both the pre- and posttest errors of measurement. "If the obtained score outpaces either of these, there can be more confidence in the significance of the gain, and the issue of the practice effects can be ruled out" (p.123). However, Lidz (2003) emphasises the time and mathematical expertise that will be needed to implement various models which can accommodate change within their models; most dynamic assessors are, after all, practitioners working in the field. Change scores which reflect intra-individual changes within a test situation were critically reflected on as early as the 1930's in Europe (Wiedl, Guthke, Wingenfeld, 1995). Early works in Europe that were veering in the more process oriented direction in terms of assessment were pressing for conceptually different techniques of assessment, although these thoughts had, as yet, to manifest in actual test application (Wiedl, Guthke & Wingenfeld, 1995). Early on Cronbach and Furby (1970) championed the residualised gain score recommended by DuBois (1957 in Cronbach & Furby, 1970). Here, the posttest score is expressed as a deviation from the posttest-on-pretest regression and the remainder or residual information of the posttest score that is not linearly predictable from pretest scores is taken as the change score. This was done in an attempt to appease the then current testing establishment in terms of extracting true gain from pre to posttest measures. Residuals remained free from the linear predictors made on the basis of the pretest scores. Change-based measurement, should by its very nature, estrange itself from stable trait measurement. Yet another reason for dynamic assessment to disassociate itself from mainstream measurement (Meier, 1994).

4.4.2.2.2 Generalizability theory

A gradual development of theory resulted in CTT's vague conceptualisation of random error into a more contextualised account of error specific to the context at hand and although contributed to by many psychometrists, generalizability theory's origins is cited as culminating in the work of Cronbach, Gleser, Nanda and Rajaratnam (1972) (in Marcoulides, 1999; McDonald, 1999; Murphy & Davidshofer, 1998; Suen, 1990). The theory allows for a greater distinction between multiple simultaneous sources of error by resolving error into various components and determining the degree to which each component contributes error and yields more information about the measurement procedure as opposed to CTT (Cronbach, 1990; Marcoulides, 1999) hence its reference as a theory of multifaceted errors of behavioural measurement (Marcoulides & Drezner, 1997). Clearly, the influence of analysis of variance and the role that this method of data analysis played in the development of generalizability theory is obvious in its emphasis on levels of analyses (or facets within generalizability theory). In essence it is a statistical theory which is concerned with the dependability of measurement (Marcoulides, 1998) and has developed new ways of thinking of reliability (Crocker & Algina, 1986). It seeks to know how generalizable such measurement is (Lunz & Lincare, 1998) and is an extension of CTT (Embretson, 1999) but is more liberalised in its approach (Marcoulides, 1999). Apart from generalizability theory, the other most common model for analysis involving judgements is multifaceted Rasch measurement (Lunz & Lincare, 1998; Marcoulides & Drezner, 1997) which accounts for variations in such a way that bias within various contexts is removed from score interpretation. Reliability is considered within the test situation and error is considered the product of the test scenario.

Generalizability theory makes a less strong assumption about parallel tests and refers to this assumption as the domain-sampling approach. As a theory of reliability, CTT assumes that tests are random samples of items from a domain (Nunnaly, 1978). Underlying this assumption is the notion that tests can be said to be randomly parallel if the items are random samples emanating from the same set of possible items (Suen, 1990). The ability to generalize from one set of results to another is the main focus of this theory and its initial aim sought to answer the question of the degree to which generalizations could be made from one set of results to another in various conditions (Murphy & Davidshofer, 1998). Analysis of variance is employed in generalizability theory in an attempt to account for multiple sources of error as opposed to the more conservative CTT approach (Crocker & Algina, 1986; Marcoulides & Drezner, 1997; Kerlinger, 1981). Other methods estimating variance components (a major aspect within this theory due to the information provided on error source variance) exist other than analysis of variance, such as Bayesian methods and restricted maximum likelihood methods, but analysis of variance has proved to be the easiest to handle (Marcoulides, 1999). Generalizability theory can account for various reliability queries whereas CTT can account for only one reliability measure, namely, Pearson's correlation. Variance plays a major role in this theory due to the manner in which it accounts for accuracy (reliability) of measurement across scores; in other words how error is treated (Murphy & Davidshofer, 1998). Within CTT the rationale behind measurement is to assess people whereas generalizability theory accounts for the assessment of multiple sources of error because it is able to handle more dimensions than CTT. If measures change or new

dimensions are added, generalizability theory is capable of withstanding these extra sources of error, however the analysis will change depending on the context. By generalizing from one scenario to another, decisions are based on pre-empting measurement errors in new applications which is where the familiar G-study (from the generalizing of error) and D-study (the decisions subsequently carried out from g-studies; Crocker & Algina, 1986; McDonald, 1999) originates. Error source is extrapolated back from the use of the specific test, underlying the call to understand why and how the test is utilised rather than merely looking at scores (Murphy & Davishofer, 1998; Rust & Golombok, 1992; Suen, 1990). This is perhaps its biggest drawback however as few practitioners have the resource to continuously revalidate tests (Rust & Golombok, 1992; Kaufman, 2004). Moreover, it is considered a conceptually different approach to CTT as opposed to being statistically different (Murphy & Davidshofer, 1998). These added dimensions are accounted for in this theory. Variation across one such dimension can be construed as true variance and the rest are due to measurement error or facets of measurement. These facets are the equivalent to factors as utilised within analysis of variance. Facets are applicable within and across domains and hence the theory accommodates multiple levels (McDonald, 1999).

Depending on the number of dimensions, the number of facets of measurement will vary. Research designs employing multiple levels of assessment are amenable to such analysis especially where more than one rater is necessitated by a test across more than one level of assessment; an issue which has plagued dynamic assessment inter-judge reliability when attempting to identify cognitive deficits (Reschly, 1997). Although it is strongly advocated that IRT be looked at with greater regard by dynamic assessment researchers, generalizability theory can offer qualitative and more clinically aligned dynamic assessment practitioners greater scope in adjudicating reliability of scores across multiple levels of analysis depending on how the dynamic assessment intervention proceeds. *X* raters can assess *y* individuals on *z* instruments (including observations) across a number of occasions. Embretson (1999) points out that unlike most applications of IRT, generalizability theory considers the impact of various measurement conditions on the trait level and thus can analyse the differential effect of various measurement conditions on the psychometric properties of the test. Marcoulides (1999) highlights his work into models of generalizability theory which he views as an extension of the theory as well as a special instance of IRT especially where judgments play a role in scores. His extended generalizability theory is able to account for latent traits such as ability and item difficulty as well as rater severity. Wilson and Wang (1995) investigate via their model of IRT the issue of different modes of assessment and include rater estimates with some very interesting results as they compare multiple choice items, rater estimates and open ended questions and the resulting IRT functions. Embretson (1999) cites this new work in generalizability theory as not having been thoughtfully considered by the broader community and although there is literature on the matter, it is relatively recent. It is emphasised that attention be turned towards the newer models of generalizability theory as well as IRT models within dynamic assessment, specifically generalizability as it can accommodate qualitative ratings which is an area of concern to dynamic assessment in its more clinical forms.

4.4.2.3 Modern test theory

The ideas behind modern test theory as exemplified in item response theory are not new and extend back to Spearman's 1904 paper on factor analysis, who in addition also pioneered many procedures within CTT (Borsboom, 2005; Embretson & Reise, 2000). The notion of a latent trait underlying various observed behaviours has remained with test theory in a variety of guises (having been theoretically and practically constructed in more primitive form in the 1920's in Thurstone's work; Embretson, 2004). Its modern renditions still assume a latent trait to operate meaningfully within instruments assessing for various traits (Borsboom, Mellenbergh & Van Heerden, 2003) and in many instances this underlying trait is considered unidimensional (Hambleton & Rogers, 1991). However below will be shown the need for and development of multidimensional trait models specifically to estimate changes in ability, or potential. Interest in latent trait models dates to the late 1970's and early 1980's but due to its perceived technicalities only later filtered through slowly into the realm of intelligence research (Whitely, 1980). Kormann and Sporer (1983) for instance comment on the intractable problem of dynamic assessment test validity and the lack of a general theory of change measurement even though attempts within IRT had already started to address these issues; but in all fairness, the techniques were decidedably under-developed as far as technological implementation was concerned and the authors do note the then-available contributions to the debate.[95]The main difference between latent trait models and CTT is that the former attempts to account for the mechanisms which generate the data (Borsboom, 2005). In other words, it is a top-down model or process as latent trait models or theory illustrates how parameters relate the latent variables to the data according to the data generated under the particular model (model fitting) (Borsboom, 2005). This is of particular importance currently as psychometric theories of intelligence have mostly prescribed construct definition via statistical and measurement techniques rather than the construct defining the necessary psychometric structures (Hunt, 2005). However, in all its sophistication the broader field of IRT is nevertheless concerned with mathematical models of ability traits (Carroll, 1996). Moreover, scaling is

[95] Interestingly enough, one of the researchers cited as having contributed to the measurement of change debate in Kormann and Sporer's (1983) article is one R. Schwarzer whose meta-analytic software programme is utilised in Appendix 1. Also of interest is the fact that as early as 1945 analyses at the level of item responses was considered as warranting attention from intelligence researchers (Rapaport, Gill & Schafer, 1945 in Kamphaus, Petoskey & Morgan, 1997).

justifiably interval-level based, parameter estimates are invariant and a common scaling of items and individuals is made possible (Embretson, 1996).

Paralleling trends within the factor analytic tradition, IRT models developed in various ways resulting in models that could deal with dichotomous variables, polytomous variables as well as categorical variables (Borsboom, 2005). Conditionalising on the latent trait scores (assumed to underlie the observed variables) will result in the statistical independence of the observed variables and is referred to as the principle of local independence, specifically within a subpopulation located at a "single point on the latent trait scale" (Crocker & Algina, 1986, p.343). Local independence allows for information pertaining to the conditional individual responses to be sufficient in determining the conditional probability of the response patterns for a set of items (Glas, 1997; McCollam, 1998; Van Schuur, 1997). In other words, responses to items are independent of one another (Embretson, 1983; Pfanzagl, 1994). If the latent variable is held constant the probability of endorsing two or more items is the product of the probabilities of endorsing each item (Coombs, 1967). Factor analytic treatment, confirmatory factor analysis, the parallel development of latent variable analysis with dichotomous variables and the resultant growth of item response theory utilised for both continuous and categorical and polytomous observed variables can all be traced back to the fundamental notion of a latent construct (Borsboom et al., 2003; Heck, 1998). The notion of a latent construct is a burdensome assumption to make especially as so much theory growth and development has centred around it and Cronbach and Furby (1970) mentioned similar comments as to the unlikely event of ever truly defining constructs via psychometric models despite their sophistication and development. This was stated over thirty years ago and is still being heatedly discussed in the literature today. For instance how is the measurement of latent traits to be interpreted? Borsboom (2005) offers two interpretations, which he states, is largely semantic.

i. Stochastic subject interpretation
   a. The person is central to this interpretation. Item responses are regressed on the latent trait and the probability of a correct response will change as the latent trait changes
   b. Similarly to CTT though, the score on item *j* for subject *i*, $\Sigma (U_{ij})$ is exactly the same as *i*'s scores on *j* IF the latent variable model is true
   c. Characteristics of individuals are being modelled

ii. Repeated sampling interpretation
   a. Here the latent trait model is reconceptualised as a repeated measures interpretation where the repeated sampling is central
   b. The focus is on populations as opposed to individual subjects as in the stochastic interpretation
   c. This interpretation focuses on the repeated samplings of item responses from populations evidencing probability distributions which are conditional on the latent variable. These populations all have the same position on the latent variable and the parameters are thus related to the latent variables
   d. Repeated sampling from the same population will thus result in sub-population means
   e. The crux: it is not the individual subject's probability that is of concern; rather the focus in this model interpretation is the probability of drawing a person that endorses the item from a population with the latent trait level sampled
   f. In other words; population means on θ is distributed *f* over item responses u_{ij}; hence the expected item response is $\Sigma (u_{ij} \mid \theta_l)$
   g. Person-level random variation is not accounted for
   h. Sub-populations means are being modelled

Substantiating the theoretical entities which are assumed to underlie the observables (learning potential) is the task of psychological assessment but doing so through conventional means, such as IQ test concurrent and predictive validity (Embretson, 1992; Fernández-Ballesteros & Calero, 2000; Guthke, Beckmann & Stein, 1995; Hamers, Hessels & Tissink, 1995; Hessels, 2000; Jensen, 2000; Resing, 2000) or correlating results from widely used intelligence batteries is dangerous in terms of ill-defined initial constructs present in these original studies or criteria (Guthke, 1982; Kline, 1998; Pennings & Verhelst, 1993; Wallis, 2004), as has been argued in this thesis thus far. This manner of validating the construct of intelligence is circular in argument (Heitz, Unsworth & Engle, 2005). Ratther, performance change criteria should be utilised (Hamers, Hessels & Pennings, 1996). Guthe, Beckmann and Stein (1995) conclude that an external concurrent valid criterion is unlikely ever to be found for learning potential tests and as such construct validation should proceed along the lines detailed by Cronbach and Meehl (1955). This author maintains that such external learning potential criteria can be located or developed and will not necessarily have to resort to nomological networks of findings. As mentioned, researchers such as D. Borsboom concur. One need only look towards the reductionists within physiological psychology to find common ground or other process-based measures of intellectual development. The utilisation of change-based IRT models is also another avenue to pursue (more on this below). Criterion validity within dynamic assessment thus poses a difficulty within the mainstream set-up (Lidz & Thomas,

1987). Dynamic assessment would perhaps be better off in utilising ipsative[96]assessment, which is of course a move away from standardisation. If, as the Spearman effect attests to, tests loading higher on *g* tend to discriminate more regarding differences between people, such *g*-loaded test are already biased against the supposedly disadvantaged groupings dynamic assessment intends to assess as it has been reported in the literature over a span of many decades that certain groupings tend to perform poorly on intelligence tests in comparison to other groups (Loehlin, 2004). Why would we stack the odds against ourselves by utilising such *g*-loaded test in the first place? Of course there is no simple answer to this particular dilemma. Embretson (1983) refers to this concurrent construct validating procedure as the nomothetic span of the test which is indicative of the degree to which the test can differentiate between individuals. Intelligence tests usually have very good nomothetic span[97] but are not always able to specify the underlying construct being assessed. This is in addition to the literature replete with near zero correlations between tests assumed to measure the "same" construct (Guilford, 1982). If the underlying construct of initial ability, intelligence, is different to the underlying construct of learning potential then it is hardly surprising that the two do not correlate meaningfully in many instances as the constructs are supposedly measuring different traits (Guthke, Beckmann & Stein, 1995).

Does the substantive theory necessitate the model (Borsboom & Mellenbergh, 2004)? This is echoed in the same plea for substantiating the theoretical hypothesis within null hypothesis significance testing, which as we have seen above, is not a falsifiable manner of progress. IRT models are in fact falsifiable (one of the rare occasions within the psychological discipline where such as situation prevails) as the IRT model employed to fit observed data may not necessarily fit the data and thus not adequately predict or explain the data (Hambleton, Swaminathan & Rogers, 1991). Does the latent trait underlie the observable variable or do we construct the latent trait from the observable? Borsboom et al, (2003) discuss at length the philosophical issues surrounding just such a notion by ascribing to various interpretations (constructivist and instrumentalist) but agree that realist interpretations of this notion can be the only solution. The latent trait can never be assessed directly (otherwise there would be no use for tests!) but it can be assessed in terms of the joint probability of the results it implies via joint measurements made indirectly (Michell, 2000). We still have not adequately addressed the issue of which is paramount. The trait's manifestation through observables or our constructing of the trait via the observables. Addressing the latent trait from a constructivist position entails a dependent trait and not an independent trait which is of course not sound science and secondly, addressing the trait from an instrumentalist position results in numerous operational definitions for the very same construct (as each operation will need to be defined for each test) (Borsboom, Van Heerden & Mellenbergh, 2003). Entity realism ascribes to a correspondence of truth view and as mentioned is the view taken by the authors. Purely mathematical and statistical analysis will relinquish the need to even describe the trait in any form other than one of mathematical concern. However, psychology is clearly not a sub-discipline of mathematics and it is very easy to see how the discipline often loses itself in a morass of statistical and mathematical trickery.[98]

The Rasch model as exemplified within item response theory is, as mentioned above, a form of conjoint measurement as it ascribes to the necessary requirements for additivity over and above its ordinal status. The Rasch model is not a model constructed in isolation from its historical context. "Test-free linear measures were latent in Campbell's 1920 concatenation, In Fisher's 1920 sufficiency, in Levy's 1937 and Kolmogorov's 1950 divisibility; clarified by Guttman's 1950 conjoint transitivity; and realized by Rasch's 1953 additive Poisson model" (Wright, 1997b, p.43). Borsboom and Mellenbergh (2004) state that the Rasch model is not the only IRT model to avail of itself of conjoint measurement and that there are other probabilistic latent variable models which can be used. Ability and difficulty level are both continuous quantitative attributes in a monotone model of which the Rasch is one such example. If Jane's probability of endorsing an item increases with the difference between her ability and the difficulty of the item, then so it must be with Peter also, dependent on the fact that the difference between Peter's ability[99] and the difficulty item is not less than the difference for Jane. This is referred to as a monotone item response model (Cliff, 1992; Michell, 2003a). Algebraically this looks as follows:

$$P(x_{ij} = 1) \geq P(x_{hk} = 1) \equiv (a_i - d_j) \geq (a_h - d_k)$$

It can be seen how the Rasch model and its additive conjoint characteristics allow for two simultaneous attributes to be quantitatively concatenated, thus allowing for fundamental measurement. This notion of double cancellation is extremely important in this argument and is the reason why it was explained at length above. The following argument is Michell's (2004) but is considered so imperative in this discussion on dynamic assessment's status that it needs to be précised. Recall in the discussion on classical test theory above that error is unknowable and that true score derivation and interpretation is tentative at

---

[96] Ipsative is being referred to in its educational sense in which assessment is compared to the individual's past performance and does not refer to self-ratings by the individual.
[97] From a *g* point of view this is hardly surprising. The retort to this would be that *g* is the common factor.
[98] Is there really still a "psycho" in psychometrics? Consistent with Blinkhorn's (1997) concern about the tenuous link between test theory and psychological theory. Suen (1990) maintains that there is a consistent move towards cognitive theory-based psychometrics but that psychometric education programmes will need to consider more closely the issue of mathematical programming into subsequent courses.
[99] Ability is used throughout the discussion, but IRT is successfully utilised in personality assessment as well where "ability" is merely reflective of endorsing an item or not and in this context does not refer to ability as is understood within intelligence assessment (Hershberger, 1999).

best and egregiously wrong at worst. The one parameter Rasch item response model is a probabilistic model which is predicated on the probability of an individual endorsing an item dependent on the individual's purported ability level and item characteristics. The individual's ability as well as random error is taken into account. If error was not considered and the individual's response to an item was endorsed then it could be said that the ability equaled the difficulty parameter of the item. Thus one only knows an ordinal fact. However, item responses modelers assume quantitative information (and not merely ordinal - see the above discussion), from the distribution of the error component. If the model is true then the error distribution reflects the quantitative structure of the attribute but if the attribute is not quantitative (and so far, potential hasn't been proven as such) then the supposed distribution of error is completely vacuous. "Here, as elsewhere, psychometricians derive what they want most (i.e measures) from what they know least (i.e the shape of "error") by presuming to already know it" (Michell, 2004, p.126). Now, if random error is retained, but the shape is admittedly unknown then the best we can do is to assume ordinal structure of the attribute in question (potential) unless the double cancellations as discussed above manifest and hence, even utilising IRT models is suspect when it comes to assessing for psychological constructs. A partial solution to this has been put forward by Mokken (1971) (in Michell, 2004) in the utilization of a non-parametric item response model, which, according to Michell, has been largely ignored. The core differences inherent in IRT (new rules) and CTT (old rules) can be summarised as follows (Embretson, 1997b):

i. Old rules - CTT
   a. Rule 1 – standard error of measurement applies to all scores within a population
   b. Rule 2 – if reliability needs improvement, the test is usually made longer
   c. Rule 3 – in order to successfully compare tests utilising different forms, test parallelism and/or test equating is necessitated
   d. Rule 4 – score distributions are necessary if particular scores are compared
   e. Rule 5 – if the data yield normal raw score distributions then interval level interpretations can be made
ii. New rules - IRT
   a. Rule 1 - standard error of measurement generalises to the population but differs according to response patterns
   b. Rule 2 – as IRT is item-based, making the test longer is ineffectual. In fact it can be shown that shorter tests are more reliable
   c. Rule 3 – test equating is optimal for different ability groups
   d. Rule 4 – item based scores are paramount. The distance from items are required and not an entire score distribution
   e. Rule 5 – interval level scales are inherent in the model (conjoint measurement ensures a defensible and justifiable manner of quantifiable measurement)

Can dynamic assessment which is predicated on change-based measurement and the derivation of potential not avail of itself a manner of such measurement? In this way, at least adherence to robust modeling is accounted for, it remains within a scientific nomological network of scientific progress, pursues replicable research, and still maintains it's qualitative nature of clinical intervention assessment mode. This is one such attempt to place dynamic assessment within mainstream intelligence assessment whilst still participating fully (and perhaps even more so) in a scientific sub-discipline. What is being addressed here is firstly the need to determine whether psychological constructs are measurable and secondly to determine a method suitable for allowing such quantitative structure to become manifest which is precisely the dictates set forth by Michell (2004). His argument centres around the need to substantiate the requirement for additivity and to thoughtfully ask if such requirements are being met. His answer to this is that for much mainstream psychometric practice this issue has been ignored. It is not so much that quantification of attributes cannot be made as much as the fact that these questions are never asked in the first place.

Dynamic assessment can embrace conceptions of the nature laid down by Michell (2003, 2004) and can utilise the recommendations set forth. Such an avenue is available and taking cognisance is what is now necessitated. However, one must not fall into the same trap and declare that dynamic assessment constructs are themselves amenable to quantification. After all, adhering to a model which requires two simultaneous attributes and in so doing making manifest a third measure does not mean that the two original attributes are quantifiable. *All that is being done is arguing for the quantifiability of the model and not the traits*. In other words, quantification can never be detected directly but only via a means of monotonic transformations of the observed variables which fit the model (Borsboom & Mellenbergh, 2004; Cliff, 1992). This is highlighted because it is the crux of the argument. Is potential, like intelligence, measurable? The monotone Rasch model may well avail of its structure to make allowances for the additive conjoint structure of the attributes but what of the attributes? "At present, psychometrics does not possess theories of sufficient richness" (Michell, 2004, p.124). Items are usually constructed using face validity (it is an art rather than a science; Chipman, Nichols & Brenna, 1995; Embretson, 1983), of what we think measures something, of what we assume to be the case and what experience has presumably taught us. When items are dropped from item banks (item attrition hovers between 30% - 50%; Embretson, 2005) rarely is it the case that the actual reason for its being dropped is known or even questioned (Meier, 1994). If this is the case for the dropping of items, what can be said of the retention of others? What makes these items suitable? Luck, chance or statistical spread? It is unnerving to think that the situation presents itself in this fashion.

The unresolved question of validity makes its appearance in this particular segment of the argument. Utilising criterion measures as correlations of predictive success is hardly much of an answer to the question of what it is we measure (Barrett, 2005). How do we know for sure what it is the construct is? To unequivocally state that test A predicts test B is saying absolutely nothing about what underlies test A or B at all. Our tools may be robust, developed and highly advanced, but the point of origin which is the fundamental question of quantification is not being addressed, at least not adequately. The threefold problem looks as follows (Barrett, 2005):

i. We do not know the nature of the criterion against which we wish to test for predictive or construct validity
ii. We assume it is additive and hence measurable
iii. We causally link the new construct to an "established" criterion and do so within what is perceived to be a normative framework (there is no normative framework, recall our lack of progress in even defining what is meant by intelligence). A nomological network does not entail construct validity (Barrett, 2005; Borsboom, Mellenbergh & Van Heerden, 2004), because just by saying that we all agree $x = y$ does not make it so. Science does not work that way and should not work this way for psychology. One need only think about item bias and how items which are considered biased are removed from item banks for certain groups. Are the items ever looked at in terms of why they may have led to under/overprediction in the first place? This would entail going back to the nature of the construct. Most often these items are merely removed (Van de Vijver, 1991)

How do we adhere to a scientific discourse as discussed in chapter 3? Such deep philosophical issues are also found pertinent by Borsboom and Mellenbergh (2004). What of the discipline in this regard? Once again, the need is felt to state that assessment as sub-discipline should lodge itself firmly within one camp or the other, not both:

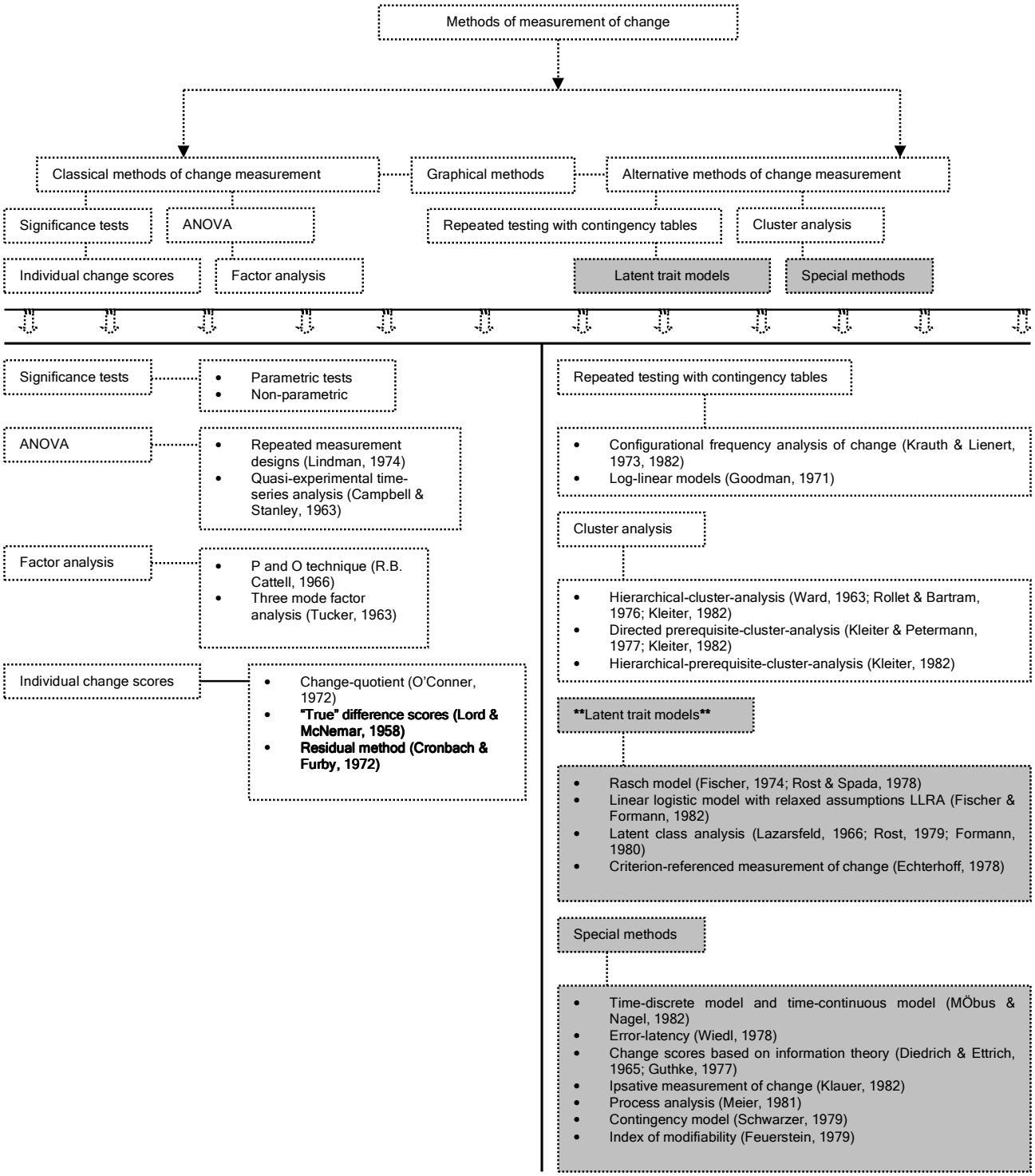
a. it is either a scientific field of enquiry molding closely to the natural sciences and is able to prove or at least spring from an axiom of measurability (recall the discussion on improvable axioms in the section on mathematics)
b. or it should retract its position within this realm and focus on its exclusive qualitative clinical interventionist strategy format to aid individuals in reaching what is now considered to be their potential but in a newer light.

Regarding axioms in (a) above, Borsboom and Mellenbergh (2004) say as much when they state that certain axioms will just have to be adhered to in order for the continued use of a probabilistic model such as certain IRT models. The accepted "fact" within the model is its assumption that what it measures is in fact quantitative. But is this good enough? After all this is precisely the concern that Michell (2003, 2004) has and it is not a concern to be shrugged at and swept under an assumption within a model. Does the model purport to *represent* (an isomorphic relation that is mapped from the attribute to the numerical?) or does it purport to *explain*? (Borsboom & Mellenbergh, 2004). This takes us back to chapter 3's discussion on the veracity of psychology as a formal science – does it seek to state or explain? This depends on its nature: instrumental, realist, constructivist and so forth. One cannot isolate such a thesis by engrossing oneself in the level of explanation attained only in this chapter. The need to look at a broad spectrum is warranted which is precisely what has been done thus far. It is at this point, where the issues discussed in chapters 2, 3 and 4 cohere. Right here in the debate on the nature of quantification, the nature of the tools used and the nature of what it is we psychologists are trying to do within science. The question asked here cannot and should not be asked only of assessment or dynamic assessment in particular but of science in general (Borsboom & Mellenbergh, 2004).

4.4.2.3.1 Modern test theory change models – an answer for dynamic assessment's change score problem

As prelude to this section on modern test theory's attempts to address change within a test, Kormann and Sporer (1983) offer a description of the state of affairs regarding the measurement of change. Even though theirs is a dated one the historical record is nevertheless accurate and hence warranted as is depicts the gradual process of development of ideas towards the understanding of change assessment originating from classical ideas to modern test theory ideas of change. Figure 68 below illustrates this.

Figure 68 Kormann and Sporer's (1983) overview of change measurement as viewed both classically and alternatively[100]



[100] Citations in this figure are not referenced in this thesis.

- *Embretson's multidimensional Rasch model for learning and change (MRMLC and MRMLC+)*

Reflecting on the gain score issue as discussed above in 4.4.2.2, the need for dynamic assessment to answer psychometric questions has been partially addressed since the late 1980's but has been known for many years prior (Guthke, 1982; Embretson & McCollam, 2004; Wilson, 1989) with modern test theory models allowing for changes in ability within a reliable model which attends to the issue of the change score paradox (Bereiter, 1962, 1967; Embretson, 1992; 2000; Embretson & Prenovost, 2000; Lidz, 1991). It is not a novel idea and perennially resurfaces when aided by the requisite technology but which will hopefully not fall in to the trap envisaged by Cole and Valsiner (2005) who summarily state that old and worn-out theoretical ideas resurface when in fact they should not. The most important aspect within these general models is their original concern with theory. Specific cognitive developmental theories (and more emphasis on cognitive processes assessment in general; Roberts, Markham, Matthews & Zeidner, 2005) posited by developmental psychologists serve as point of origin which is becoming increasingly focused on within modern psychometrics (Draney, Pirolli & Wilson, 1995; Embretson, 1994, 1997a) as opposed to the traditional (and still current) notion of measurement implying inference from observations to theory (Ippel, 1991). Theory is attended to via the tool and not the other way round (Wilson, 1989) as evidenced from early work on intelligence assessment which occurred largely in theoretical voids (Kaufman, 2004). Confirmatory IRT models assume that performance is determined by underlying abilities or traits which can be accounted for through the model (Kline, 1998), such as is the case with Embretson's multidimensional Rasch model for learning and change (MRMLC) (Embretson & McCollam, 2004) which is a special case of structured latent trait models (SLTM) (Embretson, 1997a, 2000). The development of confirmatory factor analysis has led IRT researchers to utilise confirmatory as opposed to exploratory factor analytic models. The former were typically the models utilised in early intelligence research (Keith, 1997) and is rightly stated to be an imperative tool towards the understanding of intelligence research in general (Schukze, 2005) precisely because of the link between theory and model and its respective philosophical underpinnings (Mulaik, 1988). Confirmatory models allow for hypothesis testing, whereas exploratory models' underlying nature and number of factors is unknown (Nunnally, 1978) relating back to sciences' origins as inductive and deductive knowledge-acquisition device.

Theory predominates over model and confirmatory models allow researchers to model the underlying theory (Heck, 1998). Due to the MRMLC'S constraining of item discrimination to unity, evidence of its familial tie to the Rasch model[101]can be seen (Embretson, 1991a, 1991b) as the Rasch model is well suited to incremental development within individuals (Bond, 2000). The need to impose such linear constraints is due to the need to allow hypothetical factors (latent trait, ability, intelligence or potential) to influence the observed responses (Fischer & Tanzer, 1994). However, similar constraining of modifiability averages is not permitted as their increases or decreases (Sijtsma, 1993b) over time are reflective of the latent response of potential (Embretson, 1991a). It is pertinent to recall the vast literature on ability change over tasks where correlations decrease between ability and task performance as skills develop (Ackerman, 2005). Its multidimensional status is implicitly implied within dynamic assessment as there is more than one ability accounted for (Embretson, 1987). Dimensionality of a test refers to the number of traits necessary to achieve local independence, however, local independence and dimensionality are not the same concept. The one follows on from the other. Validity within psychological constructs is based on the assumption that there exist x amount of latent traits which account for items being locally independent (Crocker & Algina, 1986). It is worth reiterating that no matter what model, no matter how sophisticated the mathematical modelling, no matter how grand the statistical techniques; construct validity remains with the realm of theoretical speculation. Rasch himself emphasised this in his book on probabilistic models where it is stated that only an *estimate* is given of θ, "in fact the best estimate obtainable but θ itself can never be determined empirically" (1980, p.77). Note his confidence in that the facility available to estimate θ is the best one available (and at the time is was) but that this recognition in no way impinges on empirical construct validity. This loops back to the discussion on psychology's methodology within science. Empirical testing of constructs might well be possible and indeed valid within a realm concerned with the physiological aspects of physiological measurement (see chapter 2). However, the case for dynamic assessment and its agenda is far from such empirical methodology and should maintain focus on qualitative endeavours, yet remain scientific in its own right. Rasch, it seems, was saying just this. Models aid, simplify and ease a number of issues, but even extensions of generalizability theory and various IRT models cannot yet account for valid constructs. Dynamic assessment's "potential" constructs would be better founded and validated within a more qualitative set-up. In essence what is necessitated by such a model of change is what Fischer (1997)[102] describes as an ability which is viewed "by a point in a multidimensional parameter space, and development being understood as a migration through that space" (p.189).

---

[101] The traditional unidimensional Rasch model did not allow for item subset comparisons "which are required to operationalise cognitive processing variables"(Embretson, 1997a, p.225). Hence the improvement along with the need to measure multiple dimensions (learning potential or change in ability).

[102] Fischer (1997, p.190) gives a compact outline of the various models he and his co-workers have devised. Dichotomous, polytomous and frequency data are modelled according to the number of occasions individuals are assessed as well as the nature of the items utilised (same or different items).

Unlike its predecessors such as Fischer's (1972) linear logistic latent trait model with relaxed assumptions (LLRA), Embretson's MRMLC is able to account for individual differences within training. The LLRA cannot measure individual differences and assumes all subjects learn at equal rates which is clearly not in keeping with dynamic assessment theory (in Embretson, 1991a). Likewise, Anderson's (1985) multidimensional Rasch model for repeated administration of the same items to the same individuals over a period of time was unable to account for change parameters for individuals although it was able to determine the impact of time or treatment on the theta distribution (in Embretson, 1991a). Initial ability along with modified abilities are assessed in which the initial ability and subsequent modified scores are related via the Wiener process[103] (the design structure for the MRMLC; Embretson, 2000) which increases over time and across conditions (variance will increase alongside decreases in correlations). The Wiener process allows for each ability level to be assessed seeing as it is being modified throughout (Embretson, 1996). As with the Rasch one parameter logistic, the MRMLC is defined by an initial ability level and the difficulty level of the item. However, in addition to the one parameter model, successive modified abilities are weighted between successive conditions and via the Wiener process ability levels are differentially involved. The generic structure Λ would like something as follows:

Table 18 Structure change dependent on modified ability through successive occasions

		Ability		
	Occasion	**θ_1**	**θ_2**	**θ_3**
Λ	1	1	0	0
	2	1	1	0
	3	1	1	1

If a typical dynamic assessment intervention worked perfectly, the structure in Table 18 would manifest, reminiscent of the perfect Guttman scalogram (Coombs, 1967; Kerlinger, 1981). Due to the model's similarity with the Rasch-family model, measurement is justifiably interval-level applicable. The initial ability θ_1 is utilised in all three occasions; θ_2 is the second ability which is θ_1 in addition to modification and θ_3 is θ_2 in addition to modification. θ_1 remains involved in each subsequent condition as does θ_2 and θ_3 in their subsequent inclusions (Embretson, 2000). The initial ability level (pretest score) is thus utilised throughout the successive modifications. The general structure for Embretson's (1991, 1992, 2000; Embretson & McCollam, 2004) MRMLC is as follows:

$$P(X_{i(k)j} = 1 \mid \theta_j, \beta_{I,} \lambda_{i(k)m}) = \frac{e^{(\Sigma m\, \lambda i(k)m\ \theta jm - \beta i)}}{1 + e^{(\Sigma m\, \lambda i(k)m\ \theta jm - \beta i)}}$$

[104]

θ_{j1} is the initial ability level; $\theta_{j2, ..., \theta jm}$ represent the modifiable abilities across different conditions; β_i is the difficulty for item *I* which is considered as constant across the various conditions; $\lambda_{i(k)m}$ is the weight of the latent ability *m* in item *i* within condition *k* which is specified as either 0 or 1. The structure Λ represents "matrix weights for the items within the *k* conditions in estimating the *m* abilities" (Embretson, 2000, p.513) as seen above in table 18. Dynamic assessment theory pivots the central notion of change of ability and here the assumption is being made that potential can be thought of as an ability. This is an issue some mainstream psychometrists contest (Embretson, 1987, 1995), hence the term multidimensional in the model, it is assessing more than one dimension; namely potential (Embretson, 1995). Originally, unidimensional models were considered robust enough to manage multidimensional traits especially when traits were highly correlated[105] but this is not always the case especially in change-models (Adams, Wilson & Wang, 1997). Ability and potential represents two variables, pretest scores are unidimensional and posttest scores are multidimensional (due to the introduction of modifiability) (Smith, 1997). The most common reason for the use of change scores is to operationalise the concept of such an ability (Cronbach & Furby, 1970; Pennings & Verhelst, 1993). It is essential to do so as dynamic assessment construes learning potential as an ability (Hamers & Sijtsma, 1995). The question now raised would be: if intelligence is not definable as a construct, is learning potential? Moreover, if learning potential is linked to intelligence in some form or another, the questions asked of intelligence research can be directed at dynamic assessment. If

[103] Embretson (2002) notes that although the Wiener structure is used in the MRMLC it is not the only system that can be used. Others include the Helmert structure which may posses better measurement properties because of the greater number of items in the estimation of ability and modifiability but these modifiabilities are more global as opposed to condition-specific. "The Wiener simplex model is particularly appropriate for structuring parameters in a Rasch model since the properties of equivalency of measurement scales between tests and additivity of the effects are compatible between the models" (Embretson, 1991, p.499). Another design embeds an orthogonal polynomial design where abilities are represented as global, linear, cubic or quadratic change. Data fit for each system is not a concern but the choice of system is determined by measurement error which varies across the systems and secondly, theory should dictate which system should be used.

[104] For the sake of consistency, notation representing ability (θ) and item difficulty (β) will be kept consistent throughout even though authors may choose to utilise other notational forms.

[105] For intelligence assessment is this yet again reminiscent of *g*?

dynamic assessment ceases to conform to mainstream assessment it might not have to even answer the question at all. Embretson's (1992) mathematical modelling of these manifest changes from pre to posttest has evidenced the change in the construct representation of the ability which results in changes of test validity as well. Embretson's (1991a, 1992, 1995, 2000) model:

- Predicts pre-posttest reliability by factoring in modifiability within the multidimensional scale through additional dimensions in the posttest results and does so simultaneously using maximum likelihood[106] for all item responses over time. Via maximum likelihood estimation, error variance for abilities is known. As a repeated measures design, means and standard deviations increase across occasions
- Accommodates varying meanings attributable to changes in raw scores at different initial ability levels. In other words ability estimates are invariant across the sets of items and these ability estimates are not biased due to the difficulty level as they would prove to be in CTT and because the scale is no longer an ordinal one, interval-level changes are meaningful. Irrespective of initial ability levels, modification has the same impact on log odds on any of the items which is a direct consequence of utilising the Rasch model. Initial ability and modified abilities are now additive (see section 4.4.1.2 above on the importance of additivity for quantification of constructs)
- As per dynamic assessment theory and because of its IRT model status, there is an expectation of different raw scores for persons who manifest equal levels of potential or modifiability when their initial ability levels differ. In addition to this, different abilities can be estimated from the same items because it is the response pattern that is assessed (Wright & Stone, 1979). This allows for change to be viewed as a separate ability
- Removes scaling artefacts that contribute negative spurious elements within the correlation change scores and are thus partialled out. Problems associated with measuring change are essentially due to scaling and regression (Klauer, 1993; Schöttke, Bartram & Wiedl, 1993)
- Performs the functions inherent within unidimensional IRT models as performance is ascertained from response patterns and is not based on linearly derived total scores. Due to the original Rasch unidimensional model's lack of flexibility (Wang, Wilson & Adams, 1997) the MRMLC provides more robust initial ability estimates as well as introducing modifiability estimates which are corrected for differences in initial ability
- Allows for confidence intervals to be estimated for each person's ability and modifiability estimations and these confidence intervals can be compared across design structures (recall the plea in section 4.3.1.2 where confidence intervals were increasingly called upon to perform the function of power in NHST)
- Yields different results from CTT when viewing change scores, "statistical results depend on how change is measured" (Embretson, 2000, p.518)
- Does not repeat items. It does so to avoid the classic problems associated with repeated measures which include practice effects, retention of items and response consistency effects. Local independence, one of the main characteristics of IRT is thus upheld

The model would:

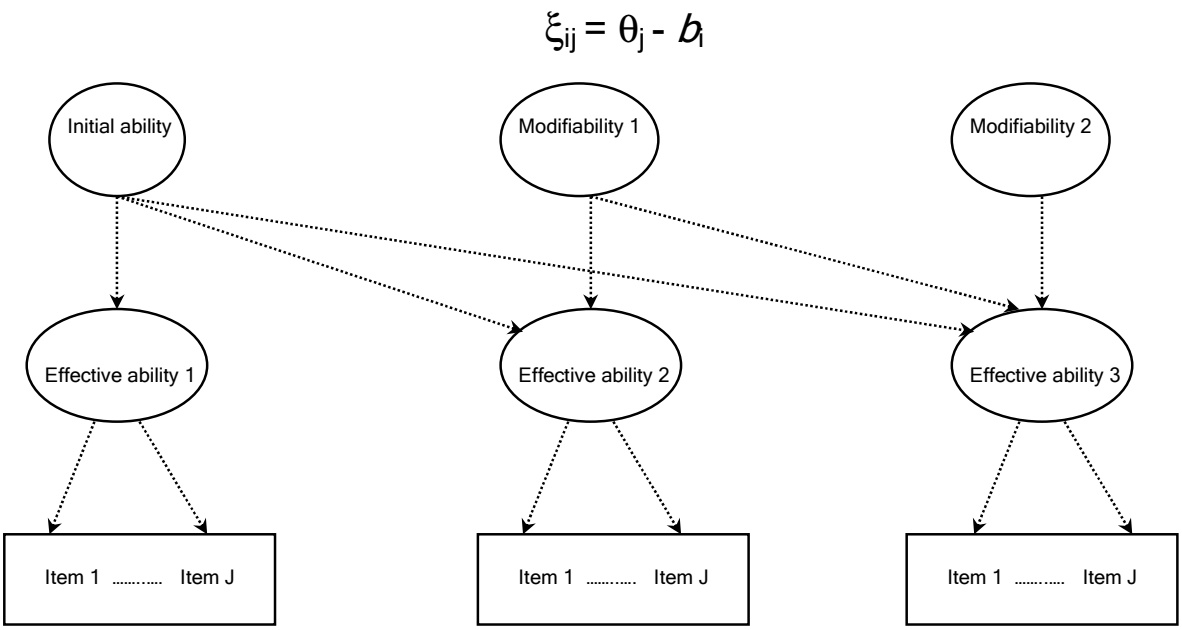
- be better utilised within situations where modifiability was the main target of change measurement
- not necessarily yield greater predictive validity with fixed content tests, as pretest and posttest scores would suffice
- need multiple groups so as to observe each item within each condition
- need to condition on the ability at each occasion within each group in order to estimate item parameters
- be limited in its efficacy to link ability changes or learning to requisite changes in the cognitive processes underlying such changes. In other words, the theory-substantive fit could be better defined

Embretson's (1995) MRMLC+ extends functionality of the MRMLC discussed above by resolving issues surrounding individual change and allows for adaptive testing so as to estimate learning which it links to the substantive changes in processing and knowledge. In other words it links the difficulty of the task to features that influence cognitive processing (Embretson, 1996). The improved version of MRMLC includes structural models of item difficulty and includes the facility to link changes in the model to changes within the construct (Embretson, 1995). MRMLC originally estimated the difficulty parameters for each item whereas MRMLC+ contains models for both person and item. Difficulty item parameters are replaced with new model predictions. Figure 69 illustrates the relation between initial ability level, modified abilities and the subsequent effective ability which results from the accommodation of initial and modified abilities in the model (Embretson, 1991b). Cliff (1991a) refers to this process as a path

[106] Recall Bayes' theorem. The maximum likelihood approach is a special case of Bayes and is used under certain circumstances (Suen, 1990). Three methods of maximum likelihood are utilised for Rasch models: joint maximum likelihood (point estimates for persons and items), conditional maximum likelihood (item estimates only) and marginal maximum likelihood (item parameter estimates and sample distribution parameters) (Roberts & Adams, 1997). Bayesian parameter estimation is also another model which uses a uniform prior function but it is recommended that the distributions for θ and b be normally distributed (Suen, 1990). Once again it is evident how important Bayes theorem is to modern day test theory; a theorem, which in Bayes' own time, was all but forgotten. Maximum likelihood is not the only mechanism through which estimations of parameters can be made. Heuristic or approximate procedures can also be utilised (Crocker & Algina, 1986).

model due to each stage in the process being reliant on the values of the previous stage. The current score reflects the previous score in addition to an increase that presumably arises from mediation (or some treatment or in the case of lifespan development, maturation) that has occurred in the interval since the previous measurement. It is worth asking the question at this stage: what happens if mediatory intervention makes no difference or results in a decrement? This echoes Collins' (1991) question as to whether development is reversible and how longitudinal change-based models are able to effectively handle such data.

Figure 69 Effective ability as consequence of initial and modified ability within the Rasch model (Embretson, 1991b, p.189)



The South African study by De Beer (2000) which, as mentioned above, utilised IRT as means of constructing and validating a dynamic assessment test recognised the importance of Embretson's model but differed in its emphasis on learning potential as exclusively defined by the difference score. De Beer's (2000) study utilised the pretest and difference score in defining learning potential and not just the change score. De Beer's (2000) study utilised computer adaptive testing (CAT) as platform for IRT response modelling for a dynamic assessment and as Embretson (2005) states, the future of CAT has already proven expedient (Drasgow & Olson-Buchanan, 1999; Kingsbury & Houser, 1999). The next phases in this development are underway, namely, adaptive item generation which generates new items which avail of the most information pertaining to the individual. Not only does the computational system, which is premised on artificial intelligence, choose the next appropriate item but it does so by merging psychometric theory with substantive cognitive theory. It is paramount that the stimulus value of the item in terms of its cognitive consequences on item functioning be understood. Embretson's (2005) item generation system is based on "the cognitive design system approach" published in 1998 (in Embretson, 2005). Cognitive design emphasises the impact of the underlying processes involved in solving items, the impact that this process has on performance as well as considering the impact of the stimulus features on the process (Embretson, 2005). This approach takes information from both worlds so to speak:

- ⇒ the psychometric world gives to the design information on item difficulty and item discrimination
- ⇒ the cognitive world gives to the design, information processing difficulty which is derived from an information-processing approach to the study of cognitive. This stresses problem solving, use of abstract reasoning and working memory capacity for instance
- ⇒ items which embody the same sources and levels of cognitive complexity emanate from the same structure type but these structures can differ qualitatively from item type to item type. Research is ongoing in an effort to bridge the development of these structures to variables within the cognitive model

Of note to this study in particular, is that this approach has been extensively applied to advanced matrix tests including the Raven Advanced Progressive Matrix Test, a test often utilised within dynamic assessment studies and also considered a good indication of intelligence (of both *g* and *Gf*, Wilhelm, 2005) within mainstream intelligence assessment. The elements thus far:

- ⇒ development and deployment of substantive cognitive theory (see chapter 3 and its discussion on substantive theory as well as chapter 4 which argues for substantive theory development alongside its

statistical entity status within statistics). Embretson (2005) supports Cronbach and Meehl's (1955) treatise on construct validity and how it is defined within a nomological framework. This is, however, contrary to the views espoused by Borsboom, Mellenbergh and Van Heerden (2004) for instance. The latter authors prefer to plumb epistemological and ontological depths when it comes to this issue, whereas Embretson shows a preference for immediate solutions in a practical world. Both sides have much meaningful insight

⇒ development and deployment of robust psychometric theory (further developments in measurement foundations)

⇒ the utilisation of both CTT and IRT indicators of validity and reliability (in so far as this can possibly be attained)[107]

⇒ further development and deployment of multidimensional IRT models which are able to account for changes in underlying ability (the perennial change-based approaches which are increasingly becoming psychometrically justifiable)

⇒ the advancement in CAT which utilises information and practice from various disciplines such as artificial intelligence, programming languages and neurology. This echoes the stance taken in chapter 2 in which reductionism as approach was applauded in its efforts to understand various sub-systems to a greater extent. Once each individual system is more fully understood, the closer the goal of unification appears in terms of bringing much of this knowledge together in a framework which will serve to elucidate the underlying processes within intelligence assessment. The discipline of dynamic assessment especially should take heed of these novel approaches

- *Wilson's SALTUS model*

Wilson (1989) developed his developmental Saltus model (from the Latin "leap" and by implication "leaping" from one stage to the next) which is also an extension of the Rasch model (Embretson & McCollam, 2004). Utilising various tasks at each developmental level, state changes are measurable but Wilson (1989) assumes that development proceeds in a stage-like manner and that it is not continuous (Wilson & Draney, 1997). This is contrary to available evidence which suggests cognitive development is continuous or at the very least undecided with much research involving mathematical theories of complexity (complex systems theories), chaos and catastrophe[108] as well as sophisticated computational models (Brainerd, 1993; Feldman & Fowler, 1997; Fischer & Rose, 1998; Gelman, 2000; Halford & Wilson, 1980; Johnson & Gilmore, 1996; Mareschal & Shultz, 1997; McCollum, 1997; Molenaar & Raijmakers, 2000; Preece & Read, 1995; Schulze, 2005; Shayer, 2000; Suizzo, 2000; Van der Maas & Molenaar, 1992). Modern theory and practice have called into question some of Piaget's results[109] (Gamlin, Elmpak, Franke-Wagar, Lotfabadi, Nourani & Wang, 1996; Nicolopoulou, 1993). That critical periods of development exist is not in question (Carlson, 2002). Non-linear models of development including catastrophe theory model the development process as one of continuous, stable yet abrupt change and is characterised by attractors (stable aspects of development) which can change radically and unpredictably (Sternberg & Grigorenko, 2002). It is vital to understand the cognitive theory behind the use of various models as they clearly effect interpretation and understanding of the underlying rationale. Was the model built to mould the theory or has theory been tweaked to mould to the model? The assumption here is that if researchers or practitioners do not agree with the stage-like development as advocated by Piaget for instance, then perhaps this model is not the one to pursue within a dynamic assessment set-up positing change. How exactly does the change occur? This is one example of a question that needs to be addressed before models are used. These changes are defined on two types of discontinuities; first and second-order changes. First-order changes are sudden changes which occur within a single ability and second-order changes are those manifesting across at least two domains. The Rasch one parameter model is utilised for the first-order changes and Saltus model is utilised for second-order changes. These first and second-order changes are the result of Fischer, Pipp and Bullock's (1984) (in Wilson, 1989) work on discontinuities within cognitive development; an instance, as stated above, where model was derived from theory. Wilson's (1989) Saltus model is a refinement of Guttman's scalogram model (as can be seen in Table 69 above) in which one item per level exists (Embretson & McCollam, 2004).

---

[107] Recall the thesis running throughout: no amount of technical application, development and sophistication can alter the one remaining variable. That variable is the origin of test items which are devised by people for people by hand on an intuitive basis and there is nothing scientific about this in the strictest sense of the practice of science. As argued throughout this thesis though, a science can be practiced within psychology without necessarily having to adhere to natural science tenets.

[108] Which does tend to make one sceptical of over simplistic modelling of change!

[109] It is at times unfair to be overly judgmental towards researchers in years gone by especially when their original contributions were so ground-breaking at the time. Piaget was situated with a context of his own and should be judged accordingly and perhaps not so ruthlessly discredited (Campbell, 2000; Kuhn, 2000; Perre-Clermont, 2000). Piaget's constructivist account of development is relevant in approach even today (Karmiloff-Smith, 1994) although much of his theory has been criticised (Anderson, 1992; Bates & Elman, 1994). His framework can serve as point of departure for comparative behavioural analysis regarding the development and trajectory of primate cognitive development (Doré, 1991). This highlights the similar evolutionary background shared among primates but this is a discussion for another time and place. But as an aside, it is remarkable to see the differences in animals who have led lives enclosed in poor environments and who have received subsequent exposure to better environs; is this not a manner of allowing the animal to negotiate in a zone of yet-to-be development?

Due to problems within Guttman scaling, Wilson (1989) utilised Rasch's probabilistic approach thus circumventing the adherence problem which manifests when persons not adhering to the scalogram ordinal model are discarded. In addition to the usual parameters (person ability and item difficulty) Saltus includes parameters influenced by cognitive theory which account for stage influences on item types. Wilson (1989) investigated data sets according to three cognitive developmental models in an attempt to apply his model and determine the fit with each set according to each theory. Wilson's (1989) reiterative concern echoes Guttman's similar concern; scale analysis does not define content, only substantive theory can! Recall this study's preoccupation with the need to re-look philosophical issues within psychological concerns. Refining and tuning tools of measurement will never solve a substantive problem, no matter how sophisticated the instrumentation is. Utilising modern IRT multidimensional change models can only aid in better fit to data[110] and theory provided the theory is, via quantification, amenable to such measurement in the first place. Wilson's (1989) structure for Saltus is governed by a logistic function as in the Rasch one parameter model but the ability / item parameter (θ_i - β_j) consists of additive elements for person, ability, item difficulty and level and is expressed as such:

$$\theta_i - \beta_j + \Sigma_h \phi_{ih} \tau_{hk},$$

$$i = 1,\ldots,N,\ j = 1, \ldots,L.$$

k = item type is a known function of j with at most k types; τ_{hk} is the level parameter for the subject group h and item type k. ϕ_{ih} is the selection factor which indicates group membership with $\phi_{ih} = 1$ if person i is in group h and 0 if not. Depending on the number of person groups, h, $\phi_i = (\phi_{i1},\ldots, \phi_{ih})$ must be estimated from the data. The person group $h(i) = h$ and item type $k(j) = k$, the probability of response y_{ij} is as follows:

where $\psi(\sigma_{ij})$ is the logistic function

$$p(y_{ij} \mid \theta_i, \beta_j, \tau_{hk}, \phi_{ih}) = [\psi(\sigma_{ij})]^{y_{ij}} [1-\psi(\sigma_{ij})]^{1-y_{ij}}$$

$$\psi(\sigma_{ij}) = \frac{E^{(\sigma ij)}}{1 + E^{(\sigma ij)}}$$

$$= \frac{E^{(\theta i + \beta j + \Sigma h\, \phi ih\, \tau hk)}}{1 + E^{(\theta i + \beta j + \Sigma h\, \phi ih\, \tau hk)}}$$

Each person from each group, it is assumed, will apply strategies pertinent to that group across all items and under the assumption of local independence the conditional probability of a response patter $y_i = (y_{i1},\ldots,y_{il})$ is as follows: (see section on Bayes conditional probability above) Wilson (1989) emphasises again the importance of substantive theory because the function $h(j)$ is known *a priori*. Group membership is determined from the data based on observable responses but the ability that led to that observed function was not known):

$$p(y_i \mid \theta_i, \beta, \tau, \phi_l) = \Pi_j [\psi(\sigma_{ij})]^{y_{ij}} [1-\psi(\sigma_{ij})]^{1-y_{ij}}$$

- *Other models*

  i. Multidimensional random coefficients multinomial logit model

Other IRT models which encompass change in their structure have been developed, and include among others, the multidimensional random coefficients multinomial logit model which contains Embretson's MRMLC structure in addition to many other models developed since Rasch's 1960 model (Adams, Wilson & Wang, 1997). This model is an extension of the original unidimensional model developed by Adams and Wilson (1996) (in Adams et al., 1997). The unidimensional random coefficients multinomial logit model (RCMLM) as precursor to the multidimensional random coefficients multinomial logit model (MRCMLM), allows for flexible custom-designed testing situations and extends other generalised Rasch models (Wilson & Wang, 1995) which impose linear structures on item parameters, similar to Embretson's MRMLC's item constraint. It utilises marginal

[110] The author is aware of the reigning controversy surrounding the debate of fitting the data to the model (Rasch) and fitting the model to the data (one, two and three parameter IRT) (Andrich, 2004).

maximum likelihood estimates for parameter estimation (Wilson & Wang, 1995) which aids in dealing with inconsistent estimates provided that the population distribution is specified. If not, conditional maximum likelihood estimation is utilised (Adams, Wilson & Wang, 1997). Joint maximum likelihood estimates become inconsistent as the sample sizes increase thus the utilisation of marginal likelihood estimates (Wilson & Wang, 1995). It allows for different numbers of categories in different items, differences in rater estimates and avails of multilevel structures (Wilson & Wang, 1995). Given the unique nature of dynamic assessment and the underlying ability changes, it is hypothesised that the Adams et al., 1997 model can perhaps be utilised for certain dynamic assessment situations. Specifically, the immediate attraction here is that the RCMLM incorporates a scoring function which enables a flexible relationship between qualitative aspects of the performance on an item and the quantitative level of performance the response reflects; i.e. scores above 0 and 1 can be allocated. The following exposition is taken from Adams, Wilson and Wang (1997). The RCMLM model can be written as follows:

$$P(X_{jk} = 1; A, b, \xi \mid \theta) = \frac{e^{(b_{jk}\theta + a'_{jk}\xi)}}{\sum_{k=1}^{kj} e^{(b_{jk}\theta + a'_{jk}\xi)}}$$

And the response vector model as follows:

$$P(X = x \mid \theta) = \Psi(\theta,\xi)\, e^{[x'(b\theta + A\xi)]}$$

With

$$\Psi(\theta,\xi) = \left\{ \sum_{z \in \Omega} e^{[z'(b\theta + A\xi)]} \right\}^{-1}$$

Each item, i, consists of k_j+1 response alternatives (k=0,1, …, k_j). Xj is the vector valued random variable where X_{jk} is 1 if the response to item *i* is category *k* and if not, then 0. X' is the response pattern realised via the lower case x and x'. ξ are item responses of *p* parameters which as mentioned above are linearly constrained. The linear combinations result in design vectors a_{jk} (i = 1, …, n and k = 1, …, k_j) each with length *p*.These are collected into a design matrix A. The scoring function mentioned above, is represented by a response score b_{jk} and is collected as a vector b'. In the simple logistic model 0 and 1 are utilised as labels to indicate both the response category and the scores; which also includes figures above 1 in developed models. θ is of course the latent variable.

The extension of this RCMLM model is the MRCMLM[111] which takes into consideration a number of traits, D, as underlying performance, hence its multidimensional status. The D-dimensional latent space is once again represented by a vector $\theta = (\theta_1, \theta_2, \ldots \theta_D)$ (Wang & Wilson, 2005) and represents a random sample from a population. The scoring function is now a vector and not a scalar as in the RCMLM because the response category, *k*, in item *i*, corresponds to a D X 1 column. A response in category *k*, on dimension d (d=1, …, D) of item *i*, is now represented as b_{jkd}. These scores, in turn, are collected into a vector b_{jk} and are collected into an item scoring submatrix, B_i and lastly collected into a scoring matrix B. ξ and A are defined as they were for the RCMLM model where they define the RCMLM for the item (Wilson & Wang, 1995). The probability of a response in category *k* of item *i*, is as follows for the MRCMLM:

$$P(X_{jk} = 1; A, b, \xi \mid \theta) = \frac{e^{(b_{jk}\theta + a'_{jk}\xi)}}{\sum_{k=1}^{kj} e^{(b_{jk}\theta + a'_{jk}\xi)}}$$

The difference between the RCMLM model above and the one to the left (the MRCMLM), is that θ is a scalar in the former but a DX1 column vector in the latter. Also, in the former model, the response *k* to item *i* is a scalar b_{ik} and in the latter model this is a vector $\mathbf{b}_{ik}$. D = 1 + exp $(\theta_1 + \xi_1)$ + exp $(\theta_1 + \theta_2 + \xi_1 + \xi_2)$ + $(\theta_1 + \theta_2 + \theta_3 + \xi_1 + \xi_2 + \xi_3)$.

And the response vector model as follows

$$f(x; \xi \mid \theta) = \Psi(\theta,\xi)\, e^{[x'(B\theta + A\xi)]}$$

With

$$\Psi(\theta,\xi) = \left\{ \sum_{z \in \Omega} e^{[z'(B\theta + A\xi)]} \right\}^{-1}$$

[111] Wang and Wilson (2005) state that MRCMLM runs in ConQuest and is implemented with the marginal maximum likelihood estimation. The SAS NLMIXED procedure can also be used for fitting linear and non-linear models of the MRCMLM (Wang & Wilson, 2005). In addition it also utilises an empirical Bayes method (i.e. multivariate normal distribution is assumed and the vector-covariance matrix is empirically estimated (see the discussion on Bayesian statistics as under-utilised statistic within the social sciences above).

ii. Unified model for assessment

The unified model for assessment, although not developed with the change construct in mind is, according to Embretson (2000), relevant to the identification of the underlying attributes which have been successfully mastered (DiBello, Stout & Roussos, 1995). The unified model captures substantive cognitive theory as well as psychometric item response theory in one model. By utilising substantive theory the model is able to relate information pertinent to practical contexts in which specific information pertaining to cognitive areas is needed. Incorporation of educational assessment via cognitive theories into a model which is able to amalgamate both psychological research and robust defensible psychometrics is starting to come of age in models such as proposed by DeBello et al., (1995). Regarding conventional latent trait models:

- ⇒ Very few continuous latent traits are evidenced as underlying responses
- ⇒ A model is said to be multidimensional only when item responses can be shown to be conditionally independent when equal ability individuals are sampled
- ⇒ This multidimensional trait then accounts for the covariation between items
- ⇒ Such an approach, although successful with broad-based ability traits, are not able to render anything useful in terms of cognitive specificities which is needed in practice during remediation and so forth

Newer models:

- ⇒ Accommodate discrete cognitive skills leading to latent class analysis
- ⇒ When the probability distribution of classes (initially not known) is known as well as conditional item response probabilities the latent structure will be known
- ⇒ The research in this area is still being conducted and as a result, if too many classes are utilised, the model becomes unfeasible. Likewise, too few classes results in the resemblance of this model to the multidimensional models
- ⇒ The model can account for where the individual is placed within a cognitive space and can determine the accuracy of the assessment – once again it can be seen how substantive theory and psychometric modelling can be melded
- ⇒ A fine balancing act is necessary if researchers are to extract the most pertinent cognitive information from test responses in such a manner as to allow for a greater number of parameters, without the size increasing to such an extent that the model becomes unwieldy. In order to aid in making the model finer grained, certain computational mechanisms are needed, such as neural networks and Bayesian inference networks
- ⇒ The trade-off between fewer parameters versus less accuracy is the model's proficiency at detecting small changes in responses
- ⇒ The model incorporates various sources of stochastic response variation (as yet unaccounted for in uni- and multidimensional trait models, factor analytical models) such as:
    - Differential strategy usage – i.e. the model may predict certain strategies and not others, but if the item is endorsed utilising unaccounted for strategies this needs to be factored in
    - Completeness – which reflects the as yet to be completed list of skills, procedures or other attributes which are used by individuals but are not listed in the model
    - Positivity – where for some reason the correct strategy is not used even though the ability is present, or the requisite ability is not present but the correct strategy was used
    - Slips – random errors made by individuals for any number of reasons
- ⇒ The model is discussed in detail in DiBello et al., (1995) from which the above has been summarised. The research presented by these authors, although now over ten years old, is indicative of where future latent trait models are heading and is currently cited by eminent researchers in the field (Embretson, 2005). At the time of the 1995 publication, though, there were still various issues which needed attention in the model such as model calibration and the need to incorporate more cognitive operation necessary for diagnostic assessment. The research was still at simulation phase

iii. Linear Partial Credit Model

Ponocny and Ponocny-Seliger (1997) have developed a programme which manages various models within the area of change; $L_{PC}M$ (Linear Partial Credit Model). It fits models for two or more time points and can do so for unidimensional, multidimensional and mixed dimensional items sets with both dichotomous and polytomous data in an attempt to measure change. The programme is written in C and initially ran in DOS, but was, during 1997, being developed for a Windows based environment.

Dénouement

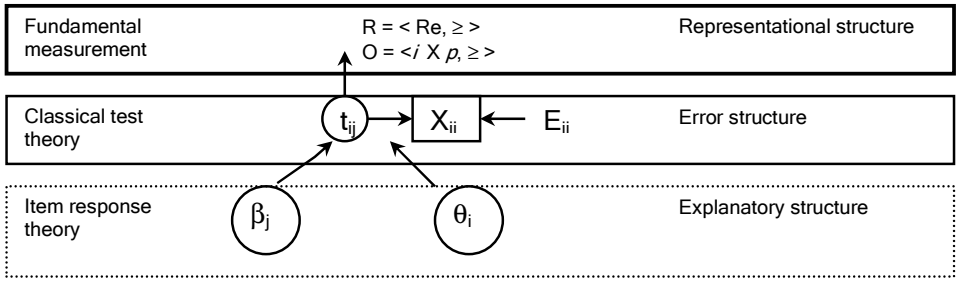
The culmination of mathematics, statistics and measurement and what these seemingly disparate areas of concern mean for test theory cannot be more clearly seen than in Wright's (1999) table which he has entitled "An anatomy of inference" which

follows below in table 19. Test theory is an evolving discipline with a large literature attesting to various models' strengths. It is worth citing directly from Borsboom (2005) regarding the three main theories concerned with the measurement of the mental: "classical test theory is basically about the test scores themselves, representationalism is about the conditions that should hold among test and person characterises in order to admit a representation in the number system, and latent variable theory is about the question [as to] where the test scores come from" (p.121). This statement can be visually reflected in the following figure, 70.

Table 19 An anatomy of inference (from Wright, 1997, 1999)

An anatomy of inference		
Obstacles of *raw data* inference	**Solutions to *inference* from raw data**	**Inventors**
Uncertainty (do not know)	*Probability (know at least something)*	Bernoulli, 1713 (section 4.2)
have => want	binomial odds	Bayes, 1764 (section 4.3.1.1)
now => later	regular irregularity	Laplace, 1774 (section 4.2)
statistic => parameter	misfit detection	Poisson, 1837 (section 4.3.1)
Distortion (from observation to conceptualisation)	*Additivity (making visual sense of observation data by making it linear in our two and three dimensional space)*	Fechner, 1860 (section 4.4.1.1)
nonlinearity	linearity	Helmholtz, 1887 (section 4.4.1.1)
unequal intervals	concatenation	Campbell, 1920 (section 4.4.1.1)
incommensurability	conjoint additivity	Luce / Tukey, 1964 (section 4.4.1.1)
Confusion (the need to separate and study one dimension at a time)	*Separability (causes as separate parameters make for our quantification; i.e. item or person parameter)*	Rasch, 1958 (section 4.4.1.3 in which Rasch (1980) is used)
interdependence	sufficiency	Fisher, 1920 (section 4.3)
interaction	invariance	Thurstone, 1926 (sections 2.7.1.1.; 4.4.1.3 and 2.8.2)
confounding	conjoint order	Guttman, 1944 (section 4.4.2)
Ambiguity (based on our still arbitrary decisions as to what underlies behaviour)	*Divisibility (there are known solutions to dealing with such ambiguity)*	Levy, 1937 (section 4.4.1.3)
of entity, interval	independence	Kolmogorov, 1950 (section 4.2)
and aggregation	stability	Bookstein, 1992

Figure 70 Relations between fundamental measurement, CTT and IRT(Borsboom, 2005, p.129)



Fundamental (representational) measurement:

- ⇒ the true score can be instrumentally compared ≥ to the product of person ability and item difficulty (*i* X *p*)
- ⇒ the true score is numerical and can so be ordered
- ⇒ the effects of latent ability and item difficulty are independent, items and subjects can be ordered instrumentally
- ⇒ this results in O = <*i* X *p*, ≥ >
- ⇒ if the model is true and because ability and item difficulty are independent, true scores can be transformed in an additive manner befitting Rasch model representation and
- ⇒ this results in the numerical relational system R = < Re, ≥ > but "the representationalist theory becomes a liability, for it prises apart the numerical concepts used in science from the empirical reality studied" (Michell, 1990, p.49)
- ⇒ together, O = <*i* X *p*, ≥ > and R = < Re, ≥ > form an additive conjoint measurement structure

Classical test theory:

- ⇒ is a theory of error structure
- ⇒ the true score is defined as a propensity distribution of expected values for individuals, *i*, and items *j*

Item response theory:

- ⇒ hypothesises the data generating process
- ⇒ variation on the latent trait produces variation on the true score
- ⇒ item difficulty produces such variation as well
- ⇒ it thus claims more than the fundamental representation, which itself cannot add anything other than a representation of the relationship
- ⇒ as mentioned, the IRT model is falsifiable on account of its not fitting the data. Fundamental measurement cannot avail itself of this even as it only functions as a construction of a representation and nothing more

4.4.3 Summary

"To measure" - this short clause is packed with underlying assumptions which are sometimes made explicit but are most often implicitly implied. This tacit agreement among psychologists as to what constitutes measurement, borders at times, on ignorance of these said assumptions. Measurement precludes quantifiable structures which, as a concept, necessitates an additive structure, units of concatenation, provable representation, fundamental qualities or extensive units or in the case of less identifiable areas of concern, conjoint measurement based on two or more intensive measures via the process of double cancellation as well as via solvability and Archimedean axioms. Do psychologists routinely consider the above? This question is often asked within texts detailing the shortcomings of a quantifiable psychology and social science. Measurement within psychology is intimately bound to its mathematical foundation, for it is upon a system of axioms which has allowed for its maturation into the discipline it is today. Also intimately bound to this measurement system are the statistical techniques employed to render inferential information to users. It is, after all, the users who are supposedly gaining from such a system of measurement. This thesis though is highly questionable as it cannot at this stage even be proven that the constructs being measured exist as defined according to the techniques used to measure them. Unfortunately it seems that constructs are defined by the measures used which is tautologous, unfair and is simply an egregious error in logical thinking.

Psychological measurement within the intelligence arena is often predicated upon a general construct seemingly prevalent in most sub-tests assessing for some sort of intelligence. There seem to be as many definitions of intelligence as there are tests yet the one pervasive finding is the statistical manifestation of a general factor. This is hardly surprising if most test batteries

assess on very similar items. Spearman's general factor and his development of a theory of mental testing attests to his insight and work but one must be mindful that there are possibly innumerable ways of assessing for constructs which do not necessarily fit the picture of a general factor underlying what we think is intelligence. Much work in this regard has been undertaken and has been gaining momentum since the early years of the twentieth century often concurrently to classical notions of test theories. The main argument within dynamic assessment, as this thesis contends is that

- Intelligence is difficult if not impossible to define
- The construct is assumed to exist in some fashion or other
- It is likewise assumed to be stable

Dynamic assessment not only has to contend with the above-mentioned three aspects but has the following added to its problems

- learning potential is difficult to define especially if one is to continuously define it according to mainstream ways of identifying it
- learning potential will need to extricate itself from the mainstream terminology otherwise it will never define for itself its own position – one can only proceed to step two once step one is complete; but step one is not even there
- learning potential assess for change which is entirely at odds with the notion of stable development and progress

The solutions to this dilemma are numerous but the following can perhaps be put forward as tentative attempts

- dynamic assessment should leave the realm of intelligence assessment as currently understood in mainstream testing
- intelligence assessment as understood in the mainstream realm is perfectly suited to further advancements on the physiological front as well as on the psychometric front
- the research efforts as evidenced by numerous erudite intelligence researchers cannot be ignored and in fact should be encouraged, however,
- dynamic assessment has its own unique philosophy and thus should contain itself within a similarly different philosophical and practical areas and hopefully in this way
- it will be able to more accurately define for itself what is meant by learning potential. The definition may well undergo radical transformation if defined within a completely different context

Measurement utilises elements as its population of manipulation. Such elements need to undergo transformations if they are to be accurately represented within a system of measurement. In order to effect this, correspondence rules need to be addressed so as to preserve the original structure of the element. Such transformations are governed by various techniques applicable to certain types of elements which is why due consideration for psychological measures is so imperative. Psychological measurement was thought about a great deal during the first four decades of the twentieth century and it is perhaps a pity that similar thought had not occurred on such a scale since for it has become routine to accept what is thought to be an accurate understanding of what it means to measure. Stevens, who is often viewed as the person starting a measurement revolution for psychology in the 1940's gave to psychological measurement, scales, which could handle various types of data. Unfortunately Stevens' ideas were not the only ones promoted but they were marketed well and were possibly understood the best. This situation is easy to understand once thought is turned towards the prevailing scientific spirit of the times, which has been discussed at length in chapter 3. Unfortunately, assigning numerals to objects is completely ill-befitting to the serious enterprise of psychology and its construct assessment. There is more to measurement than scales. Far more.

Primitive or extensive measures are quite easily come-by in the natural sciences where concatenation, representation and subsequent measurement is possible, at least more so than in the social sciences. Results in the natural sciences look robust and generally appealing to scientists who are able to measure even objects which are not observable, albeit in a indirect manner however psychology cannot even say it can measure observables let alone unobservables. Thus, in an early attempt to appease discordant opinion, measures were assumed to be possible and subsequent measures on many fronts were undertaken. Most notable are those measurements stemming from psychophysics which even in its name is redolent of a natural science endeavour. By utilising carefully worked out statistical science along with measurement theories, a robust veneer was granted to an essentially soft discipline. The unnerving aspect about this is that a perfectly scientifically valid qualitative science can evolve within psychology without its having to subscribe to natural science dictates – but it seems that the research world and the public at large still do not recognise this fact. The consequences play forth in areas such as dynamic assessment which seek to encourage and assess change in a manner quite foreign to mainstream assessment.

Physics is not psychology and psychology is not physics, so why then do we insist on utilising physics methods in psychology? It can be assured that physics does not utilise psychological methods in its studies.[112]Nevertheless in psychology's continual pursuit as robust science it devised for itself a basis of verifiable and justifiable measurement in the sense understood within the natural sciences. It accomplished this by representing constructs in a manner which allowed properties to become jointly assessed and hence amenable to measurement in the extensive sense of measurement. Intensive measures were now cased in extensive frameworks. Conjoint measurement, although still questioned as to its facility to really assign objects to measurement, has become an accepted part of modern test theories. Axioms of measurement are carefully considered within their mathematical systems and have proofs derived for them in order to validate their postulates. Much recent work in this field has been conducted, among recognised others, by Joel Michell, a researcher in Australia and Paul Barrett in New Zealand. Criticisms and thoughtful concerns regarding classical as well as modern theories of measurement placing special emphasis on construct validation has been conducted by, among recognised others, Denni Borsboom in the Netherlands. Their geographic locations may not, at first glance, seem important but when one considers the origins of dynamic assessment as emanating principally from Russia (Soviet Union), Israel, the Netherlands and Germany interesting questions start to be asked. Mainstream testing has had a bountiful history within Britain and America (although predicated on different philosophies throughout). These facts only serve to raise tentative questions at this stage. Why have alternative mechanisms of assessment principally arisen in countries outside the areas of mainstream assessment such as the United States and Britain for instance? South Africa, at present, is pulled both ways but seems to have sided more with the mainstream manner of assessment and basic underlying philosophy. Likewise notable criticism of theories of mental testing has emanated from countries outside the United States and Britain. Is this due to local minority and/or immigrant populations in countries outside the United States and Britain? These two countries have had to contend with an influx of immigrants especially since the Second World War whereas other countries have consistently existed with local minority groups. This brings to the fore the issue of cultural relevance and the use of tests to keep people out as opposed to letting them in. A contentious issue at the very least but a plausible reason as to why criticisms are lodged at mainstream assessment which emanate from countries outside mainstream concerns.

Newer developments within mental test theory are evidenced in the works of modern item response theory, a method whose time has now truly arrived. Due to the lack of computational power in the 1940-1970's many techniques were simply too complicated and power-hungry in order for them to be utilised, especially by mathematically unskilled psychologists.[113] Classical test theory is a powerful theory of reliability and should not be neglected in the future but rather used concurrently with other newer techniques and test theories. Generalizability theory which is considered as a branch of true score theory allows for more flexibility in terms of relaxing certain assumptions within its predecessor. Modern test theory offers solutions to various measurement issues that have confronted psychology but one troubling aspect still remains, which is unlikely to be solved by any measurement technique and that is the burning issue of construct validation. For this, substantive theory is needed which in psychology's present-day status, needs to be revisited in terms of its methodology and philosophy. Modern test theory has partially aided dynamic assessment's quest to define for itself a measure of changeability, which in the true score model is almost impossible. These change score models were discussed in some detail so as to illustrate their encompassment of change within the measurement model which represent the new rules of measurement. This was done for a reason. By seeing which representations are being utilised for change assessment, the model can be questioned directly. There are numerous Rasch related IRT models available which avail of themselves mechanisms to assess change and only five were discussed. Clearly, if dynamic assessment is to stay within mainstream assessment, its best bet at this stage would be to follow up on and make use of such change-based IRT models whilst utilising useful information and insights from classical test theories and older models of reliability. If dynamic assessment is to forge ahead on its own unique trajectory then it should avail of quantifiable non-numeric measurements. Perhaps hybrid dynamic assessment models can coexist in both realms. The debate, however, has only just begun.

4.5 Integrating Madsen - attenuating the framework

In attempting to attenuate Madsen's quantifiable systematology it cannot be ignored that criticisms have been levelled at his technique and approach towards documenting a method of theory judgement (Brandt, 1984; Ettin, 1984; Schaeffer, 1984). What is central to this study is Madsen's technique and not necessarily his actual findings in his own area of interest and research. In other words problematic issues within Madsen's work such as reliability including selection, inclusion and rendition criteria (Ettin, 1984) are simply not agreed upon by critics and are not explicated to any great extent by Madsen himself. Hence the need to attenuate and support the need for utilising his approach; not to mention the rather obsolete usage of S-R terminology[114] which is quite dated as the stimulus-response model of human behaviour is not only crude but incorrect due mainly to its simplistic

[112] The author is aware that there is a continuous flow between the disciplines in terms of importing ideas but there is not much crossover in terms of methodology; physicists do not infer unknowable causes from effects even in areas such as quantum physics!
[113] Perhaps this is not always a fair argument. After all, psychologists are not mathematicians nor are they statisticians. However, this does not mean that utilisation of techniques without concern for how they were derived should proceed without some measure of understanding of the underlying mechanisms. To state that one is not informed about certain mathematical formulations is to take the easy way out.
[114] The notation will be used throughout chapter 5 in keeping with Madsen's convention of scoring the HQ.

stance on human functioning. The author is of the opinion that Madsen offered his systematology of theories of motivation as one researcher's contribution to the field and did not, in contrast to what Ettin (1984) implies, expect that such an analysis be taken as a given in future readings of the text. This is open to speculation, yet there is sense in the criticism lodged at the lack of reliability and validity scaling for his technique and that it in fact does not, in its current form, warrant the status of empiricism (Ettin, 1984). Madsen's conception of a meta-theory has been classed as more of an heuristic aid than empirical certainty (Brandy, 1984; Ettin, 1984) and due to his approach being described as descriptively sound, cogent and parsimonious yet in need of revision in some quarters (his HQ system) the study utilises his approach.

In attenuating and building a framework for dynamic assessment and intelligence, all that has been discussed thus far in chapters 1-4 become enmeshed into the framework in such a way, as will be seen, to allow for greater flexibility within Madsen's original scheme in order to accommodate this study. Madsen, for instance, did not take cognisance of the prime considerations within psychology, namely the mathematical, statistical and measurement foundations permeating assessment; thus this will be worked into the framework alongside his original elements. The first consideration will be the culture and society in and from which various models and theories have developed. As already hinted at, different viewpoints have emanated from different countries possibly due to influences not always directly perceptible. The scientific community from which the individual researchers practise will also be highlighted. Within any model or theory, the nature of the science practised will be looked at and will include the nature of the empirical research, the theoretical thinking behind it as well as the underlying philosophy. Meta-levels of concern also focus on these three levels but extend their range somewhat to include philosophies of science, psychology and their respective histories. Recall that theories become data within the meta-theoretical framework and are treated at the empirical level. Theories are a culmination of deep-seated philosophies and hypotheses, conjectures as well as data theses whereas models function as heuristics which in turn are either theoretical, empirical or a mix of both. Theoretical heuristics serve to constrain whilst empirical serve to detail and locate. Such models are aligned according to their degree of abstraction: material (descriptive), graphic (explanatory), simulation (ontological) or mathematical (symbolic). These dimensions will serve to support the model framework. The data stratum is the level at which measurement philosophy enters Madsen's framework. As this is a vital component within the thesis in attempting to direct dynamic assessment's measurement schemes, it will need to be brought in at this stage.

Prime considerations as discussed in chapter four will thus form part of the data level analysis (even though it can be easily accommodated in the epistemological realm) and will concentrate on the statistical and measurement issues more so than on the mathematical ones. Hence the framework will proceed as follows: consideration of meta-level issues such as ontology, philosophy; hypothetical-level issues such as hypothetical terms, scientific hypotheses and the hypotheses system and data-level issues such as abstract data and concrete data issues. An attempt will be made to utilise Madsen's hypothesis quotient (HQ) and depending on the ease of application, the results should illuminate, at least at a very basic level, the testability of various models and theories. Visually the framework as illustrated in figure 71 will utilise the notation so as to ensure clarity. Thus when the models/theories are assessed according to various dimensions, notation such as "A(ii)" and "B(i)" and so on will be used. The hypothesis quotient will be worked out according to the model/theory's amenability to such quantified rendering. One very positive aspect or feature about this process, is that the same framework is utilised throughout thus yielding a standardised rendering of the models. Pitched at the same level, so to speak, there is at least some minimal basis according to which comparative assessment and judgments can be made. Thus, although not everyone may agree to the synopsis provided by working through the framework, one can at least use a common metric to assess these different models and theories and conclusions can be reached.

Figure 71 The attenuated Madsenian systematology framework

A: Meta-stratum (chapter 2 as necessary prelude)

A (i) Ontology:

- How does the theory/model comment on or is influenced by the conception of the human being, the mind/brain thesis and human freedom?

A (ii) Philosophy:

- How does the theory/model account for its epistemological modes of conception in terms of cognition and its relation to reality?
- How does the theory/model comment on or is influenced by meta-theoretical concerns such as nomothetic, hermeneutic and idiographic ideals?
- How does the choice of the theory/model's methodology impact on the outcome; what test methods and data language is utilised?

B: Hypothetical-stratum (predominant attenuation occurs within this level)

B (i) Hypothetical terms:

- How does the theory/model account for its ontological reference, existential form and function of hypothetical terms? In each of these classifications are concerns which will be noted, however should they prove unnecessary or superfluous within this thesis'context such concerns will not be dealt with, hence an attenuation of Madsen's framework.

B (ii) Scientific hypotheses:

- How does the theory/model account for its ontological classification of scientific (testable) hypotheses as well as its meta-theoretical classification of scientific hypotheses? In each of these classifications are concerns which will be noted, however should they prove unnecessary or superfluous within this thesis' context such concerns will not be dealt with, hence an attenuation of Madsen's framework.

B (iii) Hypothesis system:

- How does the theory/model account for its deductive explanatory system as well as its model explanations and what is the degree of abstraction involved? In each of these classifications are concerns which will be noted, however should they prove unnecessary or superfluous within this thesis' context such concerns will not be dealt with, hence an attenuation of Madsen's framework.

C: Data-stratum (inclusion of chapter 4)

C (i) Abstract data:

- How does the theory/model account for its functional relations between variables?

C (ii) Concrete data:

- How does the theory/model account for various forms of concrete data such as evidenced from behaviour?

C (iii) Prime considerations (chapter 4):

- How does the theory/model account for and defend its use of statistical techniques?
- How does the theory/model account for and defend its measurement models (mental test theories)?

4.6 Conclusion

Psychological assessment literature more often than not concerns itself with measurement issues as currently understood via mainstream assessment issues. Undergraduate and postgraduate assessment training very often comprises issues pertaining to classical test theory and the newer item response theories of measurement. Very rarely does the text delve into the mathematical foundations upon which these theories are predicated. Here reference is being made to philosophical mathematical issues and not only issues pertaining to test theory models per se. When psychologists purport to test, assess and in some or other manner measure constructs, they are leaning on a long history of mathematical dispute and disquiet of which many are blissfully unaware. The fact that mathematical tensions exist between what is knowable and what is not knowable hardly ever receives a mention in psychological assessment texts. Number crunching, statistical manipulations and the requisite details underlying measurement theory are often left to researchers within these respective fields. A continuous thread spans endeavours in mathematics, statistics and measurement where their culmination is manifested within assessment. Although the philosophical and historical foundations of mathematics may seem to be far removed from assessment upon first glance, it is vital that these issues are at least understood and looked at before tests are thoughtlessly applied in practice. Perhaps the most important fact for psychologists is that so-called firmly entrenched inviolable proofs are in fact predicated upon axioms which are themselves improvable. What does this mean for the practising psychologist? A great deal of angst and concern if one is to fully understand that cherished ideas of mathematical superiority and certainty are in fact not so superior nor so certain. "Axioms" as a term should perhaps be revised; perhaps we should consider them as temporary axiomatised systems in waiting. Probability as a concept has played a monumental role within social science statistics and measurement and has rightly pervaded (or invaded) various sub-disciplines to a large extent. Probability is part of mathematical thought and thus needs to be addressed when one seeks to understand what its function is within psychology. The probability of an event occurring, or that an experimental technique will work hinges on its mathematical predicates and history. The story of mathematics, statistics and measurement is a long one spanning millennia, so it can hardly be ignored. Having said this, the question is asked once again: why the need to consider mathematical foundations? It is hoped that this question has partially been answered.

The statistical foundations of the social sciences really is the work of giants on whose shoulders we gladly stand; however stagnation at such lofty heights may well prove to be our downfall. Pioneering statisticians devised schemes in manners that would leave most current statisticians and historians in disbelief. Computational power was not even a consideration during the heady days of Galton, Fisher, Neyman and Pearson (Karl and Egon) and constant reminders of the possible inadequacies of their insights has to be greatly tempered with their astounding achievements given their only source of computational power – their intellects. Hard work, perseverance, determination and considerable skill allowed such individuals, among others, to propel the nascent field of statistics in a direction which was to serve psychology for the next century. But all is not well. Consistent misrepresentation and misunderstandings of the rationale behind the use of many statistical techniques has partly resulted in an extended era of number crunching at the expense of substantive theorising. As stated in chapter 3, psychology (as with other disciplines) moves at the behest of policy, law-makers and society's needs and wants. Who are we to argue about grants which come our way if we can prove our techniques are psychometrically defensible and robust despite their logic being flawed? This decades-long stalemate will eventually lean in favour of one or the other direction. Pundits have lauded the entrance of thoughtful statistics and decried the continued use of outmoded and dated techniques within psychology specifically. Papers are cited as being peppered with *p* values, asterisked to such an extent that it becomes an almost gross violation of what it means to be significant. The crud factor has entered into a realm which can happily accommodate yet more data and more values less than 0.0001! How low can you go?[115]

The subsequent confounding of statistical significance with that of substantive significance as well as arbitrarily chosen labels at which to accept or reject findings within experiments which are themselves biased towards findings in one direction has made for experimentation within psychology an enterprise of anticipated number crunching. Where is the substantive psychology in all this? Dynamic assessment's process-based qualitative interest in the renewal of cognitive functioning, it seems, is ill-fitting to such a scheme. Can the complexities of cognitive maturation and skill development as observed within intervention processes really be contained within a framework and structure allowing for only certain levels of significance to be obtained within certain statistical tests? Brain development is not equivalent to the processes involved in determining the level at which certain metals undergo molecular decay and hence disposed of in a production line (where significance comes into its own and is especially applicable and robust; saving time and money). A re-look at the use of statistical practice among psychologists prompted numerous debates concerning the over-use of null hypothesis significance testing and under use of other manners of treating data (such as confidence intervals and use of Bayesian statistics for instance). However, this particular concern with statistics is perennial and has a history extending as far back as the 1960's. Dynamic assessment was cloaked in an already established framework of statistical assessment of data and as such, has hardly had much of a choice of techniques at its disposal

[115] Given a large enough database, shoe size really can be shown to be significantly correlated to weather patterns.

especially if it was to become more generally accepted into mainstream assessment practise. Psychology and statistics within the social sciences developed almost in tandem and both fields informed and gained information from the other. There was no fully developed statistical framework according to which data could easily be analysed and custom treated depending on the context. Psychology had to develop in its own manner and borrowed techniques from statistical sources but also stimulated the development of novel and widely used techniques, some of which have spread to other disciplines involving the more natural science orientated.

Although useful, NHST testing necessitates a concerned revamp and a renewed look at Bayesian statistics is warranted, especially within the field of dynamic assessment which resolves to understand prior and post mediation performances. Instead of basing results on tests accounting for only information written in the test, a Bayesian approach can accommodate prior information on the individual which does not necessarily have to emanate from a test per se. This is fairer and more inclusive and in keeping with numerous aspects propounded by dynamic assessment's basic philosophy and serves as a reminder than there is more to inference than frequentist approaches. Results, when they are necessary, should attest to a flexible framework in which individual case studies are assessed on their merits and which do not yield to the strict and short-sighted "accept-reject" mode of practice within NHST. Studies should also be amenable to a falsificationist approach as opposed to a verificationist one. This larger reality, however, encompasses variables which cannot be ignored and in finding for itself another way of dealing with the statistical within dynamic assessment research, the broader scientific community needs to re-arm itself with a variety of tools which can be deployed. One such tool is the choice to allow non-significant studies into publications, as it is replication upon which a science should be predicated. Replication, it might be said, is taken into account within statistical modelling as in the case of the power of a test, which is cost and time saving; but when conducting substantive research or when research is at the level requiring more substantive work, then this spirit should not prevail. Dynamic assessment is at just such a level of substantive research.

Simply put, measurement entails numericising objects. Objects can be physical entities which can be directly observed or indirectly ascertained via means of skill inference from what can be observed or what is known. Indirect inference can proceed from objects known to exist but known only through mathematical abstraction. Many objects of the latter type have been predicted and later discovered yet remain unseen in the conventional sense. Psychological variables, however, cannot be said to have attained the status of existing object whether seen or not. Representation of objects does not proceed along a manner akin to merely assigning a number to such an object but includes rules and axioms of representation. Numericising such objects has posed, what some may consider, to be an insurmountable problem. Means exist through which such unobserved entities can manifest via numerical relations; this enterprise within psychology is called test theory and encompass older and newer forms of mental test theories, each with merits and disadvantages. Measurement is, however loosely, predicated upon a mathematical set of axioms, some of which have been delineated and proved. Decades of debate concerning the nature of mental measurement has periodically resulted in committees being constituted in an attempt to further understand the nature of measurement and its requisite tools and how best to go about the business of testing and measuring. The most salient question hovering above all this has been whether psychological entities, such as they exist, can at all be quantified.

Assuming a quantified structure, classical test theory progressed primarily from a theory supporting the notion of a general factor of intelligence and was preoccupied with the reliability of scores. True scores, in this model, are never directly knowable, rather, they are computed from error in addition to the observed score. Curbing or at least accounting for error as robustly as possible by factoring random and chance factors it is assumed that the true score is attained to some degree. Item response theory jointly computes individual ability along with the difficulty of item in a probabilistic model thus allowing for direct measurement of intensive measures, thus overcoming, at least partially, the deficits of numericising intensive measures that classical test theory could not adequately manage. In addition to this, change within testing is more readily accounted for in IRT which is now able to model change utilising various models of change-based multidimensional models, many of which are based on the original Rasch models. Such early models were being considered theoretically in the 1950's but due to lack of computational power were not amenable to practice testing. Luckily, there is evidence of a trend towards increasing use of IRT models, although some researchers still question the degree to which even these techniques (although superior in many ways) are able to account for a truly representational manner of measurement. The core question remains as to whether there exist any entities which can avail of quantification in the first place. This would of course necessitate a re-look at substantive psychology encompassing renewed interest in theoretical psychology, an issue discussed in chapter 3. This chapter concludes with the attenuated Madsenian framework which will be deployed in chapter 5.

A point has been reached in this thesis in which the developed framework can be utilised for purposes of exploring a meta-theoretical framework for dynamic assessment and intelligence. Chapter 1 briefly introduced the need for such a framework within the area of dynamic assessment research. Chapter 2, although in a measure testament to the author's particular view points, was a necessary prelude to chapter 3 which dealt with some concepts that were evidenced in Madsen's framework and in order to more fully appreciate some of Madsen's views certain dimensions received attention in chapter 2. Chapter 3, in keeping with the framework, discussed the importance of identifying psychology's formal existence as having sprung from a natural science orientation and that due to historical circumstances was framed and dictated to, to a certain degree by the

prevailing climate of the times. In addition to discussing the role of natural science explanatory mechanisms as well as those of social sciences, attention was turned towards psychological explanatory mechanisms. Chapter 3 continued with the detailing of a meta-theoretical framework as developed by the Danish psychologist, K.B Madsen. In pursuance of a goal of unifying framework, the development of a framework was not yet complete. The framework necessitated additions in terms of prime concerns running throughout psychological assessment. As mentioned Madsen's framework was originally developed with theories of motivation in mind and not psychological assessment per se. The need to include epistemological and ontological measurement foundations in the area of assessment is paramount when the future of dynamic assessment within intelligence is concerned. Thus chapter 4 needed to be incorporated into the existing framework. The emphasis within the framework is placed more so on the statistical and measurement foundations as it pertains to dynamic assessment and intelligence, however the mathematical foundation discussion was necessary in order to highlight various aspects within the statistical and measurement foundations. Chapter 5 applies the framework detailed thus far and attention is now turned toward this culminating chapter.