Science and Mathematics Education Centre

Measurement of Challenge and Self-Efficacy in Learning

Michael Lopez

This thesis is presented for the Degree of

Doctor of Philosophy

of

Curtin University

June 2012

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: .....

30/5/2012

ABSTRACT

Students' perceived self-efficacy is an important factor in determining their educational success. Those with high self-efficacy are likely to engage readily in learning activities, and to persist with their studies in the face of adversity. Those with low self-efficacy may shy away from engagement in activities that they perceive as challenging, and may give up when problems are encountered. The perceived difficulty of learning tasks is also important. If there is insufficient challenge, students might not value the learning and thus disengage. However, if the material is too challenging, learners may lose confidence in their ability to master the material and may also disengage. Getting the level of challenge right is thus a key factor in designing learning experiences that will engage learners and build their confidence in their ability to master challenges in their future.

This thesis investigates the possibility of providing educators with objective evidence of students' self-efficacy and the perceived challenge of the learning activities in a course. The investigation includes both theoretical and practical perspectives.

The theoretical perspective involves the derivation of a formal measurement model, together with its theoretical and conceptual underpinnings. When the assumptions of the model hold, the model produces objective linear measurement from ordinal data, accompanied by estimates of uncertainty, conjointly for students' self-efficacy and the challenge of activities. A comprehensive framework of hypotheses is developed to test the assumptions of the model, from the theoretical quantifiability of the constructs, through to fitness for purpose. A software implementation of the model is developed and evaluated from theoretical and empirical perspectives. Evidence of construct validity is provided. The measurement model used required a reimagining of measurement theory from an information theoretic point of view. This point of view adds to the theoretical understanding of measurement when items with multiple categories are used.

The practical perspective involves the development of conceptual and representational frameworks that are readily understood and interpretable by

educators. The key elements are the use of multiple output representations, the use of appropriate analogies and metaphors, the mapping of statistical and information theoretic terms and constructs to equivalents that are more familiar to educators, and the reimagining of reliability as a measure of fitness for purpose. Two alternative approaches to the measurement of self-efficacy are investigated: one based on direct reporting, and one based on inference from engagement in activities. Both are found suitable for the measurement of self-efficacy.

The model is robust under the introduction of random noise and violation of the core assumption of local independence. Relatively few subjects and items are required to achieve useful measurement accuracy and this can be achieved in realistic educational settings. Overall, it was found that measurement of self-efficacy is both practical and useful in a realistic educational setting.

ACKNOWLEDGEMENTS

The work described in this thesis has been a challenging, but satisfying, journey for me. This journey would not have been possible without the help and support of many people.

First, I would like to thank my supervisor, Darrell Fisher of Curtin University. In particular, I appreciated his gentle way of encouraging me and keeping me focused on this journey. I valued not only his wide experience, but his way of motivating and interacting with students; he has become a role model for my own engagement with students.

I would also like to thank the “beta” group of educators who trialled the software while it was being developed. Their forgiveness when issues were discovered was appreciated, and their insights and feedback were invaluable. I would especially like to acknowledge the contribution of the many students who used the Salsa software, and participated in the research. I will remember their cheerfulness, encouragement, support, and constructive feedback for a long time to come.

I would like to thank my many educational colleagues for their support and encouragement throughout the journey and, in particular, Amit, Dave, Mike, and Alison at Christchurch Polytechnic Institute of Technology.

I would also like to thank my family for their support, and for their tolerance, throughout this journey. At times they must have felt that any conversation inevitably led to a discussion on principles of measurement. In particular, I would like to thank my mother, who brought me up to value the sense of achievement that can only be reached by hard work and perseverance, and my wife, Dobrila, who always found time to review my work, and make endless cups of tea, despite her own substantial workload.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1.
INTRODUCTION

A little railroad engine was employed about a station yard for such work as it was built for, pulling a few cars on and off the switches. One morning it was waiting for the next call when a long train of freight-cars asked a large engine in the roundhouse to take it over the hill "I can't; that is too much a pull for me," said the great engine built for hard work. Then the train asked another engine, and another, only to hear excuses and be refused. At last in desperation the train asked the little switch engine to draw it up the grade and down on the other side. "I think I can," puffed the little locomotive, and put itself in front of the great heavy train. As is went on the little engine kept bravely puffing faster and faster, "I think I can, I think I can, I think I can." Then as it near the top of the grade, that had so discouraged the larger engines, it went more slowly, but still kept saying, "I--think--I--can, I--think--I--can." It reached the top by dint of brave effort and then went on down the grade, congratulating itself, "I thought I could, I thought I could."

The quotation above is taken from *Thinking One Can*, author unknown, which was published in 1906 in Wellspring for Young People, a children's Sunday school publication. The story was later popularised in the children's book, *The Little Engine that could,* by Mabel Caroline Bragg, writing as Watty Piper (Piper, 1930). The central idea in the story is the connection between achievement and belief in one's capabilities. This belief in capabilities is now called self-efficacy. Bandura (1994) defined self-efficacy as "people's beliefs about their capabilities to produce designated levels of performance that exercise influence over events that affect their lives" (p. 71). An important distinction is made between capabilities and

abilities. Abilities refer to what a person can do in the present whereas capabilities refer to abilities that could be developed. A similar contrast was made by Vygotsky in developing his *Zone of Proximal Development* which he defined as follows:

> It is the difference between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers. (Vygotsky, 1978, p. 86)

A strong sense of self-efficacy fosters intrinsic interest and deep engrossment in activities. In contrast, people who doubt their capabilities may shy away from difficult tasks which they view as personal threats. They may slacken their efforts and give up quickly in the face of difficulties. This has clear implications for educators. If there is insufficient challenge in learning (and assessment) material, students will not value the learning and may disengage. However, if the material is too challenging, learners may lose confidence in their ability to master the material and may also disengage. Getting the level of challenge right is thus a key factor in designing learning experiences that will engage learners and build their confidence in their ability to master challenges in their future.

Educators routinely face the need to make judgements about the level of challenge in their courses[1] and the ability of learners to meet such challenge. These judgements are informed by many sources, including theories of learning, communities of practice, an educator's experience, and course specific information. This last may involve either gathering supplementary data on the learning experience, or the analysis of data that occur naturally.

---

[1] The use of the term "course" varies between tertiary institutions. This project uses the terminology of the author's host institution which is summarised here. Students undertake a programme of study which leads to an award such as a degree; other institutions might refer to such a programme as a course. Each programme is made up of a number of courses which usually represent about 150 student learning hours and are taught over one semester of study. In other institutions, these courses might be termed papers, modules, or units.

Supplementary data are often gathered routinely as part of formal course evaluations. In addition, individual research projects may also explore specific course related issues. Although both of these are valuable, it is difficult to gather data in sufficient detail, and frequently enough, to inform day-to-day decisions and detailed course design without affecting the learning experience itself or imposing a major burden on learners. The use of naturally occurring data seeks to minimise this burden. It also keeps the data close to practice and minimises the effect that gathering the data has on the learning experience. Traditionally, summative assessment has been the main objective technique that uses naturally occurring data to supplement an educator's judgement and knowledge of the cohort of students. Over time this may align learners' beliefs in their ability with those of the assessor. Analysis of responses to individual assessment items may also give the educator insights into what was well understood, and what was less well understood. However, the summative basis of the data often means that any insights gained cannot be applied until the next delivery of the course.

This raises the question of whether other naturally occurring data could be used to inform educator judgements of the level of learning challenge posed by the course, and learners' beliefs that they have the capability of meeting the challenge (self-efficacy). It may be noted that, rather than measures of ability, it is a learner's belief in their capabilities that is central to self-efficacy. This suggests that it may be promising to base measurements and educator judgements on self-report data. Many courses include such self-report elements and this research takes its starting point from a specific tool: "Salsa", which is an acronym for the *Systematic Analysis of Learner Self-Appraisal.* Salsa (Lopez, 2005) is a software tool that is used to build and support reflective and metacognitive learning skills. Students use the tool to monitor and reflect on their learning progress, and communicate this in confidence to their tutors. Regular reports from the software then form the basis of a constructive learner/teacher dialogue. The software records the progress reports in a database, giving the potential of further analysis. These reports include both engagement in activities and belief in capabilities associated with course topics; each of these is recorded on an ordinal scale.

There is a direct mapping to self-efficacy from students' beliefs in their capabilities associated with course topics. However, it is also possible to make a connection between engagement in activities and self-efficacy. Making such a connection suggests that it might be possible to estimate self-efficacy by inference from data on engagement in activities. The conceptual relationship between engagement in activities and self-efficacy is shown in Figure 1.1. This relationship suggests that inference about self-efficacy might be possible where there are data concerning starting activities, persistence in activities, and the perceived outcome of activities.

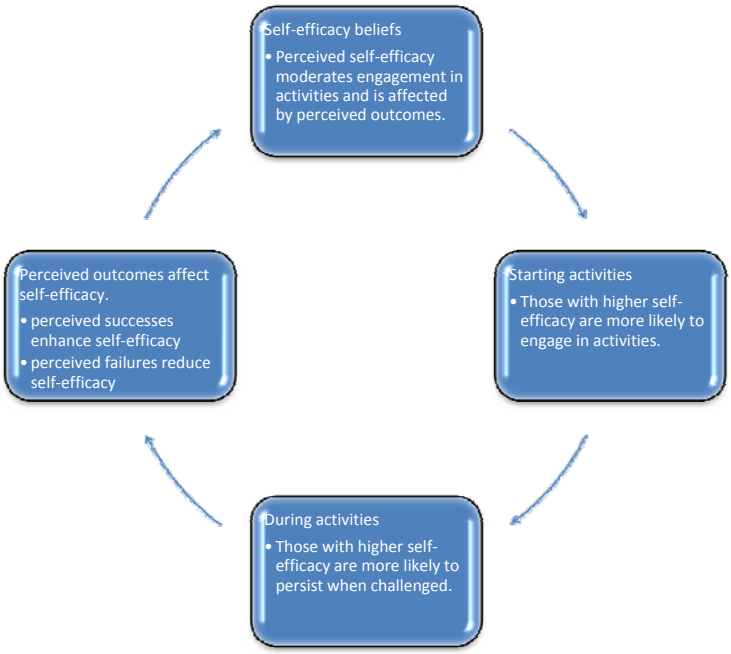


Figure 1.1: Conceptual relationship between self-efficacy and engagement

These data are available from Salsa but, perhaps more importantly, if such inference proves possible, then it may be possible to estimate self-efficacy without the need for the Salsa software, wherever equivalent data on engagement in activities are available.

The student reports from Salsa are ordinal in nature. Translating these into linear interval level measurements will enable formal investigation of the relationships set out above. However, there is no absolute unit of self-efficacy, which must therefore be defined relative to established norms or the difficulty of the tasks undertaken. This research takes the latter approach and uses the term *challenge* for the

difficulty of the tasks. In this sense, challenge and self-efficacy are complementary terms. Measuring both conjointly, and on the same metric, allows meaning for each to be constructed by reference to the other.

However, conjoint measurement places formal requirements on the constructs and data (Luce & Tukey, 1964), and it is important to verify that the assumptions are met. This leads to the first research question:

> Is it possible to develop objective measures of challenge and self-efficacy from self-report data? (Objective 1)

Provided the assumptions are met, the creation of objective measurements from ordinal data is well understood (e.g. Andrich, 1978; Rasch 1960/1980). However, the natural unit of such measurement is the log of an odds-ratio. This is unlikely to have inherent meaning to many learners and educators, who are more familiar with scores, marks or percentages. Measurement of challenge and self-efficacy will have limited usefulness unless the measures can be readily understood by both learners and educators. This leads to the second research question:

> How can these measurements be communicated clearly to educators and learners? (Objective 2)

Moreover, for the measurements to have a practical impact on learning, both learners and educators need to be able to act on the information they receive. This requires that they receive the information in a timely manner and with minimal effort. Ideally, the information should be available at a convenient location whenever they need it. Making information available at a convenient location is straightforward, but providing it in a timely manner is more problematic. Conventional approaches to objective measurement typically require all data to be collected before measurement is undertaken. At first sight, computerised adaptive testing (CAT) might appear to be an exception to this, but CAT still requires pre-calibrated item banks (Van der Linden & Glas, 2000).

In order to provide real-time feedback to both educators and learners, reasonable estimates of all parameters need to be made before all the participant data have

been collected. Essentially, the goal is that correct measurements are made and reported at all times. As more data are accumulated, the confidence limits of each measurement will narrow, giving better estimates. Students, in particular, may be used to this way of working. For example, many "Web 2.0" sites include the concepts of ratings. Viewers may have little confidence in ratings when few users have rated an item, but confidence grows as more user ratings are aggregated. To bring rigour to this concept, an information theoretic approach is taken. This approach treats responses as sparse data and accumulates information (thus reducing uncertainty) as data build up. It also requires, *inter alia*, a graph analysis of responses and connectedness to ensure that a common metric is maintained. For the software tool to be practical, it is important that meaningful estimates are produced with relatively few participant responses. This leads to the third objective:

> Develop a practical computer software implementation that will communicate the measurements in real time (Objective 3)

There are, however, some limitations to the generality of the approach taken in this thesis. For valid measurement, the constructs must be one-dimensional which, in turn, requires that:

- A course forms a coherent body of learning. This implies that it is possible, at least conceptually, to use a single metric for learning.
- There is a constructive alignment (Biggs, 2003) between course learning outcomes and the learning episodes of the course. This implies that the common metric may be used throughout the course.

The first requirement is not onerous; any course for which it is possible, at least conceptually, to have a valid assessment regime leading to marks or grades will meet this criterion. The second requirement is more problematic. If different episodes in the course relate to different constructs, the measurement model will only extract information from the commonalities between these. Provided there is some commonality (as implied by the first assumption), this does not affect the validity of the measurements. The correct estimates will be reported, but with an

increased margin of error that is associated with the contribution of the unaligned component. However, this increased uncertainty may impact the usefulness of the measures for courses with significant unaligned elements.

This last point touches on the central aim of this work which is to provide and evaluate a software tool that provides practical and useful measurement for educators. It is hoped that the provision of such a tool, with a formal objective model for the measurement of challenge and self-efficacy, will help educators gain a better understanding of the role these play in the courses they design and teach. This enhanced understanding may, in turn, lead to improved course structures, and better student engagement and outcomes. Learners, too, will get a better understanding of the challenges of the course and how well they are coping. The mechanised analysis of naturally occurring data will provide a rich and timely source of this information without imposing an additional burden on learners. The representational model developed will help bridge the gap between the formal methods used in objective measurement (Andrich, 1978; Rasch 1960/1980) and practitioner understanding. This may facilitate wider use of objective measurement techniques in education.

It is also hoped that this work will contribute to understanding of measurement theory. The development of a time-series extension to the Rasch model is unique to this work, and another key innovation is the development of an effective on-line algorithm. An algorithm that works with incomplete and progressively unfolding data has wider implications than just this project. As the world becomes more connected digitally, large amounts of data are being accumulated in databases, and on-line algorithms will become increasingly relevant. Within the education field, students may be assessed at different times and places rather than an entire cohort taking an assessment at the same time. If this happens, it may be important to analyse results for quality management before all students have completed the assessment and without the need for large pre-calibrated item banks.

The remainder of this thesis is organised as follows.

Chapter Two introduces a formal framework for measurement. Conceptions of measurement are reviewed from the perspectives of Social Science, Mathematics, Physical Science, and Metrology, and the fundamental requirements of measurement are set out. A review of measurement process sets out the historical context and establishes the need for a theoretical measurement model. A reference model is presented and used as a basis for comparison with a number of existing measurement models. The proposed model is presented and extended to cater for a time-series, as a unique contribution of this project.

Chapter Three explores the perspective of *measurement as hypothesis*. Use of a measurement model does not automatically guarantee that measurement is valid or useful. The assumptions of the model are discussed and threats to validity are identified. Statistical tests for each of these are introduced. A comprehensive framework for the evaluation of the success of the measurement exercise is presented.

Chapter Four summarises the development of the software. The project used a Design Science methodology (Hevner, March, Park, & Ram, 2004) to develop the main software artefacts. Special attention is paid to the human factors that enable the software to be usable and interpretable by educators in a typical setting. To accommodate this, an expert system paradigm was chosen and a number of output representations were developed. Evaluation of the model on small and large scale reference datasets provides basic evidence of convergent and discriminant validity.

Chapter Five presents the theoretical evaluation of the model and software. Simulated datasets were used to carry out a systematic exploration of the performance of the model under a wide range of conditions. This exploration serves to evaluate the correctness of both the implementation of the model, and the theory underpinning its construction. The robustness of the model to data that do not fit the model is explored, and practical benchmarks for use are derived.

Chapter Six introduces Bandura's concept of self-efficacy, the central component of his *Social Cognitive Theory*. It briefly outlines the relevance of self-efficacy to education and discussed some of its implications for educators. Two approaches to

the measurement of self-efficacy and challenge are introduced: the first based on direct self-reporting of perceived capabilities, and the second based on inference from an observed pattern of engagement in activities.

Chapter Seven presents an empirical evaluation of the model and software using authentic datasets: data collected by the Salsa software was used as the source of self-report data. An overview is given of the Salsa software and the self-report data available. Evidence is presented for the construct validity of self-efficacy. The question of whether useful measurement of self-efficacy and challenge is achievable in a real classroom setting is addressed, and the two approaches to measurement of self-efficacy and challenge are compared.

Finally, conclusions are drawn in Chapter Eight.

Chapter 2.
MEASUREMENT

An experiment is a question which science poses to Nature, and a measurement is the recording of Nature's answer ~ Max Planck

The connection between an experiment and a measurement in the quotation from Max Plank (1968) above serves as an introduction to the proposition that at the most fundamental level measurement is a hypothesis rather than a process of assigning numbers. Once the hypothesis that an attribute is quantifiable is tested and accepted, measurement in the sense of systematically assigning numbers can proceed. The perspective of *measurement as hypothesis* is explored in the next chapter.

This chapter is organised as follows. It begins by reviewing conceptions of measurement from the perspectives of Social Science, Mathematics, Physical Science and Metrology. From this review, the fundamental requirements are set out for measurements to meet all these conceptions. These are the need for: linear measurement, a statement of associated uncertainty, and interpretability.

The second section sets out some key developments in measurement theory from a historical perspective and then reviews the process of measurement. Attributes to be measured may exist in the real world, or be abstract, but the measurements produced are always abstract. The process of measurement thus requires a theoretical model to make the connection between the measurand and the measurement. A good model will simultaneously provide linear measurements, statements of uncertainty, and fit statistics that serve as a test of the underlying measurement hypothesis.

The third section presents an abstract measurement model that is capable of producing objective interval level measurement from comparative ordinal judgements. Although introduced primarily for purposes of exposition, this could be used as the basis of any measurement model.

The fourth section reviews a number of alternative measurement models. A reference model is derived from the abstract model, and used as a basis for comparison. A related polytomous version is also derived and compared to existing polytomous models. The model is then further extended to cater for a time-series.

A formal presentation of the model used in this project is given in the fifth section. This includes derivation of the equations used to define the detail of the model and algorithms for its solution. It also describes the key model outputs and statements of uncertainty.

Finally, the key features of the measurement model used in this project are summarised in the last section.

2.1. CONCEPTIONS OF MEASUREMENT

Stevens (1946) defined measurement in the broadest sense as "the assignment of numerals to objects or events according to rules" (p. 677) and he set out a classification of scales of measurement (p. 678), of which the main features are summarised in Table 2.1.

Table 2.1: Measurement levels from Stevens (1946)

Scale	Empirical operations determine	Preserving transformations x' = f(x)
Nominal	Equality	Permutation group: f(x) is any one-to-one substitution.
Ordinal	As above plus greater or less	Isotonic group: f(x) is any monotonic increasing function.
Interval	As above plus equality of intervals or differences	General linear group: f(x) = ax + b
Ratio	As above plus equality of ratios	Similarity group: f(x) = ax

A key insight in his scheme is the clarification of appropriate arithmetic operations that preserve the meaning of the scales. His classification is not without its critics. One set of objections arises because the classification excludes other conceptual classes that both exhibit mathematical rigour, and are widely used in practice. An

important example of such is the class of “counted fractions” which includes percentages. An alternative taxonomy of classes is presented by Mosteller and Tukey (1977). A more telling criticism is made by Velleman and Wilkinson (1993) who point out that scale types need not be fundamental attributes of the data, but rather may derive from both how the data were measured, and what is concluded from the data. Moreover, Michell (1997) argues that the classification leads researchers to focus on the *instrumental* task of constructing procedures for measurement without giving sufficient attention to the *scientific* task of establishing that the relevant attribute is quantifiable. Despite these criticisms, the classification is widely used in the behavioural sciences, and provides a useful taxonomy for the discussion in this section.

A *nominal* scale is used to categorise data. For example, rather than classifying the gender of a participant as male/female or using a letter code such as M/F, numerals could be assigned such as 1 for male and 2 for female. Such assignment of numerals neither implies that arithmetic is valid on the category labels, nor that there is any natural order; the fact that the numeral 2 is clearly greater than the numeral 1 cannot be taken to imply that female is greater than male.

An *ordinal* scale is used to categorise data for which there is a natural order between the categories. For example, items in an opinion survey could use a Likert scale with ordered categories such as strongly disagree, disagree, neutral, agree, and strongly agree. These could be assigned numerals such as {1, 2, 3, 4, 5} where the order of numerals matches the order of categories. In this case, numeric comparison of the category labels does match the natural order, but arithmetic on the labels remains invalid. Clearly the categories could have been allocated any sequence of increasing numbers so it cannot be concluded that the difference in strength of agreement between neutral and agree (4 - 3) is the same as that between agree, and strongly agree (5 - 4) even though the arithmetic difference between the category labels is 1 in both cases. Such arithmetic (including calculating a mean) is only valid if it is known that the difference in the numeric labels matches the respective difference in the underlying trait.

An (equal) *interval* scale makes the assertion that the difference between any two numerals corresponds to the difference in magnitude of the associated underlying trait. It is linear; a graph plotting scale values against the magnitude of the underlying trait would form a straight line. For example, with a series of observations taken at times {10:00, 10:05, 10:10 and 10:20}, the time could be represented on an interval scale by the numerals {1, 2, 3, 5}. Equally, they could be represented by the numerals {0, 2, 4, 8}; the choice of origin and scale unit is arbitrary. In addition, interval scales can be either discrete or continuous. For the present discussion, the term *continuous* can be interpreted as meaning that between any two numerals there is always another numeral. For example, if a and b are numerals in the scale, then $(a + b)/2$ is also in the scale.

A *ratio* scale has the properties of an interval scale with the addition of a natural zero; the ratio of any two numerals matches the ratio of the magnitudes of the underlying trait.

Interpretation of test scores can be used to illustrate the differences between the last three. If higher scores represent higher abilities, then these scores are on at least an ordinal scale. If, in addition, the difference between a score of 40% and 50% represents the same difference in ability as that between 50% and 60%, then these scores may be on an interval scale. Finally, if in addition a score of 60% represents twice the ability of that associated with a score of 30%, then the scores may be on a ratio scale. In practice, of course, test scores at best form an ordinal scale (Wright & Linacre, 1989).

From a mathematical perspective, the fundamental conception of measurement is a strictly monotonic mapping of members of a set to real numbers. Thus, for a set S of observations, with members s_i and a comparison operation $\rhd$, a function f mapping S to $\Re$ is a measurement function if and only if,

$$s_i \rhd s_j \underset{\forall\, i,j}{\iff} f(s_i) > f(s_j)$$

For example, a scoring function that represents ability is a measurement function if and only if greater ability results in a higher score, and a higher score represents

greater ability. From the mathematical perspective then, a nominal scale cannot be used for measurement because of the lack of a comparison operation. However, ordinal, interval and ratio scales all meet this minimal measurement requirement.

From a scientific perspective, attributes are measurable if they have a quantitative structure; the magnitudes of the quantity can be expressed as real numbers relative to some other magnitude or reference unit. Michell (1997) summarises the requirements for such a quantitative structure (p. 357) and his summary is given in Table 2.2.

Table 2.2: Summary of requirements for quantitative structure

1. Any two magnitudes of the same quantity are either identical or different and, if the latter, there must exist a third magnitude, the difference between them, i.e. for any *a* and *b* in *Q,* one and only one of the following is true
   - a = b,
   - there exists *c* in *Q* such that *a = b + c,*
   - there exists *c* in *Q* such that *b = a + c;*
2. A magnitude entirely composed of two discrete parts is the same regardless of the order of composition, i.e. for any *a* and *b* in *Q, a + b = b + a;*
3. A magnitude which is a part of a part of another magnitude is also a part of that same magnitude, the latter relation being unaffected in any way by the former, i.e. for any *a, b* and *c* in *Q, a + (b + c) = (a + b) + c;*
4. For each pair of different magnitudes of the same quantity there exists another between them, i.e. for any *a* and *b* in *Q* such that *a > b,* there exists *c* in *Q,* such that *a > c > b;* and
5. Given any two sets of magnitudes, an 'upper' set and a 'lower' set, such that each magnitude belongs to either set but none to both and each magnitude in the upper set is greater than any in the lower, there must exist a magnitude no greater than any in the upper set and no less than any in the lower, i.e. every non-empty subset of *Q* that has an upper bound has a least upper bound.

The density requirement (item 4) requires the variable to have continuous rather than discrete values. In general, ordinal scales do not meet this requirement. Both interval and ratio scales are consistent with Michell's conditions if continuous values are allowed.

The term measurement is commonly used both for the process of measuring and for the measurand: the quantity or amount resulting from that process. To avoid confusion, the term *measuring* will be used in this discussion to refer to the process and *measurement* to the quantity resulting from the process. Although an attribute may have a precise conceptual value, measuring inevitably introduces some uncertainty as to the resulting measurement. In the foreword to the guidelines published by the National Institute of Standards and Technology (Taylor & Kuyatt, 1994), the authors note:

> It is generally agreed that the usefulness of measurement results, and thus much of the information that we provide as an institution, is to a large extent determined by the quality of the statements of uncertainty that accompany them. (p. iv)

Measuring should therefore produce both an estimate of the quantity and a statement of the uncertainty associated with that estimate. This is commonly expressed as the standard error of the estimate, equal to the positive square root of the estimated variance.

For a measurement to be useful, both the measurement of an attribute and the process of measuring should exhibit invariance. In what might be termed *process invariance*, the output of the measurement process should not depend on when or where it was measured, or on who carried out the measurement. In what might be termed *magnitude invariance*, the magnitude of the attribute measured should be stable enough that the estimate remains appropriate when applied in the intended context. For example, it should be possible to determine the total weight of a set of objects by weighing each of the objects individually and then summing their weights. This is, of course, an ideal; physics tells us that many useful everyday measurements do in fact vary a little. For example, weights depend in part on height above the earth and the relative position of the moon; the length of a day in seconds varies slightly from one day to the next because of weather; the length of a ruler will vary depending on the tidal force applied by gravity. However, invariance

can be safely assumed in practice, provided such variation is small compared to the normal uncertainty of the estimates.

The psychologist Louis Thurstone identified the need for process invariance in the 1920s:

> A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement (Thurstone, 1928, p. 547).

The principle of magnitude invariance allows everyday measurements to be expressed as some multiple of an internationally agreed standard unit. This is the foundation of measurement in much of physical science, industry, commerce, and everyday life; attributes can be measured independently of their application. This also allows theories of relationships to be constructed and tested. The behavioural sciences, however, presently lack such standard units. In the absence of such units, the development of general theories that relate the findings of different studies requires at least the ability to place those findings on a common metric.

Within the behavioural sciences, the most easily obtainable data arise by observing and analysing the responses R to a set of stimuli S given by a set of participants P. For example, a survey instrument (set of questions S) might be administered to a sample of participants (P) and the responses (R) analysed. Alternatively, in an educational setting, an examination or test (set of items S) might be given to students (P) and the answers (R) scored. The core problem in either case is that, without independent measurements of the attributes of S and P, it is unclear whether effects should be attributable to S or to P.

For example, if a student scores 60% on one test and 70% on a later test, does that mean that the student has increased knowledge, that the later test is easier, or some combination of the two? Perhaps the tests just measure different things? Answering these questions requires some way of equating the tests, thus placing them on a common metric. One way of achieving a common metric is to repeat the measurement holding either S or P constant. Thus, for example, the same set of items could be administered to multiple groups of participants, enabling the responses of those groups to be compared. Similarly, two or more sets of items could be compared by using the same group of participants to respond to each set of items.

There are several possible approaches to test equating. Kolen (1988) outlines some approaches from classical test theory. With a *single group design*, the same participants take both tests in random order at the same time. With a *random groups design*, participants are randomly allocated to groups, each of which takes a single test. With a *common item non-equivalent groups design*, participants can take the tests at different times, but each test shares a common subset of items. However, equating tests with this last approach is problematic under classical test theory; Kolen and Brennan (1987) give a possible approach based on synthetic groups. For each of the designs, tests are equated by calculating a correction formula that translates the scores of one test to the metric of another. Each formula makes a different assumption about the distributions of scores. The simplest, *mean equating*, adjusts each score by the difference in means between the two tests; with *linear equating*, a linear transform is applied to scores of the second test to achieve the same mean and standard deviation as the first; with *equipercentile equating*, the transformation aims to achieve the same mean, standard deviation, skewness and kurtosis. However, despite a great deal of work, equating remains problematic for the classical test theory paradigm; Brennan and Kolen (1987) give an outline of some issues.

In recent years, educational practitioners have progressively moved away from the *norm-referenced* approach of classical test theory, with its distributional assumptions about subjects (P), towards a *criterion-referenced* approach which

emphasises the difficulties of the items in the test (S). The use of Item Response Theory (Lord, 1980) has grown in parallel to this shift. An accessible introduction to IRT can be found in Baker (2001). Cook and Eignor (1991) discuss test equating within an IRT paradigm; essentially this requires only calibration of item parameters for the new test. IRT equating can use the same designs as those discussed above for classical test theory. The main difference is that the *common item non-equivalent groups design* is not problematic; the shared items are simply used as "anchors" to place the two tests on the same metric.

Regardless of the approach, equating or linking tests as described above helps to generalise findings across contexts by using a consistent metric. However, the lack of a standard unit of measurement means interpretation of the measures requires familiarity with at least one of the contexts. One approach that helps with interpretation in a more general context is to publish norms for known groups. Thus, for example, a measure of physical fitness might produce scores that mean little to a wide audience, but publishing average scores for a range of age groups, say, or for a range of professions, would enable meaning to be constructed by those unfamiliar with the measure. An example of this in an educational setting is the idea of *reading age*. The Lexile framework (Stenner, 1997) allocates readability scores to text based on word frequency. Longer sentences give higher scores, as does the use of words that occur less frequently in a database of literary works. The scores have little intrinsic meaning to those unfamiliar with the framework, but Wright and Stenner (1999) show how an ability test can be constructed from selected passages of text, allowing reading ability to be measured on the same metric. It is then possible to administer the test to groups of different ages, professions etc. This allows meaning to be constructed by statements like "*the average Lexile score at age x is y which is associated with reading books like z*"; the meaning comes not from the score itself, but from the association.

Even when tests are equated as discussed above, dependencies may still exist between participants and items, particularly when a scoring approach is used. In attempting to place statements of opinion on a scale, Thurstone pointed out:

> One of the first requirements of a solution is that the scale values of the statements of opinion must be as free as possible, and preferably entirely free, from the actual opinions of individuals or groups. If the scale value of one of the statements should be affected by the opinion of any individual person or group, then it would be impossible to compare the opinion distributions of two groups on the same base (Thurstone, 1928, p. 416) .

He also drew a parallel conclusion for dependency on individual items:

> It should be possible to omit several test questions at different levels of the scale without affecting the individual score. It should not be required to submit every subject to the whole range of the scale. The starting point and the terminal point, being selected by the examiner, should not directly affect the individual score. (Thurstone, 1926, p. 446)

These two ideas might be termed *participant invariance* and *item invariance* respectively. In 1960, Rasch (1960/1980) introduced a measurement model that is now widely known as the *Rasch model*. A major contribution of this model was the concept of objective measurement. When discussing the evolution of this concept Rasch (1977) later chose the term *specific objectivity* to express the idea that, based on observed responses, any two participants, P_n and P_m, should be comparable to each other; the result of such a comparison should be unique; and the result should not depend on the particular subset of items chosen from *S*. Similarly, comparison of two items, S_i and S_j, should be unique and independent of the particular subset of participants chosen from *P*. The concept of specific objectivity can thus be seen to encompass Thurstone's notions of participant and item invariance.

Another concern with measurement is that of validity. Stated simply, a valid measurement instrument measures what it should (Brown, 1996, p. 231). Traditionally, validity was considered to comprise three facets: content validity, criterion validity and construct validity (Brown, 1996, pp. 231-249). *Content validity* focuses on the internal content of an instrument; if an instrument uses a set of items, judgements are made as to whether these items represent effectively the full range of the construct the instrument is intended to measure. *Criterion validity*

focuses on the relationships demonstrated with other external measures; that associations are found if and only if their presence accords with theory. *Construct validity* focuses on demonstrating that the instrument is measuring the construct it claims to measure. It is clear that this last facet subsumes the other two, so a unified view of validity as equivalent to construct validity is possible. Additionally, Cronbach pointed out that "one validates, not a test, but an interpretation of data arising from a specified procedure" (1971, p. 447). This emphasises the meaningfulness of the measurement results as central to validity. The 1985 Standards (AERA, APA, NCME, 1985) echo this view:

> A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. (p. 9)

More recently, there has been a trend to add a fourth element of validity: *consequential validity.* Messick (1989) argued that the social consequences of a test's use should become an important part of the validity framework. He believed that what is needed is:

> … a way of cutting and combining validity evidence that forestalls undue reliance on selected forms of evidence, that highlights the important though subsidiary role of specific content and criterion-related evidence in support of construct validity in testing applications, and that formally brings consideration of value implications and social consequences into the validity framework. (p. 20)

This view has been embraced by a number of practitioners such as Shepard (1997), but strongly challenged by others such as Popham (1997) or Mehrens (2005) who argue that although it raises a legitimate concern, it is not part of validity and only confounds the concept. From a pure measurement perspective, this is clearly true; it is possible to have a valid measurement and measuring instrument even if the subsequent use of the measurement is inappropriate. However, that is not to say

that nothing can be done to promote consequential validity. Appropriate consequential use of the measurements can be supported by: clarity about what is being measured, appropriate statements of uncertainty, systematic attempts to identify weaknesses in the measurement model, and model fit and quality metrics.

Before reviewing the process of measurement, the main points of the above discussion are summarised. Measuring is the process of estimating the magnitude of an attribute of an object, yielding a location on an interval or ratio scale (the measurement), together with a statement of uncertainty of the estimated location, expressed as a standard deviation (the standard error of the estimate). Valid measurement instruments measure what they should and allow meaningful interpretation. In the absence of standard units of measurement, meaning is enhanced by relating measurements to other objects on the scale. The process of measurement should be objective; measurements should not be affected by objects other than those being measured. The next section discusses the process of constructing measurements from observed data.

2.2. MEASUREMENT PROCESS

This section begins by outlining, from a historical perspective, the development of some of the key ideas in measurement. It then presents the need for a theoretical measurement model to connect measurements with observations, and outlines how such a model can be used to produce measurements, statements of uncertainty, and tests of the measurement hypothesis.

2.2.1. Historical Perspective

Quantification of physical attributes has long been believed to be central to the success of physical science in describing and understanding the physical world. Lord Kelvin expressed this sentiment in a lecture given in 1883:

> In physical science a first essential step in the direction of learning any subject is to find principles of numerical reckoning and methods for practicably measuring some quantity connected with it. I often say that when you can measure what you are speaking about and express it in numbers you know something about it;

> but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be. (Kelvin, 1889, p. 73)

Herbart (1877) argued not only that a mathematical approach was possible, but that it was necessary:

> The reason for such necessity is, in a word, that the aim and end of all speculation is otherwise absolutely beyond our reach, and that aim and end is: mathematical certainty. This necessity to establish our theories definitely is the more urgent, the greater the danger is that philosophy may soon relapse into the state in which it is already in France and England. It is a manifest blindness of most of the living German philosophers that they do not see this danger. If they knew mathematics, (and I mean a little more than the elements of geometry, or quadratic equations, or the signs of the differential and integral calculus) -- if they understood mathematics, they would know that an indefinite talk, interpreted differently by each individual and which only multiplies the disputes, cannot possibly -- notwithstanding the beauty of presentation and the sublimity of the subject matter -- keep abreast of a science which instructs and elevates by every proposition uttered, and which elicits never-ending admiration -- not for the vast spaces it has measured, but for the exhibition of the most stupendous human sagacity. Mathematics is the ruling science of our time; its acquisitions grow daily, though noiselessly. He who does not befriend it, will have it his enemy in the future. (p. 262)

Measurement of psychological constructs could thus be seen as contributing to the goal of elevating psychology to the status of a science. However, from times of antiquity, the corporeal world, the physical world, the world of matter, the real has been seen as fundamentally different from the world of the mind, the incorporeal spiritual world, the perceived world. This *mind-body problem* was first articulated in modern form in the writings of René Descartes in the 17th century. In meditation II of his work *Meditations on First Philosophy* (1641)**,** Descartes sets out the

proposition that our *interior* mental world is populated by ideas (memories, images, perceptions, beliefs etc.) that are separate from the things they represent in the *exterior* world, and may be accurate or false representations. In the Middle Ages, most scholars approached the issue from a theological perspective and held the Aristotelian belief that "Quantity does not, it appears, admit of variation" (Michell, 2003, p. 519). Discussing the *Sentences* of Peter Lombard, a text widely read by medieval scholars, Crombie commented that Lombard had put the question "whether the theological virtue of charity could increase and decrease in an individual and be more or less intense at different times" (1994, p. 410). Most scholars held the view that "There could be no addition or subtraction of degrees of intensity of a quality as there could be of length or a number" (1994, p. 414).

However, in the early 14th century, the Franciscan John Duns Scotus had proposed that a change in degree of a qualitative attribute could be understood as the addition or subtraction of homogeneous parts of the quality (Michell, 2003, p. 520). Despite this, the notion that measurement could apply only to the extensive, exterior world, and not to the intensive interior world, persisted at least until the 19th century, and perhaps still endures today. Against this back-drop, a number of scholars nevertheless pursued the goal of objective measurement of constructs in the interior mental world.

Like Descartes, Kant (1781/1998) made the distinction between things in themselves, and things in appearance (perception). However, anticipating the future field of *psychophysics*, he suggested that measurement was possible if one regarded reality as the cause of the sensation or perception, and if appropriate measurement techniques, which he termed *mathesis intensorum*, were developed. He noted:

> All sensations without exception have degrees, and thus what is real in all appearances has degrees. This is the second application of mathematics (*mathesis intensorum*) to natural science (Kant, 1783/2004, p. 67)

However, Kant did not pursue his *mathesis intensorum*, leaving it to empirical psychologists to gather measurements of intensive magnitudes of both the real and resulting sensations (Baumann, 2008).

Part of the perceived difference between measurement in the physical world and measurement in the mental world may be traced to a lack of clear understanding of what is meant by measurement in the physical world. The essential requirement for measurement is that the variable being measured has a quantitative structure (Michell, 1997). Hölder (1901/1996) was one of the first scholars to consider physical measurement from an axiomatic perspective, pointing out that quantitative structure was commonly simply assumed in physical measurement, rather than established formally. It is clear from his work that parallel considerations apply to both the physical and mental worlds, and that there was an opportunity to build on these ideas for psychophysical analysis. However, further development of this approach was not progressed until the second half of the 20th century.

A number of scholars attempted to study the relationship between real world measurements and perceived magnitudes. Weber (1834/1978) carried out a series of experiments designed to identify the *just noticeable difference* in the perceived magnitude of some physical quantities, such as weight or light intensity. He found that in many cases, the additional stimulus ΔI required to produce a just noticeable difference was approximately a constant fraction of the total stimulus I, over a wide operating range. This can be expressed in what is now known as *Weber's law*:

$$\Delta I = \mathrm{K}_w I \tag{2.1}$$

The constant K_w is known as the Weber Fraction. Fechner introduced the term *psychophysics* to refer to the "exact science of the functional or dependency relations between body and mind, or more general: between the bodily and the spiritual, physical and psychical, world" (1860/1966, p. 8). He reasoned that:

> It follows that each single value, or each definite fraction or each definite multiple of the magnitudes that have been found equal (no matter which), can

be taken as the unit according to which the total magnitude, or every fraction of it, can be measured. The *n* equal parts that can be thought of as composing a total magnitude of course have the same magnitude as the *n* equal parts into which the total magnitude can be thought to be decomposable. All physical measurement is based on this principle. All mental measurement will also have to be based on it (1887/1997, p. 213).

Building on Weber's work, Fechner reasoned that the just noticeable difference could form a measurement unit, and presented a generalisation of Weber's law that is known as the fundamental equation of psychophysics:

$$E = k \log(I) \tag{2.2}$$

Here, *E* is the perceived intensity of the stimulus I and *k* is a constant determining the units of the perceived intensity scale. Defining I_0 as the largest intensity that was just not noticeable, he deduced what he called the fundamental *measurement formula* (1887/1997, p. 219):

$$E = k \log\left(\frac{I}{I_0}\right) \tag{2.3}$$

Stevens investigated empirical relationships between stimuli and responses for a large range of stimuli. He proposed a power law (Stevens, 1957, p. 162) to encapsulate the relationship:

$$\psi = k\, I^n \tag{2.4}$$

Here, *k* represents an arbitrary scale constant and the exponent *n* depends on the particular stimulus investigated. He determined approximate exponents for a large range of continua from loudness and brightness (with exponents around 0.3) through to visual velocity and visual flash rates (with exponents 1.77 and 2.0, respectively). Despite arguing for the form of a power law rather than Fechner's logarithmic form, he notes that:

On continua that behave like Class I we would be closer to right if we began with the assumption that discriminal dispersion is not constant but is proportional to

the psychological magnitude in question. When the psychological magnitude is a power function of the stimulus, this assumption is equivalent to saying that psychological values separated by equal units of dispersion on the stimulus scale stand in a constant ratio to each other. (Stevens, 1957, p. 175)

Applying this insight to Fechner's equation (2.3) would yield the following form:

$$\log_e(\acute{E}) = \acute{k} \log_e\left(\frac{I}{I_0}\right) \tag{2.5}$$

However, Stephens does not develop this. He discusses the development of a logarithmic scale, based on equal ratios and similar to that used in modern measurement models, before arbitrarily dismissing the utility of such an approach:

A scale of this kind may be mathematically interesting, but, like many mathematical models, it has thus far proved empirically useless. (Stevens, 1957, p. 176)

Nevertheless, if one assumes that the number continuum used by Steven's subjects to rate perceived values is just another kind of stimulus continuum, and this perceived number dimension is represented as ψ, the modified form of Fechner's equation (2.5, p. 26) can then be written as:

$$\log_e(\psi) = \acute{k} \log_e\left(\frac{I}{I_0}\right) \tag{2.6}$$

Substituting, λ for ${I_0}^{-\acute{k}}$, this can be written in the form of Steven's power law:

$$\psi = \lambda I^{\acute{k}} \tag{2.7}$$

Thus, it can be seen that incorporating the notion that subjects make judgements about which numbers to use, as well as about the intensity of a stimulus, unifies Steven's power scale with the modified form of Fechner's logarithmic scale. A formal treatment of the unification of Steven's and Fechner's approaches is given by Kreuger (1989). Other models for the relationship between the real and the perceived have been proposed, but a logarithmic treatment is generally found to unify the models. For example, Iverson notes that "The Weber–Fechner logarithmic

solution u1 is thus common to the Psychophysical Power Law and Falmagne's law" (2006, p. 286). For the present work, however, the precise formulation of the relationship between stimuli and responses is not required. It is sufficient to note that a logarithmic transform offers a parsimonious and effective way of visualising a large physical rage of values. Indeed, Stephens uses a logarithmic transformation in his own publication (1957, p. 167) to relate magnitude estimation to loudness in decibels. Graphically, his power law becomes a straight line in these units. It is tempting to speculate that since logarithmic transformations offer such power in visualisation, the mind might actually work in this way.

Thurstone (1927a), proposed a new point of view in psychophysical analysis in which he conceptualised judgement as a *discriminal* process in the psychological dimension, with the process characterised by a normal distribution. He derived a fundamental psychophysical equation and set out experimental procedures that could be used to verify the equation. He then used the equation to demonstrate that Weber's law and Fechner's law are not identical. He later named his equation a Law of Comparative Judgement (1927b, p. 267). Two key ideas are introduced in his work. Firstly, the approach is not limited to those stimuli, such as weights or light intensity, which can be readily measured in the physical world; the approach can also be applied to more abstract ideas such as beauty or legibility. Secondly, no assumptions are made about the distribution of the values of what is being measured, just the distribution of errors in the measurement process. This is in accordance with everyday notions of measurement. For example, if one uses a ruler to measure lengths of timber, one would expect to be able to measure a single piece of timber rather than requiring concurrent measurement of a number of pieces of timber whose lengths form a normal distribution. Equally, it seems plausible that measurement errors might be distributed normally, since the true or intended length gives a natural mechanism for central tendency, and measurement errors could be the sum of a number of independent causes.

Guttman (1944) described a process of constructing a scale from ordered dichotomous judgements. He noted that "It is because all the items in the sample can be expressed as simple functions of the same ordering of persons that they

form a scale" (p. 145). He also articulated the connection with measurement: "Finding that a universe of attributes is scalable for a population means that it is possible to derive a quantitative variable from the multivariate distribution such that each attribute is a simple function of that variable" (p. 148). Guttman also recognised that scores could be a sufficient statistic: "The attributes are the important things; and if they are scalable, then the scores are merely a compact framework with which to represent them." (p. 149). He reasoned that individual responses could then be reproduced from the summary score and defined a *coefficient of reproducibility* to represent the proportion of responses for which this was true. He suggested that values of 85% or better were effective. His approach was deterministic and empirical, rather than inferential. In a later paper (1947), he described a procedure in which scores were based on weights associated with categories, and the weights were established by a process of progressive refinement with the goal of maximising the coefficient of reproducibility.

Until the 1960s the dominant approach remained empirical, as typified by Steven's work mentioned above. Rasch (1960/1980) presented a stochastic model for objective measurement which introduced a new perspective. Rather than trying to fit candidate models to data, he introduced a pure measurement model, and then investigated the fit of data to the model. From this perspective, when data do not fit the model then, rather than looking for another model, the reasons for misfit are investigated and theory and understandings are adjusted accordingly.

In their seminal work on simultaneous conjoint measurement Luce and Tukey (1964) give a formal treatment of the conditions under which ordinal observations can be transformed to measurements on interval or ratio scales. Based on this, and taking the notion of specific objectivity as a starting point, Fischer (1987) formally demonstrated not only the existence of models (his theorem 1), but that, assuming the observations are locally independent realizations of Bernoulli variables, the only models that can satisfy specific objectivity are isomorphic to a logistic model with additive parameters, thus determining an interval scale for latent trait measurement (his theorem 4). The Rasch model has this form. Irtel (1995), however, shows that there is an implicit assumption of real-valued measurements

in Fischer's work, and that a more abstract definition of measurement and comparison would allow other models to demonstrate specific objectivity. Nevertheless, the goal in this project is to produce real-valued measurements, so for reasons of parsimony, the natural choice is to use a model that is isomorphic to the logistic model derived by Fischer.

The foregoing discussion is now briefly summarised. The motivation for pursuing objective measurement is that measurement of psychological constructs may contribute to the goal of elevating psychology to the status of a science by enabling the development of refutable hypotheses, and consequent theory building. The essential requirement for measurement is that the variable being measured has values on a continuum with a quantitative structure. Luce and Tukey (1964) presented the formal conditions for such a structure. From an epistemological perspective, measurement must be grounded in the world of the mind, the abstract, the ideal. However, if one accepts, as Kant proposed, that reality may be the cause of a perception then it is possible to make inferences about the real from the measurement. This is how measurement works in the physical sciences and is the foundation of mathematics.

Guttman has demonstrated how a scale can be formed from dichotomous judgements and Thurstone's work has shown how objective measurement can be built on such comparative judgements. His work also allows a distinction to be made between measurement of attributes of individual objects and more general quantitative techniques that are based on the attributes of populations of objects. This allows measurement to be separated from distributional assumptions about the object attributes being measured. Rasch presented a practical model for objective measurement, and Fischer demonstrated that objective measurement which produces real numbered measurements requires a logistic form isomorphic to the Rasch model. The natural unit of such measurement is the log of a ratio, and the empirical work in psychophysics by Weber, Fechner, Stevens and others suggests that the connection between the real and the mind may have such a form.

2.2.2. Role of a Model

Source data in the behavioural sciences often involve observations classified on an ordinal scale. However, measurement, as discussed earlier, requires estimates to be placed on an interval or ratio scale (Wright & Linacre, 1989). The process by which the necessary translation can be achieved requires a theoretical model that connects measurements, which are abstract, to observations, which may be in the world of the mind, in the world of the real, or assumed to be caused by the world of the real. The theoretical model will use parameters (which are the measurements to be imputed) to predict, in probabilistic terms, the observations that should ensue. Estimation of the model parameters proceeds by searching the parameter space to identify the combination of parameters that gives the "best" fit with observed data; what is meant by best fit is defined below. Accordance of the predicted observations with those actually observed thus evidences and supports the theoretical model. Similarly, misfit of the observations disconfirms the model. This falsifiability is an essential feature of the scientific method (Popper, 1935/2005), and allows the measurement hypothesis to be tested.

Best fit can be defined in several ways, leading to different estimation techniques. For example, least-squares estimation (Maxwell, 1974) seeks to minimise the sum of squared differences between model predictions and the observed data. One of the most flexible techniques, and that used in this project, is the procedure of maximum likelihood estimation. This technique was developed by R.A. Fisher between 1912 and 1922. The essence of Fischer's method is presented here; a fuller summary of the technique as it evolved is given in Aldrich (1997).

Let $\{x_1, \dots, x_n\}$ be a set of independent, identically distributed, observations with density specified by a function $f(\mathrm{x}, \theta)$ controlled by a parameter θ, which may be a vector. Then the joint probability of the observations is:

$$f(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \qquad (2.8)$$

Rather than viewing the probability distribution as dependent on the parameter θ, it is possible to view the parameter as the dependent variable. The likelihood $\mathcal{L}$ of the parameter θ, given the data, can be defined as.

$$\mathcal{L}(\theta|x_1, \dots, x_n) \stackrel{\text{def}}{=} \prod_{i=1}^{n} f(x_i|\theta) \tag{2.9}$$

The maximum likelihood estimator of the parameter θ is the value of θ that maximises this likelihood. In practice, the logarithm of the likelihood is usually maximised since a logarithm is a monotone transformation, and will thus have a maximum at the same value of θ. This transformation enables the continued product to be replaced by a summation.

$$\lambda(\theta|x_1, \dots, x_n) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \log\big(f(x_i|\theta)\big) \tag{2.10}$$

Where θ is a vector, searching the parameter space is also much more efficient if the function $f(x|\theta)$ can be factored into separate parts for each component of θ:

$$f(x_i|\theta) = g(x_i|\theta_1)h(x_i|\theta_2) \tag{2.11}$$

Such separation enables each component to be maximised separately. In this project, a variant of the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is used. This proceeds by alternatively producing maximum likelihood estimates of the first component (expectation phase) and the second component (maximisation phase), progressively refining both until convergence is achieved. This approach is also referred to variously in the literature as unconditional maximum likelihood estimation (UMLE or UCON) or joint maximum likelihood estimation (JMLE).

Essentially, the maximum likelihood estimation procedure discussed above treats the model parameters as dependent variables and the observations as independent variables in order to estimate the parameters. Once the estimates have been made, it is possible to take the opposite point of view and regard the observations as dependent variables.

The model, with its estimated parameters, will predict a probability distribution for the observations. The concordance of observations with these predictions is evidence of model fit. A formal treatment of model fit will be given in the next chapter once the concrete model has been introduced. For the present, however, it can be noted that the process of measurement provides a context in which model fit can be investigated. Each measurement is thus a test of the theory underpinning the construction of the theoretical model.

It can also be noted that since the model produces a predicted probability distribution rather than just point estimates, a standard error of measurement can be produced naturally for each estimate. The measuring process thus achieves the goal of estimating the magnitude of an attribute on an interval scale, together with a statement of uncertainty of the estimated magnitude, expressed as the standard error of the estimate. Taken together, model fit diagnostics and estimates of uncertainty provide for quality control of the measurements produced.

The general operation of a measurement model can be summarised as follows. The model specifies in probabilistic terms how observations should occur based on model parameters. A maximum likelihood estimation procedure is used to identify the parameters that give the best fit with observed data. Once these estimates have been produced, it is possible to ask how well the observations fit the predictions of the model, given the estimated parameters. If the fit is acceptable, the model parameters form the output measurements. If the fit is unacceptable, the data are deemed not to fit the model, and no output measurements are produced. These “goodness of fit” tests and statements of uncertainty are thus essential parts of the measurement process.

2.3. ABSTRACT MEASUREMENT MODEL

This section introduces a conceptual model that can be used to infer measurements on an interval or ratio scale from ordinal judgements. The logic of the measurement model is presented here in its most abstract form. It is essentially an abstract form of the logistic model. A more concrete model is presented subsequently in section 2.4 below.

Measurement on a ratio scale will be discussed first. The output measurement will be a value of a variable that represents magnitudes of an attribute. Magnitudes can be represented by allocating them to positions on an imaginary line. Figure 2.1 shows a possible visualisation of such a variable with allocated magnitudes A, B, C.

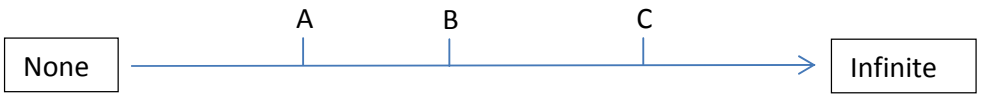


Figure 2.1: A conceptual variable on a ratio scale

There may be many observations in a dataset; those relevant to the model are those that give evidence of relative magnitudes. Let L_A and L_B be the locations on the scale of the magnitudes of the chosen attribute of two objects A and B. Let E_A be an observable event that gives evidence that $L_A > L_B$; let E_B be an observable event that gives evidence that $L_B > L_A$; and $OR(a:b)$ be the odds ratio in which the events E_A and E_B are predicted to occur. The measurement model starts with the proposition that:

$$OR(a:b) = \frac{a}{b} = \frac{L_A}{L_B} \tag{2.12}$$

The odds ratio $OR(a:b)$ can be estimated by observing the actual frequencies in which the events E_A and E_B occur in a dataset. If N_A denotes the count of events E_A and N_B denotes the count of events E_B then the relative locations L_A and L_B can be estimated as:

$$\frac{L_A}{L_B} \cong \frac{N_A}{N_B} \tag{2.13}$$

This process can be repeated for additional pairs of objects in the set, enabling the relative locations of each member of the set to be established. This model, if it holds, has some remarkable properties which will be explored below. Before doing so, it is important first of all to stress the caveat “if it holds”. As with any model it is important to verify how well the model fits the data and how this is done is discussed in detail in the next chapter. For the present, it is sufficient to note that one of the key strengths of the model is that it is readily falsifiable.

2.3.1. Comparing Items in a Test

To illustrate the model, consider a test that comprises a set of items indexed by *i*. Let L_i be the difficulty of item i; let E_{ij} be a student who gave a correct answer to item i and an incorrect answer to item j; let N_{ij} be a count of such students, or even an arbitrary subset of such students. Under the model, the relative difficulty of any two items can be estimated as:

$$\frac{L_i}{L_j} \cong \frac{N_{ji}}{N_{ij}} \tag{2.14}$$

Proceeding in this way, the relative difficulty of all items can be established from these counts. It is noteworthy that the model does not require knowledge of the ability of any students, nor does it require that all student responses are included. In particular, the model produces similar estimates whether a subset comprises less able students, or more able students. This characteristic is termed *person-free item comparison* (Wright & Stone, 1999, p. 25).

2.3.2. Comparing Students in a Test

A similar set of calculations can be carried out for estimates of student ability. Let S_i be the ability of student i; let E_{ij} be an item correctly answered by student i and incorrectly answered by student j; let N_{ij} be a count of such items, or an arbitrary subset of such items. Under the model, the relative ability of any two students can thus be estimated as:

$$\frac{S_i}{S_j} \cong \frac{N_{ij}}{N_{ji}} \tag{2.15}$$

Proceeding in this way, the relative ability of all students can be established from these counts. Again, it is noteworthy that the model does not require knowledge of the difficulty of any items, nor does it require that all student item responses are included. In particular, the model produces the same estimates whether a subset comprises easier items, or more difficult items. This characteristic is termed *test-free person measurement* (Wright & Stone, 1999, p. 26).

2.3.3. Connecting Students and Items

In the example given above, student abilities and item difficulties can be placed on a common metric if a definition of equivalence between a subject (S_x) and a notional item (I_0) can be given:

$$\frac{S_x}{I_0} \stackrel{\text{def}}{=} k \tag{2.16}$$

Here, k is the odds ratio of success that is expected when the subject ability and item difficulty are deemed equivalent. Thurstone suggested (1927a, p. 384) that a possible level of 75% correct judgments could be used to define equivalence. This would lead to an odds ratio of 3:1 correct to incorrect judgements. However, this is a subjective definition. A more natural definition from an information-theoretic perspective is to use an odds ratio of 1:1, and hence k=1. This is because the information that an observation can give about the value of the parameters S_n and I_j is at a maximum when $S_n = I_j$. With this definition, a student who encounters an item that has a difficulty of the same level as his or her ability is equally likely to get the answer right as to get it wrong.

The use of a common metric allows meaning to be constructed for the abstract measurements by relating student abilities to equivalent item difficulties.

2.3.4. Evaluation

Taken together, person-free item calibration and item-free person calibration comprise what Rasch termed “specific objectivity”. They also address Thurstone’s requirements for person and item invariance.

In the example of students and items described above, it might be expected that the estimates of relative abilities and difficulties would be very rough approximations unless the counts N_{ij} were large, leading to a requirement for tests with large numbers of items and large student numbers. However, there are many permutations of pair-wise comparisons that can be made, each of which is an

independent estimate of a relative value. Aggregating these estimates can produce a more refined estimate and a narrower margin of error.

For example, if there are *i* items and if there are existing rough estimates for each of these, a more refined estimate can be produced for any item *j* by calculating a set of estimated relative difficulties by pair-wise comparison with the other (*i – 1*) items (i.e. $i \neq j$) and averaging these. Successive iterations of this process would then produce progressively refined estimates. In a similar manner, pair-wise comparison of student ability estimates with the other (*n – 1*) students would enable student ability estimates to be refined.

2.3.5. Using an Interval Scale

Before proceeding to develop these insights more formally, two modifications to the model presented above will be made: changing from odds-ratios to probabilities, and changing to an interval scale. The first modification will be to change from odds-ratios to probabilities. Writing P_A for the probability that E_A occurs rather than E_B, or equivalently the probability of observing events supporting the proposition that $L_A > L_B$, leads to:

$$Pr(L_A > L_B) = P_A = \frac{a}{a+b} = \frac{1}{1+\frac{b}{a}} \tag{2.17}$$

The following identities can also be noted:

$$\frac{P_A}{P_B} = \frac{a/(a+b)}{b/(a+b)} = \frac{a}{b} = \frac{P_A}{1-P_A} \tag{2.18}$$

Substituting the identities from equation 2.12 (p. 33) in equation 2.17, the probability of observing events giving evidence that $L_A > L_B$ can be stated as:

$$Pr(L_A > L_B) = P_A = \frac{1}{1+\frac{L_B}{L_A}} \tag{2.19}$$

The relative estimates in the model described above are all expressed as ratios. The structure is essentially multiplicative so that, for example, averaging estimates

requires using the geometric mean. Although this is valid, it is generally more convenient to use a model with an additive structure. The second modification is therefore to change the model to produce linear estimates rather than ratios, thus producing estimates on an interval scale. A visualisation of such a scale is given in Figure 2.2.

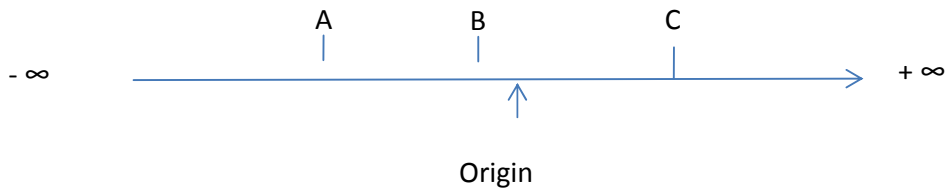


Figure 2.2: A conceptual variable on an interval scale

The location of the origin is arbitrary and can be set to any convenient location. Changing from a ratio scale to an interval scale can be achieved by using the logarithm of a ratio scale location. A logarithm to any base could be chosen. In practice, it is convenient to work with the natural logarithm, but since the choice of base is arbitrary, a scale constant α will be introduced for generality. Using the notation M_X for the location of L_X in this transformed scale, the relationship between the scales can then be written as:

$$M_X = log_e(L_X)/\alpha \tag{2.20}$$

$$L_X = e^{\alpha(M_X)} \tag{2.21}$$

Substituting these in equation 2.19 (p. 36) enables the comparison model to be expressed in terms of M_X.

$$Pr(L_A > L_B) = \frac{1}{1 + \frac{L_B}{L_A}} = \frac{1}{1 + \frac{e^{\alpha M_B}}{e^{\alpha M_A}}} \tag{2.22}$$

$$= \frac{1}{1 + e^{\alpha(M_B - M_A)}} \tag{2.23}$$

It is convenient to conceptualise the term in the exponent as $M_A - M_B$. Making this substitution, the abstract reference model for this work can now be presented:

$$Pr(L_A > L_B) = \frac{1}{1 + e^{-\alpha(M_A - M_B)}} \tag{2.24}$$

This model expresses the probability of observing events as dependent on the difference in linear measurement. Alternatively, the difference in measurement can be treated as the dependent variable. By rearrangement of this equation, the difference in measures M_X can be expressed in terms of the odds ratios.

$$\frac{a}{a+b} = \frac{1}{1 + e^{-\alpha(M_A - M_B)}} \tag{2.25}$$

$$1 + e^{-\alpha(M_A - M_B)} = \frac{a+b}{a} \tag{2.26}$$

$$e^{-\alpha(M_A - M_B)} = \frac{a+b}{a} - 1 = \frac{b}{a} \tag{2.27}$$

$$-\alpha(M_A - M_B) = \log_e\left(\frac{b}{a}\right) \tag{2.28}$$

$$M_A - M_B = -\log_e\left(\frac{b}{a}\right)/\alpha = \log_e\left(\frac{a}{b}\right)/\alpha \tag{2.29}$$

Since the logarithmic transformation is monotonic, the probability that $L_A > L_B$ must be that same as the probability that $M_A > M_B$. The abstract probabilistic model for comparison of two objects can therefore be stated as:

$$Pr(L_A > L_B) = Pr(M_A > M_B) = \frac{1}{1 + e^{-\alpha(M_A - M_B)}} \tag{2.30}$$

2.3.6. Scale Transformations

To explore the properties of the model further, it is illustrative to consider the effect of Stevens' (1946) allowable scale transformations on the model described so far. As shown in Table 2.1, multiplication by a constant c will preserve meaning for a ratio scale such as that used for the locations L_X. Thus, the transformation f(x) = cx will also produce a ratio scale. The effect of this can be explored by replacing L_X with cL_X in the equations given above. In these equations, the locations always appear as a ratio L_A/L_B. It can be noted that:

$$\frac{L_i}{L_j} = \frac{cL_i}{cL_j} \tag{2.31}$$

Thus the same relative estimates will be produced whatever value of c is chosen. Intuitively, this makes sense. Transformation by a constant c transforms the ratio scale from one measurement unit to another. Without knowledge of an absolute scale measure, the best that can be done is to estimate the relative locations of the magnitudes. Conversely, if a magnitude of some location L_X could be obtained externally, then the scale could be calibrated by "anchoring" this location and calculating all other locations relative to this.

The measures M_X form an interval scale. Transformations that preserve meaning for an interval scale take the form f(x) = ax + b (see Table 2.1). Applying this to the relationship between M_X and L_X given in equation 2.21 (p. 37),

$$L_X = e^{\alpha(aM_X+b)} = e^{\alpha b}e^{\alpha aM_X} = \mathrm{c}e^{\alpha a(M_X)} \tag{2.32}$$

From this, it can be seen that the transformation parameter **a** changes the value of α to $\alpha\mathbf{a}$. However, the parameter α is an arbitrary parameter in the model representing the scale unit. Consequently, the parameter **a** in the transformation changes the scale unit. The transformation parameter **b** has the same role as c in the underlying ratio scale; it is equivalent to setting c to $e^{\alpha b}$. The effect is thus to change the unit of the underlying ratio scale or, equivalently, the origin of the interval scale. From this, it can be seen that the linear metric on which M_X is measured has a fixed unit, determined by the parameter α and an origin which functions as a calibration constant and sets the unit of the underlying ratio scale.

2.3.7. Summary of the Abstract Model

A summary is now given of what has been set out so far. The abstract model takes, as input, observable events E_X that are based on comparative judgements; each event gives evidence that the magnitude of some scale location is greater than another or less than another. From these inputs, the model constructs estimates of the locations of the magnitudes M_Y on an interval scale, or equivalently L_Y on a ratio scale. Locations are fully determined apart from:

1) The scale origin of the interval scale, or equivalently the unit of the underlying ratio scale, which must be established by external calibration and
2) The arbitrary scale unit α that is used for the interval scale.

The model exhibits specific objectivity and is of the logistic form required by Fischer (1987). Although the model has been illustrated by reference to the familiar context of students and tests, it is a generic model and can be used to measure any attributes of any objects. Any dataset in which appropriate comparative judgement events are present can be used. Thus, it applies equally to measurement of the physical world and to the measurement of mental constructs.

2.4. OTHER MEASUREMENT MODELS

The abstract model presented above provides an outline of the logic required to construct measurements from arbitrary object comparisons. This section reviews a number of existing measurement models. To provide a consistent context for the review, some constraints and notation will be applied. Firstly, measurements will be based on the responses of subjects to items. Subjects will have an *ability*, which will be denoted as θ_n where *n* indexes the subjects. Items will have a *difficulty*, denoted as β_i, where *i* indexes the items. Measurements of θ_n and β_i will be placed on the same interval scale. The probability that a subject with ability θ_n will be respond in a category above the difficulty level β_i will be termed P_{ni}. This is equivalent to the probability that θ_n will be judged greater than β_i. Making these substitutions in equation 2.29 (p. 38) gives:

$$P_{ni} = \frac{1}{1 + e^{-\alpha(\theta_n - \beta_i)}} \tag{2.33}$$

This will be used as the reference model for the review in this section. It should be noted that the use of the terms ability and difficulty above is consistent with the literature, but other terms, such as *strength of endorsement* could be substituted, depending on the meaning of the particular scale used.

By appealing to equation 2.29 (p. 38), equation 2.33 can be stated equivalently as:

$$\theta_n - \beta_i = log_e\left(\frac{a}{b}\right)\Big/\alpha = log_e\left(\frac{P_{ni}}{1-P_{ni}}\right)/\alpha \qquad (2.34)$$

or $$\propto (\theta_n - \beta_i) = log_e\left(\frac{P_{ni}}{1-P_{ni}}\right)$$

An important characteristic of this model is parameter separation. Given the linear nature of the left side of equation 2.34, item difficulties can be eliminated from the estimation of relative subject abilities, and subject abilities from the estimation of relative item difficulties, thus establishing specific objectivity.

Another characteristic of the reference model presented here is that a count of the number of thresholds passed is a sufficient statistic. The term *sufficient statistic* was coined by R.A. Fisher (1922) to characterise a statistic that exhausts the information in a dataset about a parameter. Formally, a statistic *T* is said to be sufficient for the parameter ϑ if the distribution of the data conditional on the value of the statistic *T* does not depend on ϑ. Informally, knowing the sufficient statistic tells us all that we can know about the parameter from the dataset; it is just as informative as the full data. In the model presented here, all thresholds are equal in weight; intuitively, if all the objects contributing are equal in some attribute then counting them gives all the information available. The use of scores as a sufficient statistic enables easier calibration of model parameters. This was important when these models were first introduced. In the present day, however, commodity computers are several orders of magnitude more powerful than the computers of that era, and parameter estimation poses few practical challenges. From a measurement perspective, there is no empirical reason why each threshold should have equal weight if these weights can be established a-priori, or by an estimation procedure. On the other hand, there is no compelling a-priori reason to use weighted thresholds rather than the natural scores associated with counting thresholds.

2.4.1. Relationship to the Naïve Bayes Classifier

Comparison with the *Naïve Bayes Classifier* will clarify some of the underlying assumptions of the model. From conditional probability definitions, the conditional

probability of observed evidence *E*, given a classification *C,* can be expressed in terms of the joint probability $E \cap C$ as:

$$P(E|C) = \frac{P(E \cap C)}{P(C)} \tag{2.35}$$

Thus $$P(E \cap C) = P(E|C)P(C)$$

The probability of a classification, given the evidence, can therefore be written as:

$$P(C|E) = \frac{P(E \cap C)}{P(E)} = \frac{P(E|C)P(C)}{P(E)} \tag{2.36}$$

This is one form of *Bayes theorem*. If the evidence comprises a number of features E_x then the joint probability can be written in terms of these features:

$$P(E \cap C) = P(E_1, \dots, E_n) \tag{2.37}$$

For Bayes classifiers, the term *naïve* is used to highlight the assumption of conditional independence. Under this assumption the joint probability can be expressed as a continuing product. Thus, where $f_x(.)$ encapsulates the probability distribution of E_x,

$$P(E_1, \dots, E_n) = \prod_{1}^{n} f_n(E_n) \tag{2.38}$$

For the dichotomous classification, Y = 1 for an upper category and Y = 0 for a lower category, then by reference to equation 2.35 (p. 42), the odds ratio can be expressed as:

$$\frac{Pr\ (Y = 1|E)}{Pr\ (Y = 0|E)} = \frac{Pr\ (Y = 1)}{Pr\ (Y = 0)} \prod_{x=1}^{n} \frac{Pr\ (E_x|Y = 1)}{Pr(E_x|Y = 0)} \tag{2.39}$$

Introducing the notation of the reference model, and using C_Y to denote the prior probability this can be rewritten as follows:

$$C_Y \stackrel{\text{def}}{=} \frac{Pr\ (Y = 1)}{Pr\ (Y = 0)} \tag{2.40}$$

$$f_x(E_x) \stackrel{\text{def}}{=} \frac{Pr\,(E_x|Y=1)}{Pr(E_x|Y=0)} \tag{2.41}$$

$$\frac{P_{ni}}{1-P_{ni}} = C_Y \prod_{x=1}^{n} f_x(E_x) \tag{2.42}$$

$$log_e\left(\frac{P_{ni}}{1-P_{ni}}\right) = log_e(C_Y) + \sum_{x=1}^{n} log_e\big(f_x(E_x)\big) \tag{2.43}$$

It can be seen that the additive structure on the right hand side of equation 2.43 is a consequence of the assumption of conditional independence; it is important to note that no other assumptions about probability distributions have been made. It can also be noted that the terms take the form of a log of an odds ratio. Comparing equation 2.43 to the reference model in equation 2.33 (p. 40), it can be seen that the reference model uses only two features: E_1 for subject ability and E_2 for item difficulty. The derivation above suggests that the reference model could be readily extended to accommodate additional features, provided these are in some sense orthogonal to each other so as to preserve conditional independence. Indeed this has been done; for example, Linacre (2002) has developed the *Facets* computer program to support such multi-facetted analysis.

For the present work, however, only two facets are needed: ability of subjects and difficulty of items. Using the notation $\check{C}_Y = log_e(C_Y)/\alpha$ and $\check{f}_x(.) = log_e\big(f_x(.)\big)/\alpha$, equation 2.43 (p. 43) can be re-written as:

$$\alpha\left(\check{C}_Y + \check{f}_1(E_1) + \check{f}_2(E_2)\right) = log_e\left(\frac{P_{ni}}{1-P_{ni}}\right) \tag{2.44}$$

In this form, α can be seen to identify the unit of the metric; E_1 represents the unknown subject abilities, of unknown distribution; and likewise E_2 represents the unknown item difficulties, also of unknown distribution. The functions $\check{f}_x(E_x)$ can be seen to linearize the unknowns E_X, bringing them into the common metric with the unit defined by α. A key insight is that there is no need to identify the underlying unknowns, nor the transforming functions that encapsulate their probability distributions, since the linear forms resulting from the transformations can be

estimated directly, and are more useful for measurement. Substituting the notation of the reference model, equation 2.44 can thus be rewritten as:

$$\alpha\left(\check{C}_Y + \theta_n + \beta_i\right) = log_e\left(\frac{P_{ni}}{1 - P_{ni}}\right) \tag{2.45}$$

In this form, β_i would represent item *facility*; changing the sign to negative would allow it to represent the more conventional item *difficulty*. The prior $\check{C}_Y$ expresses the relationship between the subject abilities θ_n and item difficulties β_i. If an item difficulty is defined as the magnitude at which a subject of equivalent ability is equally likely to succeed on the item as to fail, then, from equation 2.40 (p. 42), it follows that $Pr\ (Y = 1) = Pr\ (Y = 0)$ so C_Y=1 and $\check{C}_Y$= 0, leading to the reference model as stated in equation 2.33 (p. 40). If some other relationship is required, then this factor will be non-zero. For example, with Thurstone's suggestion of a possible level of 75% correct judgments (Thurstone, 1927a, p. 384); this would lead to an odds ratio of 3:1 and a term $\check{C}_Y \approx$ -1.1 / α.

The derivation in this section has been given primarily to emphasise the underpinning assumption of conditional independence, and the lack of other distributional assumptions. Clearly, though, the abstract model presented earlier can be derived directly from probability axioms using the logic presented above.

2.4.2. Dichotomous Rasch Model

It can be noted that by setting α to 1 in equation 2.33 (p. 40) and substituting the usual Rasch model notation B_n for θ_n and D_i for β_i, the model can be expressed as:

$$P_{ni} = \frac{1}{1 + e^{-(B_n - D_i)}} \tag{2.46}$$

$$P_{ni} = \frac{e^{(B_n - D_i)}}{e^{(B_n - D_i)} + e^{(B_n - D_i)}e^{-(B_n - D_i)}} \tag{2.47}$$

$$P_{ni} = \frac{e^{(B_n - D_i)}}{e^{(B_n - D_i)} + 1} = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}} \tag{2.48}$$

This last equation (2.48) is the form in which the dichotomous Rasch model is usually expressed in the literature (Rasch, 1960/1980). Setting α to 1 causes the

resulting scale to have a unit of a logit; each unit thus represents a ratio of Euler's number, or approximately 2.718, in the underlying ratio scale. The Rasch model thus corresponds to the reference model given in this work, with a scale unit of a logit. Scores based on counts of thresholds passed are sufficient statistics for parameter estimation.

2.4.3. One Parameter Logistic Model

This model is normally expressed (Harris, 1989, p. 38) as:

$$P_{ni} = \frac{1}{1 + e^{-1.7a(\theta_n - b_i)}} \tag{2.49}$$

The item difficulty parameter b_i corresponds directly to the β_i in equation 2.33 (p. 40); the scale parameter (a) is constant and is usually set to 1. This is thus equivalent to setting α to 1.7 which enables the function to approximate the cumulative normal distribution. It has been shown that with this parameterisation, the numerical evaluations differ by less than .01 over the entire domain of the variable (Johnson & Kotz, 1972). A better approximation is given by a value of 1.702 (Bowling, Khasawneh, Kaewkuekool, & Cho, 2009, p. 120) which is the value used in this project. When this value is used the scale is expressed in units that represent a ratio of approximately 5.48 in the underlying ratio scale.

The formulation thus corresponds to the reference model given in this work, with a scale unit of 1.7 (or 1.702) logits. The model is functionally equivalent to the Rasch model and a linear transformation can be used to translate between the two.

There is, however, a fundamental conceptual difference between the two models. The Rasch model approaches the task from a pure measurement perspective; there is no reason to assume that measurements will be distributed normally, or in any particular way. In contrast, the logistic models attempt to accommodate traditional conceptions of scoring, and distributional assumptions, within the model formulation. Although this has no material effect on the one parameter model, subsequent logistic models depart progressively further from the pure measurement model.

As with the reference and Rasch models, scores based on counts of thresholds passed are sufficient statistics for parameter estimation.

2.4.4. Two Parameter Logistic Model

The two parameter model is normally expressed (Harris, 1989, p. 37) as:

$$P_{ni} = \frac{1}{1 + e^{-1.7a_i(\theta_n - b_i)}} \tag{2.50}$$

The only difference between this model and the one parameter model is that the parameter a_i varies by item rather than being a scale parameter. This parameter is usually called a discrimination parameter and is intended to model the fact that some items discriminate more sharply between subjects than others. Discrimination is traditionally expressed in classical test theory as a discrimination index, calculated as the difference in proportion correct between an upper group and a lower group. Under the assumption that test scores are normally distributed, Kelley (1939) showed that the ideal group sizes for this comparison are 27% for scores and 25% for dichotomous items. The parameter a_i is intended to create an analogue of the discrimination index that is not dependent on the specification of such groups.

The difference between Item Response modelling and Rasch modelling is brought into sharp focus by this parameter. Both camps agree that item discrimination is real and causes misfit to the model. Proponents of Item Response Theory argue that because such discrimination is real, it should be accommodated in the model. Those on the Rasch side argue that accommodating a discrimination parameter is akin to mixing “apples and oranges” and that it is better to investigate the reasons for the misfit and so improve future data. Andrich (2004) gives a comprehensive review of both sides of the argument.

It is also not clear whether discrimination should be modelled as an item parameter or as a subject parameter; do items discriminate, or do people discriminate? In either case, modelling discrimination clearly requires another dimension: two parameters for items, subjects or both. This could be achieved by using vectors, or

perhaps by using complex numbers rather than real numbers. However, measurements are inherently one dimensional; ultimately, some mapping to real numbers will be required to produce meaningful measurement. There is no clear theoretical basis for doing this. Indeed, if there were, there would be no need to use two dimensions in the model.

The two parameter logistic model simply takes the item discrimination factor and applies it to both subject abilities and item difficulties. Application to item difficulties causes no major problem but application to the subject ability causes the following interaction effect. For any two items with modelled difficulties β_i and β_j, an ability threshold θ_0 will exist such that the comparative difficulties of the items will be in opposite directions for subjects with abilities below this threshold, compared to those above the threshold. In consequence, scores based on counts of thresholds passed are not sufficient statistics for parameter estimation. However, a sum of thresholds passed, weighted by discrimination, is a sufficient statistic. From a theoretical perspective, a further issue is that any interaction between ability and difficulty parameters challenges the implicit assumption of conditional independence. There is also, no clear scale unit. In effect, a separate ruler is used for each item, and ability estimates represent an undefined aggregation of such units.

For all of these reasons, and following the principle of parsimony, the model is not considered suitable for measurement in this project. However, the model is included in the software, primarily for diagnostic purposes. Examination of items with extreme discrimination parameters can give useful insights into why these items are not performing as expected.

There is also a 3-parameter logistic model (Harris, 1989, p. 36) that attempts to model responses to multi-choice questions and also includes a “guessing” parameter. However, guessing is not relevant to the present work so this model is not pursued further herein.

2.4.5. Polytomous models

The reference model, dichotomous Rasch model, and one-parameter logistic model are all isomorphic, requiring only a linear transformation to translate between them. This discussion continues with the reference model, but it should be noted that the same comments apply to all three of these models.

Dichotomous judgements can be viewed as the classification of locations on an underlying continuum into two categories: those above some threshold or "cut point", and those below. Within the reference model (2.33, p. 40), used for this discussion, the parameter β_i has a natural interpretation as such a threshold. If θ_n is interpreted as ability, and β_i as difficulty, then subjects with lower ability are less likely than those with higher ability to be judged as above this threshold and thus classified in the upper category. The parameter β_i can thus be interpreted as the threshold above which classification in the upper category is more likely than the lower category.

A number of attempts have been made to generalise the logistic model from dichotomous responses to responses on an ordinal scale. This section will introduce notation and a reference model for such a generalisation and then review some of the approaches that have been proposed.

For an ordinal scale with k categories, the underlying continuum can be conceptualised as having $k-1$ cut points or thresholds separating the categories. Let these thresholds be indexed by j. If the thresholds are placed on the common metric used for items and subjects, then thresholds β_{ij} can be defined with the same interpretation as the β_i for a dichotomous model: the threshold above which classification in the one of the categories above the threshold is more likely than classification in one of those below the threshold.

The categories chosen in responses can then give evidence about the locations of these thresholds. Response categories will be translated into a set of dichotomous responses, one for each threshold. Let t index the thresholds and k index the chosen response category. Then a dichotomous response can be defined as

$$DR(k,t) = 1\ whenever\ k > t; and\ 0\ otherwise \tag{2.51}$$

For example, the scheme set out in Table 2.3 shows the dichotomised response pattern for an item with 5 categories.

Table 2.3: Dichotomisation of response categories

Scale category	Threshold 1	Threshold 2	Threshold 3	Threshold 4
Category 1	0	0	0	0
Category 2	1	0	0	0
Category 3	1	1	0	0
Category 4	1	1	1	0
Category 5	1	1	1	1

If τ_t represents the threshold indexed by t, then by the definition of an ordinal scale, it follows that $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$.

Louis Guttman (1944) set out the proposition that if the items in a scale have a unique ordering, then it should be possible to reproduce the individual responses from the overall score. Guttman (1947) used the term the *Cornell Technique* for this approach, but it is now widely known as a *Guttman scale*. It can be observed that the pattern in Table 2.3 follows a Guttman scale; if a response for a threshold is coded as 1, then all lower thresholds are also coded as 1; if a response for a threshold is coded as 0, then all higher thresholds are also coded as 0.

The reference polytomous model is now stated. Let P_{nij} be the probability that response of subject n to item i will be in a category above that item's threshold j.

$$P_{nij} = \frac{1}{1 + e^{-\alpha(\theta_n - \beta_{ij})}} \tag{2.52}$$

To express the probability of responding in a specific category k, rather than in any category above a threshold j, the following definition can be applied.

$$\check{P}_{nik} = \begin{vmatrix} 1 - P_{ni1} & when\ k = 1 \\ P_{ni(k-1)} & for\ the\ last\ category \\ P_{ni(k-1)} - P_{nik} & otherwise \end{vmatrix} \tag{2.53}$$

If required, an item difficulty β_i can be inferred by finding the central point of the item response function:

$$\beta_i \stackrel{\text{def}}{=} \zeta, s.t. \sum_{j=1}^{k-1} \frac{1}{1 + e^{-\alpha(\zeta - \beta_{ij})}} = \frac{k-1}{2} \tag{2.54}$$

Two important properties of the reference model can now be stated. Firstly, the thresholds β_{ij} are on the same metric as the item difficulties β_i in the standard dichotomous model, and have the same interpretation. This means that polytomous items and standard dichotomous items can be freely mixed in the same scale. Secondly, consecutive categories can be collapsed by removing the threshold separating them, without affecting either the meaning or estimation of the remaining thresholds. This is an essential measurement property. For example, in an educational context, a judgement of competency might be recorded on an ordinal scale: *not yet competent*, *nearly competent*, or *competent*. *Nearly competent* can be seen as a subdivision of the class *not competent*, indicating the subject is close to the upper end of this continuum. Clearly, combining the categories *not yet competent* and *nearly competent* should not affect the threshold level at which a subject is deemed competent.

It can be noted that the treatment of k categories as $k - 1$ dichotomous thresholds may break the assumption of conditional independence that is fundamental to the model as explained in 2.4.1. For example, if the ordinal pattern is to be preserved, it is not possible for a higher threshold to be coded as 1 if a lower one is coded as 0 and thus the thresholds are clearly not independent. For the present, the review will proceed with the naïve assumption that the thresholds of an item are independent of each other in the same way that responses to an item are independent of other items. A formal treatment of the corrections needed for polytomous models will be given in 2.5.7.

A number of polytomous models will now be reviewed and contrasted with the reference model.

2.4.5.1. Partial Credit Model

Masters (1982) introduced a polytomous model, using the principles of the Rasch model, which is now commonly known as the *Partial Credit Model* (PCM). He illustrates his motivation for the model by reference to an approach to scoring solutions to an arithmetic problem (2.55) that requires a number of intermediate steps in its solution.

$$\sqrt{(7.5/0.3) - 16} = ? \tag{2.55}$$

He suggests that the scoring system shown in Table 2.4 is appropriate to give partial credit for the correct calculation of intermediate steps. He then argues (p. 157) that allocating scores in this way conflicts with the requirement for specific objectivity in the reference model; essentially, the requirement to complete an answer in steps introduces dependencies between the thresholds.

Table 2.4: A scoring schedule for the Partial Credit Model

Not done	0 marks
Correct calculation of 7.5 / 0.3 = 25	1 mark
As above AND correct calculation of 25 – 16 = 9	2 marks
As above AND correct calculation of √9 = 3	3 marks

He then develops a model based on Rasch principles that attempts to model these dependencies. Using the notation of the reference model, replacing the threshold difficulty parameter β_{ij} with a model step parameter δ_{ij}, and using k to index a response in category k of m categories, the Partial Credit Model can be expressed as shown in equation 2.56:

$$P_{nik} = \frac{e^{\sum_{j=1}^{k}(\theta_n - \delta_{ij})}}{\sum_{x=1}^{m}\left(e^{\sum_{j=1}^{x}(\theta_n - \delta_{ij})}\right)} \tag{2.56}$$

He argues that,

> Unlike the item "levels" λ_{ik} [β_{ij} in the reference model], each of which represents the difficulty of reaching performance level k in item i, the individual step difficulties δ_{ij}, (k = 1, 2, ..., m) can be separated from, and estimated independently of the person parameters. (p. 158)

He then asserts that the model can be generalised to model arbitrary rating scales as a stepped process. For example, he argues that responses on a four point Likert scale of agreement can be viewed as a process with steps, so that a person who chooses the AGREE category:

> … can be considered to have chosen DISAGREE over STRONGLY DISAGREE (first step taken) and also AGREE over DISAGREE (second step taken), but to have failed to choose STRONGLY AGREE over AGREE (third step rejected). (p. 156)

It is by no means clear that this is how people actually choose a category. Andrich (1998) challenges this conception and argues forcibly that measurement principles should be consistent with physical measurement:

> The word "step" has no place in measurement. You can take steps over thresholds, but you can have steps without thresholds, and thresholds without steps, and measurement is not about modelling people taking steps, but modelling the responses to check where they are. The word "step" has a connotation of a local distance in the sense that the next step starts where the previous one finished and on movement. It is never used when we are trying to explain physical measurement. (p. 648)

This criticism is, perhaps, too harsh. One could consider the task of measuring the width of a sheet of A4 paper, using a 30 cm ruler marked with numbered centimetre divisions, each subdivided into unlabelled one millimetre units. The width could be established by aligning the zero mark with one side of the paper and counting the number of millimetre divisions passed until the other side is reached. Equally, one could start at the 30cm end and count backwards until the edge of the paper is reached. Perhaps more likely than each of these is that one would align the

ruler with one side of the paper and then look at the other side of the paper, identify the centimetre marks either side of the boundary and refine the estimate by counting millimetres.

The important point is that each of these measuring processes should produce the same measurement. From a measurement perspective, therefore, there is no fundamental reason why individual steps in a process should not be modelled, provided it can be shown that the same measurement would result from the modelled process as from other possible processes. Conversely, unless this can be demonstrated, its value as a measurement instrument is questionable.

Unfortunately, the PCM places its step estimates on a distinct local metric for each modelled process and produces different measurements for different modelled processes. This assertion will be justified by the following example. Assume that there are 4 categories as in Master's Likert scale example. To simplify the calculations, assume that there is a set of subjects all of whom have the same ability = 0 and that P_{x} represents the probability of any of these subjects responding in category x. If the steps are conceptualised as proceeding as proceeding in category order as suggested by Masters, then:

$$P_1 = e^{(-\delta_1)}/\Psi \tag{2.57}$$

$$P_2 = e^{(-\delta_1-\delta_2)}/\Psi$$

$$P_3 = e^{(-\delta_1-\delta_2-\delta_3)}/\Psi$$

$$P_4 = e^{(-\delta_1-\delta_2-\delta_3-\delta_4)}/\Psi$$

where $\Psi = e^{(-\delta_1)} + e^{(-\delta_1-\delta_2)} + e^{(-\delta_1-\delta_2-\delta_3)} + e^{(-\delta_1-\delta_2-\delta_3-\delta_4)}$

However, if the steps are conceptualised as proceeding as proceeding in reverse category order, then:

$$P_1 = e^{(-\delta_1-\delta_2-\delta_3-\delta_4)}/\Psi \tag{2.58}$$

$$P_2 = e^{(-\delta_2-\delta_3-\delta_4)}/\Psi$$

$$P_3 = e^{(-\delta_3-\delta_4)}/\Psi$$

$$P_4 = e^{(-\delta_4)}/\Psi$$

where $\Psi = e^{(-\delta_4)} + e^{(-\delta_3-\delta_4)} + e^{(-\delta_2-\delta_3-\delta_4)} + e^{(-\delta_1-\delta_2-\delta_3-\delta_4)}$

For the model to fit the data, these probabilities must be the same. Therefore, for these two processes to produce the same estimates δ_x, the following equalities must hold:

From equations 2.57	From equations 2.58	Conclusion	
$P_1/P_2 = e^{(\delta_2)}$	$P_1/P_2 = e^{(-\delta_1)}$	$\delta_2 = -\delta_1$	$\delta_1 + \delta_2 = 0$
$P_2/P_3 = e^{(\delta_3)}$	$P_2/P_3 = e^{(-\delta_2)}$	$\delta_3 = -\delta_2$	$\delta_2 + \delta_3 = 0$
$P_1/P_3 = e^{(\delta_2+\delta_3)} = e^0$	$P_1/P_3 = e^{(-\delta_1-\delta_2)} = e^0$	$P_1 = P_3$	

It is possible to continue in this way to develop more constraints, but the point is that the model is not free to set the parameters appropriately if it is required to produce the same estimates for different definitions of steps. The above is not meant to suggest that the model is in any way incorrect, but rather that the measurements are inextricable tied to the assumed sequence of steps taken. It is clear then that the partial credit model does not generalise to arbitrary rating scales in as much as the measurements produced depend on the sequence of steps modelled; the "ruler" produces different measurements when counting forward from the start; counting backward from the end; or starting somewhere in the middle.

Even with the motivating example given by Masters, it is not clear that a candidate must actually carry out the modelled steps. For example, an equally valid approach to the arithmetic problem he poses, that involves one more step and would produce different measurements under the Partial Credit model, even though the final step is the same, is set out below.

$$\sqrt{(7.5/0.3) - 16} = ? \qquad (2.59)$$

$$= \sqrt{7.5/0.3 - 4.8/0.3}$$

$$= \sqrt{2.7/0.3}$$

$$= \sqrt{9}$$

$$= 3$$

In summary, the partial credit model may be useful where the scoring system is closely tied to a well-defined process and all subjects are likely to follow that same process. Interpretation of the step parameters is closely tied to the process modelled and may therefore give useful insights when there is a misfit between actual performance and the modelled steps. However, it is also clear that collapsing categories changes the estimates of other categories so its value is questionable as a generic measurement model.

2.4.5.2. Graded Response Model

Samejima (1969, 1972) introduced her *Graded Response Model* (GRM) in which an item has ordered response categories, separated by thresholds. Using the notation of the reference model, her model can be expressed as:

$$P_{nij} = \frac{1}{1 + e^{-\alpha_i(\theta_n - \beta_{ij})}} \tag{2.60}$$

The parameter α_i represents an item discrimination parameter similar to that used in the two-parameter logistic model discussed in 2.4.4. Since the discrimination parameter is constant across all the item's thresholds, there is no problem with threshold disordering within an item. However, disordering will still apply between items and the remaining issues discussed in 2.4.4 will also apply, as they do with all models derived from the two-parameter logistic model.

Her model also allows for a variable score to be attached to each category. Where m_i is the number of categories for item i and u_{ik} is the score associated with category k of item i, the score is defined as:

$$S_{nik} = \sum_{k=1}^{m_i} u_{ik} \check{P}_{nik} \tag{2.61}$$

Raw scores or a count of thresholds passed are not sufficient statistics for this model, but a weighted sum of thresholds passed forms a sufficient statistic. If the discrimination parameter is constrained to be identical across all items, the model becomes isomorphic to the reference polytomous model.

2.4.5.3. Rating Scale Model

Muraki (1983, 1990) introduced a variant of Samejima's Graded Response Model which is now known as the *Rating Scale Model*. In this model, the item thresholds β_{ij} are replaced by two parameters β_i and δ_j. Items are represented by a single difficulty parameter β_i and the category parameters δ_j are assumed to be constant across all items.

This approach has some properties that are intuitively appealing. Firstly, sharing the category parameters across items reduces the total number of parameters to be estimated, thus according with the principle of parsimony. Secondly, item difficulty can be represented readily by a single item difficulty parameter whereas the reference model requires this to be inferred, although this poses little problem with modern computers. Thirdly, the category parameters can be estimated independently of abilities and item difficulties.

The main issues with the model are that items with different numbers of response categories or different meanings for each category cannot be used and that the model assumes a constant width of each category across the items. This last point is problematic for the present work because the variation of these widths across items is of central interest.

2.4.5.4. Andrich Rating Formulation

Andrich (1978) published a polytomous model which he termed a *Rating Formulation*. The model was based on ideas originally presented by Rasch (1968). He illustrates the logic of his model by reference to a *Likert* type of scale with two thresholds and three categories: Disagree, Neutral, Agree. Using a traditional threshold model similar to the reference model used in this work, he speculates on the process carried out by a subject responding to the item. He suggests that a

respondent first makes two independent judgments; whether the response should be above or below the threshold between disagree and neutral; and whether the response should be above or below the threshold between neutral and agree. The respondent then tries to bring these two judgements together. He describes the process thus:

> Now let us suppose that after the instantaneous reaction in the space Ω, that the subject brings the two processes together. To bring these together, we presume he recognizes that Ω is composed of the set Ω' = {(0, 0), (1, 0), (1, 1)} of legitimate responses and its complement the set {(0, 1)} containing an illegitimate response. We presume that he recognizes that {(0, 1)} is not legitimate because it would imply the inconsistency that his response reflects a position simultaneously below τ_1 and above τ_2 where $\tau_1 < \tau_2$. That is, he recognizes the ordering of the categories. Therefore we must suppose that even when the event in the complement of Ω' has happened instantaneously, that the response is reconsidered and redistributed in Ω'. (p. 567)

It is by no means clear that this is what respondents actually do. Intuitively, it seems much more likely that a respondent who chooses disagree over neutral will simply tick the disagree box and go no further, rather than proceed with further comparisons, identify conflicts, and reappraise the choices already made. Similar comments can be made to those in the discussion of the Partial Credit Model in 2.4.5.1; it is not safe to assume that participants follow a particular process unless it can be shown that such a process is the only one possible, or that different processes will result in the same measurement.

Andrich's model is presented (1978, p. 569) as:

$$\phi_{i0} \stackrel{\text{def}}{=} 0, \qquad \phi_{ix} \stackrel{\text{def}}{=} \sum_{y=1}^{x} \alpha_{iy} \tag{2.62}$$

$$\kappa_{i0} \stackrel{\text{def}}{=} 0, \qquad \kappa_{ix} \stackrel{\text{def}}{=} -\sum_{y=1}^{x} \alpha_{iy}\, \tau_y \tag{2.63}$$

$$P_{nix} = \frac{e^{(\kappa_{ix} + \phi_{ix}(\theta_n - \beta_i))}}{\sum_{y=0}^{m} \left(e^{\left(\kappa_{iy} + \phi_{iy}(\theta_n - \beta_i)\right)} \right)} \tag{2.64}$$

In this model, *m* is the total number of thresholds, α_{iy} represents a discrimination parameter for a threshold, β_i an item difficulty, τ_y an item threshold relative to item difficulty, and *x* a count of the number of ordered thresholds passed for a response. Although, in general, it is not safe to assume that a subject follows a specific process when responding, valid measurement may still be possible if different processes produce the same measurement. This line of investigation will be pursued by considering Andrich's simplest case of three categories (1978, p. 567). Substituting the notation of the reference model, Andrich's model can be expressed as:

$$\mathrm{P}_A = \frac{1}{\gamma} \tag{2.65}$$

$$\mathrm{P}_B = \frac{1}{\gamma}\, e^{(\alpha_1(\theta - \beta_1))} \tag{2.66}$$

$$P_C = \frac{1}{\gamma}\, e^{(\alpha_1(\theta - \beta_1))} e^{(\alpha_2(\theta - \beta_2))} \tag{2.67}$$

$$\gamma = 1 + e^{(\alpha_1(\theta - \beta_1))} + e^{(\alpha_1(\theta - \beta_1))} e^{(\alpha_2(\theta - \beta_2))} \tag{2.68}$$

If the first two categories are collapsed, the probability P_C should remain the same, and likewise, the threshold β_2 and discrimination α_2 should remain the same. This requires that:

$$P_C = \frac{e^{(\alpha_1(\theta - \beta_1))} e^{(\alpha_2(\theta - \beta_2))}}{1 + e^{(\alpha_1(\theta - \beta_1))} + e^{(\alpha_1(\theta - \beta_1))} e^{(\alpha_2(\theta - \beta_2))}} = \frac{e^{(\alpha_2(\theta - \beta_2))}}{1 + e^{(\alpha_2(\theta - \beta_2))}}$$

The following set of manipulations will show this is not possible.

$$x \stackrel{\text{def}}{=} e^{(\alpha_1(\theta - \beta_1))}$$

$$y \stackrel{\text{def}}{=} e^{(\alpha_2(\theta - \beta_2))}$$

$$P_C = \frac{xy}{1 + x + xy} = \frac{y}{1 + y}$$

$$xy(1 + y) = y(1 + x + xy)$$

$$x(1 + y) = (1 + x + xy)$$

$$x + xy = 1 + x + xy$$

$$0 = 1$$

This contradiction demonstrates that collapsing the first two categories will result at least in a different value of either the threshold β_2, or the discrimination α_2. As with Master's partial credit model, the measurements produced are dependent on the specific categories chosen. It can also be noted that unless the discrimination parameters are constrained to be identical, there are similar issues to those for the two-parameter logistic model in 2.4.4.

2.4.5.5. Summary of polytomous models

There are several other polytomous models. Items with an agree/disagree format or preference data may sometimes be assumed to follow an unfolding or ideal point process (Coombs, 1964). In this perspective, respondents endorse statements that are close to their position on a latent metric and disagree with statements either from above or below. Examples of these models are the hyperbolic cosine model (Andrich D. , 1996) and the generalized graded unfolding model (Roberts, Donoghue, & Laughlin, 2000). These models are not directly relevant to the current project which requires a dominance model rather than an ideal point process.

Within the dominance models, the most general polytomous model is the nominal response model (NRM) proposed by Bock (1972). If the categories are constrained to have an ordinal sequence the NRM becomes equivalent to the generalized partial credit model (GPCM) proposed by Muraki (1992). Master's PCM and Andrich's RFM can be seen as special cases of the GPCM with appropriate choice and interpretation of parameters. These models can be derived directly from the dichotomous Rasch model under the assumption that the natural scoring function is a sufficient statistic. The consequence of this, however, is that the item parameters

sit on a metric that is local to the item and need not even be sequentially ordered (Muraki, 1990, p. 164).

From a technical perspective, Andrich's RFM is intuitively appealing because of its close connection with the dichotomous Rasch mode. There are two practical concerns, however, for the current project. The first concern is that the item threshold parameters are difficult for educators and learners to understand; they are parameters on a separate metric from the main latent trait. The second concern is the requirement for all categories to have adequate response counts. This is likely to limit the use to relatively large classes in an educational context. The reference model, Samejima's GRM and Muraki's RSM avoid both these concerns. The main conclusions of the foregoing discussion are summarised in Table 2.5.

Table 2.5: Summary of polytomous models

Criterion	Reference Model	Masters PCM	Samejima GRM	Muraki RSM	Andrich RFM
All estimates are on the same metric and are consistent with the standard dichotomous model.	Yes	No	No, But Yes if α_i constant	No, But Yes if α_i constant	No
Adjacent categories can be collapsed without affecting other estimates.	Yes	No	Yes	Yes	No
Measurements are independent of subject's process.	Yes	No	Yes	Yes	Yes

From this table, it can be seen that neither Masters' partial credit model, nor Andrich's rating formulation model fit the measurement criteria required by this project. Samejima's graded response model and Muraki's rating scale model fit the criteria if the discrimination parameter is held constant. If this is done, Samejima's model becomes isomorphic to the reference model. Muraki's extension of Samejima's model, while intuitively appealing for general measurement, does not

allow study of the variation of category thresholds which is one of the goals of the current project. The reference polytomous model is therefore used as the basis for this project.

2.4.6. Time Series

It is often useful to investigate how measured attributes change over time. For example, educators may be interested in tracking the growth of students' skills and knowledge. A specific goal of this project is to produce multiple successive measurements of subjects in a time-series. This section presents an extension to the measurement model that supports such a series of measurements over time.

Wright (2003) describes an approach to using the Rasch model in a situation where a pre-test and post-test is used to investigate changes. He uses the term *stacking* to refer to the entry of before and after observations as separate subjects so that the data file contains twice as many cases as subjects. The data set is then analysed to construct measures on all items and cases concurrently. Once the cases estimates are completed, a before/after comparison can be made. He then addresses the question: "Doesn't putting the same subjects in twice introduce dependency?" (p. 905). He concludes that it probably does in a small way, but argues that the subjects have changed so they are not identical, and suggests that such dependencies might not be greater than other dependencies that will inevitably exist in real datasets. He then discusses in general terms the expected effect of such dependencies on model misfit, and consequently on reliability and estimation. He also introduces the term *racking* to refer to a corresponding duplication of items; in essence, *stacking* identifies who has changed and *racking* identifies what has changed.

Wright's stacking approach offers a promising starting point for the development of a time-series model, but a more formal analysis of the issues raised is required before this can be justified. A revised statement of the reference model will be given to give context to the discussion. Let t index time for a subject. Then the time-series model can be stated as:

$$P_{ntij} = \frac{1}{1 + e^{-\alpha(\theta_{nt} - \beta_{ij})}} \quad (2.69)$$

Here, P_{ntij} models the probability that subject n at time t will respond to item i in a category above threshold j; θ_{nt} models the ability of subject n at time t and, following Wright, this will be termed a case. The parameter β_{ij} retains its meaning as an item threshold. It can be noted that the only difference between this model and the previous model (2.52, p. 49) is the change from n subjects to nt cases. It can further be noted that two sets of measurements are made by the model: abilities of cases and difficulties of item thresholds. Two numbers are produced for each of these measurements: the estimated location of the ability or difficulty, and the modelled standard error of the estimate. The following discussion reviews the effect of the change from subjects to cases on each of these four numbers.

It is illustrative to begin the discussion by investigating what would happen if there were no time series but the data in the standard model were simply duplicated by entering each subject twice. Although this section makes reference to some specific details of the final measurement model which is described in section 2.5, the outline of the logic should be clear without the need for full details of that model.

It can be noted that the ability and difficulty parameters are separated in the model and that the maximum likelihood estimation procedure alternately produces ability estimates from its current threshold difficulty estimates, and threshold difficulty estimates from its current ability estimates, progressively refining both until convergence is achieved. From this, it is clear that both the estimated location and standard error of each (duplicated) subject will be the same provided the same set of threshold difficulties is produced. The same is not true for threshold difficulties, however. Focussing first on the standard error, it can be seen that the duplication of subjects would result in the understatement of the standard error by a factor of 1/√2. Effectively, the model is assuming that the information content of the dataset is twice what is really there. Provided all subjects were duplicated, however, the threshold difficulty estimates would remain the same and a simple correction factor could be applied to correct the standard error estimates.

If this example is extended to the more general case where some subjects have no duplication, and others have possibly multiple replications, again subject estimates would be correct if item estimates were correct. For items, however, both standard errors and locations could be affected. Although the measurement model makes no direct assumptions about the distribution of abilities, the extra weight given to the duplicated subjects could affect the estimates if, for example, the level of replication was associated in some way with the subjects' abilities.

Fortunately, it is easy to correct for both of these in the estimation procedure if the degree of replication is known. The simplest approach would be to discard the second and subsequent response of each subject from the threshold estimation procedure. The effect of this would be that all threshold and subject estimates would be correct. This approach is, however, less than ideal for concepts to be introduced later, such as connectivity, so an equivalent but more flexible approach will now be presented. This flexibility will be achieved by allowing the level of replication to be specified by response rather than by subject. Let D_{ni} be the level of replication of a response of subject n to item i, where D_{ni} can take the values {1, 2, …}. Each observation will be assigned a weight $W_{ni} = 1/D_{ni}$ in the threshold estimation procedure. Using this weighting will produce the same estimates of threshold location and standard error as the unduplicated estimates.

To extend the model to a time-series, the remaining issue is to identify the level of replication in the dataset. It can be noted that this replication, or equivalently the dependency identified by Wright, is what is usually termed a *carry-over effect* in the context of a time series or a repeated-measures experiment. Carry-over effects take many forms, but of particular interest to the current project is a fatigue or learning effect. Specifically, where a subject interacts with the same item on several successive occasions, it is not clear whether an active judgement has been made on each occasion, or whether the subject has simply carried forward a previous judgement without further thought. Clearly, it is reasonable to assume that an active judgement has been made if the response differs from a previous judgement, but if the same judgement is recorded, there is no empirical way of assessing

whether this is the result of an active process, or just a carry-over of a previous judgement.

Accordingly, a conservative approach is taken in this project, so as not to understate standard errors of estimates: each observation in a sequence of consecutive identical observations is assigned a replication level of the number of repeated observations in that sequence.

2.5. SPECIFICATION OF THE MODEL

A formal presentation of the measurement model used in this project will now be given. Let α be a parameter identifying the unit of the measurement scale; n index a set of subjects with modelled abilities θ_{nt} at time t; and i index a set of items with modelled difficulty thresholds β_{ij}, where j indexes the thresholds for item i. The probability that subject n at time t will be placed in a category above threshold j on item i is defined as:

$$P_{ntij} \stackrel{\text{def}}{=} \frac{1}{1 + e^{-\alpha(\theta_{nt} - \beta_{ij})}} \tag{2.70}$$

The values of θ_{nt} and β_{ij} which give the best fit to this model are estimated by a maximum likelihood procedure. A derivation of the maximum likelihood equations is given in 2.5.1. An approach to the solution of this set of equations is set out in 2.5.2, and the estimation process used is described in 2.5.3. The relationship between raw scores and the measurements produced by the model is non-linear. This mapping, the instrument's *characteristic curve*, is described in 2.5.4. In addition to producing measurements, the procedure used also produces estimates of the uncertainty associated with the measurements. How this is done is described in section 2.5.5.

2.5.1. Maximum Likelihood Equations

The maximum likelihood equations are derived as follows. First, an expression for the likelihood of the data is derived from the model. Second, this expression is differentiated with respect to each of the model parameters. Third, since these

derivatives will be zero at each maximum, equating these derivatives to zero will define the set of equations that characterises the solution.

For the purposes of this derivation, a more compact notation will be introduced. The nt cases will be indexed by u; likewise the ij item thresholds will be indexed by v; A_u will represent the term $\alpha\theta_{nt}$; and B_v will represent the term $\alpha\beta_{ij}$. With these substitutions, the model can be restated as:

$$P_{uv} = \frac{1}{1 + e^{-(A_u - B_v)}} \tag{2.71}$$

Multiplying the numerator and denominator of the expression on the right by $e^{(\mathrm{A_u - B_v})}$, this can be restated as:

$$P_{uv} = \frac{e^{(\mathrm{A_u - B_v})}}{1 + e^{(\mathrm{A_u - B_v})}} \tag{2.72}$$

For convenience the complement of P_{uv} can be defined as:

$$Q_{uv} = 1 - P_{uv} = \frac{1}{1 + e^{(\mathrm{A_u - B_v})}} \tag{2.73}$$

Let X_{uv} be 1 if the response of case A_u is above item threshold B_v and 0 otherwise. Then, assuming conditional independence, the likelihood of the observed data can be expressed as:

$$\mathcal{L} \stackrel{\text{def}}{=} \prod_u \prod_v (X_{uv} P_{uv} + (1 - X_{uv}) Q_{uv}) \tag{2.74}$$

Substituting for P_{uv} and Q_{uv}, this can be written as:

$$\mathcal{L} = \prod_u \prod_v \left(X_{uv} \frac{e^{(\mathrm{A_u - B_v})}}{1 + e^{(\mathrm{A_u - B_v})}} + (1 - X_{uv}) \frac{1}{1 + e^{(\mathrm{A_u - B_v})}} \right)$$

$$= \prod_u \prod_v \left(\frac{X_{uv} e^{(\mathrm{A_u - B_v})} + (1 - X_{uv})}{1 + e^{(\mathrm{A_u - B_v})}} \right)$$

It can be noted that *X* is either 0 or 1. When *X* is 1, the numerator is $e^{(A_u - B_v)}$; when *X* is zero, the numerator is 1. The above equation can thus be expressed more compactly as:

$$\mathcal{L} = \prod_u \prod_v \left(\frac{e^{X_{uv}(A_u - B_v)}}{1 + e^{(A_u - B_v)}} \right) \tag{2.75}$$

The goal of the estimation procedure is to find the values of A_u and B_v that maximise this. Since a logarithm is a monotone transformation, the maximum of the log of the likelihood will occur at the same values of A_u and B_v. The log-likelihood can be expressed as:

$$\lambda \stackrel{\text{def}}{=} \log(\mathcal{L}) = \log\left(\prod_u \prod_v \left(\frac{e^{X_{uv}(A_u - B_v)}}{1 + e^{(A_u - B_v)}} \right) \right) \tag{2.76}$$

$$\lambda = \sum_u \sum_v \log\left(\frac{e^{X_{uv}(A_u - B_v)}}{1 + e^{(A_u - B_v)}} \right) \tag{2.77}$$

$$\lambda = \sum_u \sum_v \left(\log\left(e^{X_{uv}(A_u - B_v)} \right) - \log\left(1 + e^{(A_u - B_v)} \right) \right) \tag{2.78}$$

$$\lambda = \sum_u \sum_v \left(X_{uv}(A_u - B_v) - \log\left(1 + e^{(A_u - B_v)} \right) \right) \tag{2.79}$$

This definition of the (log) likelihood of the observed data from the model completes the first part of the derivation. The second stage is to differentiate this likelihood equation with respect to each of the model parameters. From equation 2.79, the partial derivative of the log likelihood with respect to A_u is:

$$\frac{d\lambda}{dA_u} = \sum_v \left(\frac{d}{dA_u}\left(X_{uv}(A_u - B_v) \right) - \frac{d}{dA_u}\left(log\left(1 + e^{(A_u - B_v)} \right) \right) \right) \tag{2.80}$$

Taking, in turn, each of the terms on the right,

$$\frac{d}{dA_u}\left(X_{uv}(A_u - B_v) \right) = X_{uv} \tag{2.81}$$

$$\frac{d}{dA_u}\left(log\left(1+\ e^{(A_u-B_v)}\right)\right)=\frac{\frac{d}{dA_u}\left(\ 1+e^{(A_u-B_v)}\right)}{1+\ e^{(A_u-B_v)}}$$

$$=\frac{e^{(A_u-B_v)}}{1+\ e^{(A_u-B_v)}}=P_{uv} \tag{2.82}$$

Substituting these into equation 2.80, the partial derivative of the log likelihood with respect to A_u is:

$$\frac{d\lambda}{dA_u}=\sum_v(X_{uv}-\ P_{uv}) \tag{2.83}$$

Similarly, from equation 2.79, the partial derivative of the log likelihood with respect to B_v is:

$$\frac{d\lambda}{dB_v}=\sum_u\left(\frac{d}{dB_v}\left(X_{uv}(A_u-B_v)\right)-\frac{d}{dB_v}\left(log\left(1+\ e^{(A_u-B_v)}\right)\right)\right) \tag{2.84}$$

Taking, in turn, each of the terms on the right,

$$\frac{d}{dB_v}\left(X_{uv}(A_u-B_v)\right)=-X_{uv} \tag{2.85}$$

$$\frac{d}{dB_v}\left(log\left(1+\ e^{(A_u-B_v)}\right)\right)=\frac{\frac{d}{dB_v}\left(\ 1+e^{(A_u-B_v)}\right)}{1+\ e^{(A_u-B_v)}}$$

$$=\frac{-e^{(A_u-B_v)}}{1+\ e^{(A_u-B_v)}}=-P_{uv} \tag{2.86}$$

Substituting these into equation 2.84, the partial derivative of the log likelihood with respect to B_v is:

$$\frac{d\lambda}{dB_v}=\sum_u\left((-X_{uv})-\ (-P_{uv})\right)=\sum_u(P_{uv}-X_{uv}\ ) \tag{2.87}$$

This completes the second stage of this derivation. The third stage is to set out the set of equations for which the derivative is zero, thus defining the solution. From equation 2.83 (p. 67), the solutions for parameters A_u are thus those for which:

$$\sum_{v}(X_{uv} - P_{uv}) = 0 \tag{2.88}$$

Similarly, from equation 2.87, the set of solutions for the parameters B_v requires:

$$\sum_{u}(P_{uv} - X_{uv}) = 0 \tag{2.89}$$

These two sets of equations define the maximum likelihood solution.

2.5.2. Solution of the Equations

There is no closed form solution to these equations. However, a numerical solution is possible by successive approximation. Where a function $f(\theta)$ is differentiable, the *Newton-Raphson* method can be used to derive an improved estimate $\hat{\theta}$ from an existing estimate θ by the formula:

$$\hat{\theta} = \theta - f(\theta)/f'(\theta) \tag{2.90}$$

In this formula, $f'(\theta)$ represents the derivative of $f(\theta)$. To apply this method, an initial estimate is made of the parameter θ and the formula is used to derive an improved estimate $\hat{\theta}$. The process is then repeated successively on each improved estimate until sufficient precision is achieved. For the present project, precision is considered sufficient when the magnitude of the improvement $(\hat{\theta} - \theta)$ in any iteration becomes less than 10^{-6} (six decimal places).

Before deriving the Newton-Raphson equations, it can be noted that:

$$P_{uv} = \frac{e^{A_u - B_v}}{1 + e^{A_u - B_v}} = \frac{e^{A_u}}{e^{A_u} + e^{B_v}} \tag{2.91}$$

$$\frac{d}{dA_u}(P_{uv}) = \frac{e^{A_u}(e^{A_u} + e^{B_v}) - e^{A_u}\, e^{A_u}}{(e^{A_u} + e^{B_v})(e^{A_u} + e^{B_v})}$$

$$= P_{uv} - P_{uv}^2 = P_{uv}(1 - P_{uv}) \tag{2.92}$$

$$\frac{d}{dB_v}(P_{uv}) = \frac{0 - e^{B_v}\, e^{A_u}}{(e^{A_u} + e^{B_v})(e^{A_u} + e^{B_v})}$$

$$= -(1 - P_{uv})P_{uv} = -P_{uv}(1 - P_{uv}) \tag{2.93}$$

Equations for the parameters A_u are derived below:

$$f(A_u) \stackrel{\text{def}}{=} \sum_v (X_{uv} - P_{uv}) \tag{2.94}$$

$$f'(A_u) = \frac{d}{dA_u}\left(\sum_v -P_{uv}\right) = -\sum_v \left(P_{uv}(1 - P_{uv})\right) \tag{2.95}$$

$$\hat{A}_u \stackrel{\text{def}}{=} A_u - \frac{f(A_u)}{f'(A_u)} \tag{2.96}$$

$$\hat{A}_u = A_u - \frac{\sum_v (X_{uv} - P_{uv})}{-\sum_v \left(P_{uv}(1 - P_{uv})\right)} \tag{2.97}$$

$$\hat{A}_u = A_u + \frac{\sum_v (X_{uv} - P_{uv})}{\sum_v \left(P_{uv}(1 - P_{uv})\right)} \tag{2.98}$$

Similarly, the Newton-Raphson equations for the parameters B_v are:

$$f(B_v) \stackrel{\text{def}}{=} \sum_u (P_{uv} - X_{uv}) \tag{2.99}$$

$$f'(B_v) = \frac{d}{dB_v}\left(\sum_u P_{uv}\right) = -\sum_u \left(P_{uv}(1 - P_{uv})\right) \tag{2.100}$$

$$\hat{B}_v \stackrel{\text{def}}{=} B_v - \frac{f(B_v)}{f'(B_v)} \tag{2.101}$$

$$\hat{B}_v = B_v - \frac{\sum_u (P_{uv} - X_{uv})}{-\sum_u \left(P_{uv}(1 - P_{uv})\right)} \tag{2.102}$$

$$\hat{B}_v = B_v - \frac{\sum_u (X_{uv} - P_{uv})}{\sum_u \left(P_{uv}(1 - P_{uv})\right)} \tag{2.103}$$

For conciseness, the derivation above used the compact notation introduced in 2.5.1. This will now be expressed in the notation of the reference model. Let X_{ntij} be 1 if the response of case θ_{nt} is above item threshold β_{ij} and 0 otherwise;

let W_{ntij} be the correction factor for replication of responses in the time series. Substituting in equation 2.98 gives:

$$\hat{\theta}_{nt} = \theta_{nt} + \left(\frac{1}{\alpha}\right)\frac{\sum_{ij}\left(X_{ntij} - P_{ntij}\right)}{\sum_{ij}\left(P_{ntij}\left(1 - P_{ntij}\right)\right)} \tag{2.104}$$

In practice, it is convenient to represent the scaling term $1/\alpha$ as α/α^2. With this change, equation 2.104 can be restated as:

$$\hat{\theta}_{nt} = \theta_{nt} + \left(\frac{\alpha}{\alpha^2}\right)\frac{\sum_{ij}\left(X_{ntij} - P_{ntij}\right)}{\sum_{ij}\left(P_{ntij}\left(1 - P_{ntij}\right)\right)} \tag{2.105}$$

$$\hat{\theta}_{nt} = \theta_{nt} + \frac{\sum_{ij}\left(\alpha\left(X_{ntij} - P_{ntij}\right)\right)}{\sum_{ij}\left(\alpha^2\left(P_{ntij}\left(1 - P_{ntij}\right)\right)\right)} \tag{2.106}$$

This change is convenient for two reasons. First, in this form, the denominator becomes the *Fisher Information* associated with the estimate. Fisher Information is discussed later in 2.5.5, and the change allows it to be produced as a side effect of the estimation procedure. Second, the change allows the software to handle more readily additional models, such as the two-parameter logistic (2PL) model, in which the discrimination (α) varies by item or item threshold. Although it is not required for measurement in this project, the 2PL model has been implemented for diagnostic purposes.

Similarly, substituting the notation of the reference model in equation 2.103 (p. 69) gives:

$$\hat{\beta}_{ij} = \beta_{ij} - \frac{\sum_{nt}\left(\alpha\left(X_{ntij} - P_{ntij}\right)\right)}{\sum_{nt}\left(\alpha^2\left(P_{ntij}\left(1 - P_{ntij}\right)\right)\right)} \tag{2.107}$$

Applying the *time series* correction, discussed in 2.4.6, to this formula yields the estimation equation used in this project:

$$\hat{\beta}_{ij} = \beta_{ij} - \frac{\sum_{nt}\left(\alpha\left(X_{ntij} - P_{ntij}\right)W_{ntij}\right)}{\sum_{nt}\left(\alpha^2\left(P_{ntij}\left(1 - P_{ntij}\right)\right)W_{ntij}\right)} \quad (2.108)$$

These equations define how subject case abilities can be determined, if item threshold difficulties are known (2.106, p. 70), and how item threshold difficulties can be determined if subject case abilities are known (2.108). However, an additional constraint is required for a solution to be fully determined. By inspection of the reference model in equation 2.70 (p. 64), it can be seen that if a set of θ_{nt} and β_{ij} are solutions to the equation, then $\hat{\theta}_{nt} = \theta_{nt} + \delta$ and $\hat{\beta}_{ij} = \beta_{ij} + \delta$ are also solutions. Without an additional constraint, there are thus infinitely many solutions. Such a constraint is equivalent to setting the scale origin and there are two possible approaches to doing this.

With a *norm-referenced* approach, the assumption is made that each sample of subjects is equivalent and that, consequently, variation in responses on different occasions should be attributable to different items or mixes of items. If this is the focus, then the ability estimates of subjects can be constrained, typically by requiring the mean to be zero. The resulting item estimates are then expressed relative to this mean ability, enabling direct comparison of item difficulties.

With a *criterion-referenced* approach, a constraint is applied to item threshold difficulties. This could be achieved by requiring the mean item difficulty to be zero, or by anchoring item difficulties against some external reference or criterion. The resulting measurements are thus expressed relative either to mean item difficulty, or to an external reference or criterion. The focus with this approach is on the measurement of subject attributes.

For the current project, the issue of setting the origin has been separated into two parts: defining a scale centre, and setting constraints relative to that centre. Choice of the scale centre, which is termed C in this work, depends on the representational scale chosen; typically, it is assigned a value of zero, or the centre of the representational scale. Options for representational scales are discussed in section 2.5.4. Four approaches to setting the constraints have been investigated.

Each involves adjusting case abilities or item threshold difficulties to achieve a given criterion. The criteria are:

- The mean case ability is C
- The mean item threshold difficulty is C
- The median item threshold difficulty is C
- The expected score at ability C is 50% of the maximum

With any of these constraints, the maximum likelihood equations fully define the solution. The process of applying the Newton-Raphson estimation equations is discussed in the next section.

2.5.3. Estimation Process

The overall approach to the estimation process is as follows:

1) An initial set of estimates for the parameters θ_{nt} is produced.
2) If *norm-referencing* is being used, an adjustment is applied to the parameters θ_{nt} to set the origin of the measurement scale to the mean case ability.
3) A set of estimates for the parameters β_{ij} is produced, using the case parameters θ_{nt}, by successive application of equation 2.108 (p.71).
4) If *criterion referencing* is being used, an adjustment is applied to the parameters β_{ij} to constrain and determine the origin of the measurement scale.
5) A revised set of estimates for the case parameters θ_{nt} is produced, using the item threshold parameters β_{ij}, by successive application of equation 2.106 (p. 70).
6) Steps 2, 3, 4 and 5 are repeated until model convergence is achieved.

The choice of initial estimates (step 1) is not critical for the operation of the estimation procedure, but convergence is more rapid if reasonable initial estimates for the set of subject abilities. Initial estimates for subject abilities are made according to:

$$\theta_{nt} = C + log(A_{nt}/B_{nt})\ /\alpha \quad (2.109)$$

In this equation, C is the intended centre of the scale, A_{nt} is the sum across responses of the count of categories below each chosen response category (the natural score) and B_{nt} is the sum across responses of the count of categories above each chosen response category (the complement of the natural score). The intended centre of the scale will depend on the chosen representational model. Options for this are discussed in section 2.5.4.

In step 2, case constraints are applied by calculating the implicit centre ($\hat{C}$) as the mean of the current case ability estimates (θ_{nt}). The difference between this and the intended centre C is calculated as $\delta = C - \hat{C}$. Each parameter estimate is then adjusted to align the centre. Item difficulty thresholds are adjusted by the equation $\hat{\beta}_{ij} = \beta_{ij} + \delta$ and subject case abilities are adjusted by the equation $\hat{\theta}_{nt} = \theta_{nt} + \delta$.

In step 3, each threshold parameter β_{ij} is estimated in turn. Each of these parameter estimations involves the successive application of equation 2.108 (p.71) until sufficient accuracy (six digits) is achieved.

In step 4, item constraints are applied by calculating the implicit centre ($\hat{C}$) according to the current item threshold parameter estimates (β_{ij}) and the specified criterion. The difference between this and the intended centre C is calculated as $\delta = C - \hat{C}$. Each of the parameter estimates is then adjusted to align the centre. Item difficulty thresholds are adjusted by the equation $\hat{\beta}_{ij} = \beta_{ij} + \delta$ and, similarly, subject case abilities are adjusted by the equation $\widehat{\theta}_{nt} = \theta_{nt} + \delta$.

In step 5, each subject case ability (θ_{nt}) is estimated in turn. Each of these parameter estimations involves the successive application of equation 2.106 (p. 70) until sufficient accuracy (six digits) is achieved. The change of each estimated ability (θ_{nt}) in the iteration is tracked.

The model is deemed to have converged (step 6) when the average change across all subject cases is less than 10^{-6} (six decimal places). Once the model has

converged, the final abilities (θ_{nt}) and item threshold difficulties (β_{ij}) are the maximum likelihood estimates that form the output measurements of the model.

2.5.4. Characteristic Curve

The characteristic curve expresses the relationship between scores and the measurements produced by the model. The model allows two approaches to scoring: implicit and explicit. With *implicit scoring*, the k categories are assigned the scores: {0, 1, ... , k-1}; these natural scores are thus a count of the number of categories below the chosen category. Equivalently, they may be viewed as a count of the number of thresholds that separate the categories which are crossed or passed when moving from the lowest category in the scale to the chosen category. With *explicit scoring*, arbitrary scores can be assigned to categories, subject only to the constraint that no category has a lower score than a lower category. This constraint is required from the definition of an ordinal scale. Regardless of the approach, a score can be defined for each threshold as the difference between the scores of the categories on either side of the threshold. With implicit scores, this difference is always 1; with explicit scores, the difference will vary, but will always be non-negative. Where a non-zero score is allocated to the first category, it will be treated as zero by the model. With this convention, if a score S_{ij} is associated with each item threshold β_{ij}, then the expected score for a case with ability θ is:

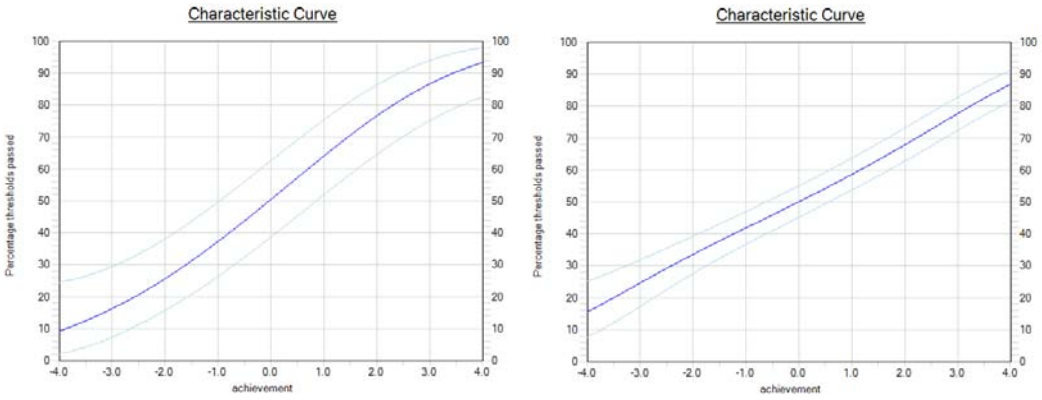
$$E(Score|\theta) = \sum_{ij} \left(\frac{S_{ij}}{1 + e^{-\alpha(\theta - \beta_{ij})}} \right) \qquad (2.110)$$

This can be standardised and expressed as a proportion by dividing by the total score:

$$E(Proportion|\theta) = \frac{\sum_{ij} \left(\frac{S_{ij}}{1 + e^{-\alpha(\theta - \beta_{ij})}} \right)}{\sum_{ij} S_{ij}} \qquad (2.111)$$

This approach thus unifies both implicit and explicit scores. Both can be expressed as a proportion of the maximum score, with zero representing the smallest possible score. The mapping of ability θ to scores is known as the measurement

instrument's *characteristic curve*. Two examples of characteristic curves are shown in Figure 2.3.



(a) (b)

Figure 2.3: Two examples of characteristic curves.

In these examples, the dark blue line shows the main mapping and the light blue lines show 95% confidence limits. For both of these characteristic curves, achievement is shown on the x-axis and measured in logits. The y-axis in both cases shows the expected *implicit score* for each ability measure. As an alternative, *explicit scores* could be shown. Each scale shown has an operating range of -4 to 4 logits. Other options are discussed later in this section. Both of the scales shown are centred to achieve an expected score of 50% at the zero point. Other options are also discussed at later in this section. The first characteristic curve (a) was from a "high stakes" assessment that was designed to produce maximum discrimination at the pass mark of 50%, which is shown here at the zero position. The second curve (b) was from a diagnostic test that was designed to produce more even discrimination over a wide range of abilities.

Measurement on an interval scale is inherently infinite. Mapping an infinite measurement scale onto a finite score scale results in the typical "S curve" that is especially noticeable in the first of these graphs. Since scores are widely used as a proxy for ability measurements, ideally one would like this curve to be close to a straight line, thus allowing concordance between linear measurement and percentage score representation. In practice, careful construction of the instrument can achieve a curve that is close to linear over a reasonable operating range. The

linearity cannot hold, however, at the upper and lower ends of the scale, because the underlying scale is inherently infinite. Although the curve will become arbitrarily close to zero and 100% expectations at sufficiently extreme ability values, no finite limit can be set on the ability estimates.

This is in agreement with everyday notions of measurement. If a quantity exceeds the maximum of the operating range of an instrument, one can say with some confidence that the quantity is above the maximum, but cannot determine an accurate estimate of the quantity. Likewise, if the quantity is below the minimum of the operating range, one can say with some confidence that the quantity is below that threshold, but cannot determine an accurate estimate of how much it is below that minimum.

In order to provide a consistent framework for representing such extreme values, the following convention is used. Reported measurements are constrained to the reporting range of the scale. If a measurement exceeds the maximum of the scale, it is assigned a value of the scale maximum and can be interpreted as meaning that value *or higher*. If a value is below the scale minimum, it is given a value of the scale minimum and can be interpreted as meaning that value *or lower*. Thus, with the characteristic curves shown, a value of 3.99 would be interpreted as a point estimate of ability, and a value of 4.00 would be interpreted as meaning 4.00 or higher. In practice, “clamping” measurements to the operating range in this way should have little consequence on measurement if the instrument is well targeted to the measurement purpose. However, if many measurements are affected by this, the fitness for purpose of the instrument may be in doubt.

Rasch or logistic measurements are conventionally reported as positive or negative offsets relative to a scale centre of zero. However, educators and learners are more familiar with scores on a finite scale, often expressed as a percentage of the maximum possible score. To accommodate both perspectives several predefined operating ranges are supported by the model, as set out in

Table 2.6.

Table 2.6: Supported operating ranges and intended use

Scale	Range	Centre	Intended use
Standard	-4 to 4	0.0	Rasch and logistic analysis
High resolution	-6 to 6	0.0	Rasch and logistic analysis
Decimal	0 to 10	5.0	General context
Percentage	0% to 100%	50%	Educational and general context

The *standard* option is intended for Rasch and logistic measurement. The *high-resolution* option extends the scale to cater for situations where a very wide range of abilities is being measured. The *decimal* option is intended for general and informal use. It removes the need for reporting negative values which may be confusing in some contexts. The *percentage* option simply expresses the decimal scale in percentage terms where each unit corresponds to 10%; this allows an informal interpretation of the measurements in an educational context as percentage scores, "corrected" for the different difficulties of the items contributing to the scores.

The operating range described above sets out the number of scale units used to represent measurements. The meaning of each unit is given by the ratio in the implicit underlying ratio scale. Several scale units are supported, as set out in Table 2.7.

Table 2.7: Supported scale units and typical use.

Unit	Alpha	Typical use
Base 2	$\log_e 2 \approx 0.693$	Information theoretic models
Logit	1.000	Rasch model and general purpose
Normal approximation	1.702	1PL and 2PL logistic models

With a *base 2 unit*, each measurement unit corresponds to a doubling of ability in the underlying ratio scale. This may be the easiest unit to interpret where the focus is on the meaning of the underlying ratio scale. With a *logit unit*, each measurement unit corresponds to an increase of ability in the underlying ratio scale by a factor of Euler's number (≈ 2.718). This may be the most appropriate unit

when the focus is on the linear measurements produced. With the *normal approximation unit*[2], each measurement unit corresponds to an increase of ability in the underlying ratio scale by a factor of approximately 5.719. This may be the most appropriate unit when the focus is on comparison of measurements with those produced by a conventional 1PL or 2PL logistic model.

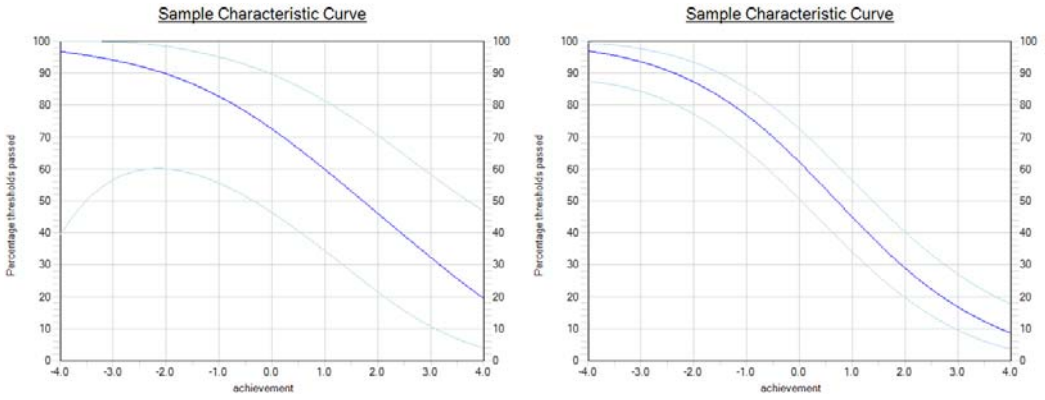
The foregoing discussion made the implicit assumption that the purpose of the measurement instrument is to assess the abilities of subjects. However, it is also possible to view the measurement process as using a sample of subjects to measure the characteristics of the instrument or items. From this perspective, a *sample* characteristic curve can be defined. This expresses the success or endorsement expected in the sample for an item threshold of a given difficulty. This can be written as:

$$E(Success|\beta) = \frac{\sum_{nt}\left(\frac{S_{ij}}{1 + e^{-\alpha(\theta_{nt}-\beta)}}\right)}{\sum_{nt}(S_{ij})} \qquad (2.112)$$

$$= \frac{\sum_{nt}\left(\frac{1}{1 + e^{-\alpha(\theta_{nt}-\beta)}}\right)}{\sum_{nt}(1)}$$

Two examples of sample characteristic curves are given in Figure 2.4. These are from the same datasets used to produce the instrument characteristic curves shown in Figure 2.3.

[2] It should be noted that the term normal approximation applies only to the individual threshold response functions. It does not imply that the characteristic curve approximates a cumulative normal distribution.



(a) (b)

Figure 2.4: Two examples of a sample characteristic curve.

It can be noticed that, in case a, an expected score of 50% corresponds to approximately 1.65 units of achievement. Since this was a high stakes test, this can be interpreted as meaning that 50% of the sample could be expected to pass a test that was composed of items which were 1.65 logits more difficult than the items used. Equivalently, since the test was targeted at a criterion based pass mark of 50%, it can be concluded that the majority of students passed the test. In case b, an expected score of 50% corresponds to approximately 0.65 units of achievement. Similar conclusions may be drawn, but it may be noticed that there is a closer match between subject abilities and item difficulties. Case b was a diagnostic test and, for such a test, a closer match provides better information about a wider range of subjects. Such a diagnostic purpose highlights the contrast between norm referencing and criterion referencing options for the centre of the scale. With a criterion referenced approach, as discussed above, the focus is on a diagnostic of subject abilities. An alternative perspective is the diagnosis of items. From this perspective, the centre of the scale can be set to the mean subject ability. Item difficulties can then be interpreted as how difficult an “average” subject finds the item. In an educational context, this can help identify items and topics which are poorly understood by a class and could benefit from additional learning and teaching time.

Both the instrument’s characteristic curve and the sample characteristic curve summarise the relationship between measurements and scores. There is also a

relationship between measurements and standard errors which is discussed in the next section.

2.5.5. Fisher Information

When exploring the properties of efficient statistics, R.A. Fisher (1925) introduced the term *intrinsic accuracy*, which he defined as the inverse of the variance. He then noted:

> What we have spoken of as the intrinsic accuracy of an error curve may equally be conceived as the amount of information in a single observation belonging to such a distribution. (p. 709)

After exploring the properties of such information, he concluded that:

> The amount of information provided by a combination of two or more independent observations is thus merely the sum of the amounts of information in each piece separately. (p. 710).

From the definition, it is clear that the variance and thus the standard error can be readily recovered from the information. The additive nature of the information allows the variance and standard errors of arbitrary sets of observations to be established simply by summing the information across those sets and then deriving the corresponding variance and standard error. For the model, the *Fisher Information* associated with a single observation is:

$$I_{ntij} = \alpha^2 W_{ntij} P_{ntij}\left(1 - P_{ntij}\right) \tag{2.113}$$

The first part of this expression (α^2) is a constant which allows the information to be expressed in the unit of measure of the scale. The weight (W_{ntij}) is a correction factor for the replications in a time series. There is no duplication when estimating case abilities from item responses, so the weight will be one in this case. However, when estimating item difficulties from cases in a time-series, there may be some replication as discussed earlier; if an observation is repeated *m* times in a time-series, then each observation is assigned a weight of *1/m*. The remaining term $P_{ntij}\left(1 - P_{ntij}\right)$ represents the entropy discharged by the observation. The

corresponding variance (V_{ntij}) and standard deviation (σ_{ntij}) associated with a single observation are:

$$V_{ntij} = 1/I_{ntij} \tag{2.114}$$

$$\sigma_{ntij} = \sqrt{V_{ntij}} \tag{2.115}$$

The *Fisher information* associated with subject case estimates (I_{nt}) and item threshold estimates (I_{ij}), together with the associated variances (V_{nt} and V_{ij}) and standard errors (σ_{nt} and σ_{ij}) can be derived by summing the information associated with individual observations across item thresholds. These statistics are set out in Table 2.8.

Table 2.8: Information and uncertainty statistics

	Subject cases	Item thresholds	
Fisher Information	$I_{nt} = \sum_{ij} I_{ntij}$	$I_{ij} = \sum_{nt} I_{ntij}$	(2.116)
Variance	$\mathrm{V}_{nt} = 1/I_{nt}$	$\mathrm{V}_{ij} = 1/I_{ij}$	(2.117)
Standard Error	$\sigma_{nt} = \sqrt{V_{nt}}$	$\sigma_{ij} = \sqrt{V_{ij}}$	(2.118)

The difference between a measurement model and a quantitative model based on classical test theory with its distributional assumptions is brought into contrast by these statistics. With the classical model, standard errors are seen as characteristic of the overall instrument. With a measurement model, they are seen as characteristic of individual measurements. A consequence of this is that the information (I), variance (V) and standard error (σ) vary across the range of the instrument. The expected values of these, conditioned on ability (θ), are given below:

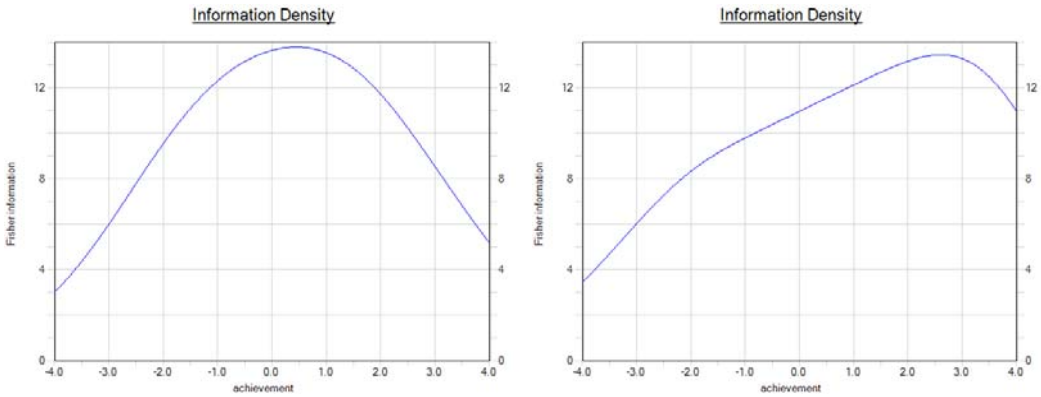
$$P_{ij}(\theta) \stackrel{\text{def}}{=} E\left(P_{ntij}\middle|\theta\right) = \frac{1}{1 + e^{-\alpha(\theta - \beta_{ij})}} \tag{2.119}$$

$$E(I|\theta) = \alpha^2 \sum_{ij} \left(P_{ij}(\theta)(1 - P_{ij}(\theta)) \right) \tag{2.120}$$

$$E(V|\theta) = \frac{1}{E(I|\theta)} \tag{2.121}$$

$$E(\sigma|\theta) = \sqrt{E(V|\theta)} \tag{2.122}$$

Two examples of *information density* curves are shown in Figure 2.5.



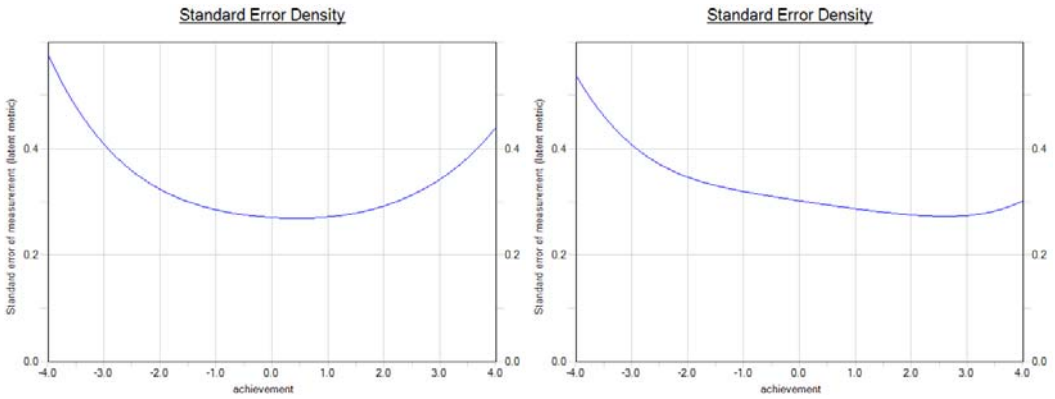
(a) (b)

Figure 2.5: Two examples of information density curves.

These two examples correspond to the datasets used earlier to illustrate the characteristic curves discussed in 2.5.4 above. The first curve (a) was for a "high-stakes" test which was designed to produce maximum information, and hence minimum uncertainty at the pass fail boundary which was at the zero position. The second curve (b) was from a diagnostic test that was designed to produce useful estimates over a wide range of abilities.

It can be noticed that the entropy term $P_{ntij}\left(1 - P_{ntij}\right)$ in equation 2.113 (p. 80) is at a maximum when P_{ntij} = 0.5, and this will occur when the difficulty of an item threshold matches a subject's ability. It is clear then that the information provided by an observation will be at a maximum at the same point. This is in accordance with the intuition that a well-designed test should have sufficient items with difficulties matching expected student abilities, and sufficient items with difficulties targeted at the pass/fail boundary. From this perspective, the information, considered as a function of ability, as defined in equation 2.120 (p. 82), is of

particular interest to test designers. A test designer can maximise the information, and thus reduce uncertainty, at critical regions of a test by removing items with difficulties in non-critical regions and adding others with difficulties in the critical regions. The associated standard errors, considered as a function of ability are defined in equation 2.122 (p. 82). The corresponding *standard error density* curves for the two datasets are given in Figure 2.6.



(a) (b)

Figure 2.6: Standard error densities for two datasets

As with the characteristic curve discussed earlier, it is possible to view the measurement model from an alternate perspective. This perspective seeks to use subjects (participants) to evaluate items, rather than using items to evaluate subjects. From this perspective, equations for the *sample* Fisher Information, and the corresponding variance and standard error, can be derived:

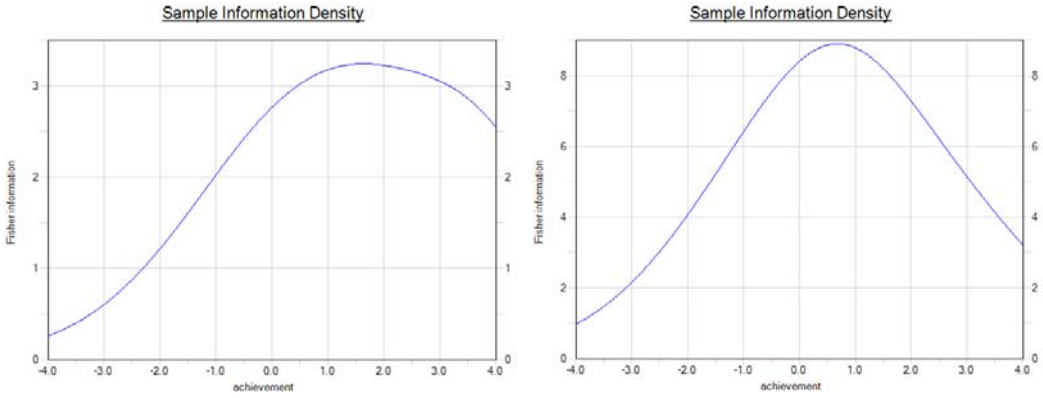
$$P_{nt}(\beta) \stackrel{\text{def}}{=} E\left(P_{ntij}\middle|\beta\right) = \frac{1}{1 + e^{-\alpha(\theta_{nt}-\beta)}} \tag{2.123}$$

$$E(I|\beta) = \alpha^2 \sum_{nt} (P_{nt}(\beta)(1 - P_{nt}(\beta))) \tag{2.124}$$

$$E(V|\beta) = \frac{1}{E(I|\beta)} \tag{2.125}$$

$$E(\sigma|\beta) = \sqrt{E(V|\beta)} \tag{2.126}$$

These equations model the information that responses in the sample give about items of a particular difficulty. Sample information density curves corresponding to the two datasets used in this section are shown in Figure 2.7.



(a) (b)

Figure 2.7: Sample information density graphs

The use of both sample and instrument information density curves is discussed in the next section.

2.5.6. Tailoring Measurement Instruments

The foregoing discussion described how the measurements of subject abilities and item threshold difficulties are established, together with the associated uncertainty in their magnitudes. Where a measurement instrument is used in a context of continuous quality improvement, the sample and instrument *characteristic* and *information density* curves can also inform refinement of an instrument for a particular purpose.

Where the purpose is educational testing, the focus is normally on producing estimates of subject ability. In general, the greater the number of items, the more accurate such ability estimates will be. However, each item represents workload for both subject and assessor, so using too many items leads to inefficiency. There is an evitable trade-off here between workload and accuracy. It can be noted that an item is most informative about a subject's ability when there is a close match between its difficulty and the subject's ability. An efficient test seeks to minimise the number of items required to produce sufficiently accurate ability estimates. The

interpretation of sufficiently accurate depends on the specific aim of the test. Where the aim is accreditation, commonly termed a “high stakes” test, it is especially important to have confidence in pass/fail decisions. Such tests should thus have sufficient items targeted near the pass/fail boundary to produce acceptable accuracy. Where the aim is diagnostic, it may be more appropriate to target items over a wider range. For both of these types of tests, interpretation may be enhanced by reporting measurements as scores or percentages. The test’s *characteristic curve* can be used to evaluate the mapping of scores to ability and the *information density curve* the uncertainty associated with ability levels. Tests may be tailored by adding items in regions where there is insufficient information and removing items from regions which have more information than is needed.

Where the purpose is on understanding items and the relationships between them, the focus is normally on producing estimates of item difficulty. In general, the greater the number of participants, the more accurate such estimates will be. However, each participant represents additional cost and workload, so using too many participants again leads to inefficiency. Once more, there is an inevitable trade-off between cost or workload, and accuracy. It can be noted that a participant is most informative about an item’s difficulty when there is a close match between ability and difficulty. An efficient study seeks to minimise the number of participants required to produce sufficiently accurate difficulty estimates. A well-targeted sampling frame can thus produce more efficient estimates of item difficulties. Where the items are criterion-referenced, the *sample characteristic curve* indicates the match between items and sample and the *sample information density* the associated uncertainty. These can therefore help evaluate the targeting of the sampling frame and the adequacy of sample size respectively. Where norm-referencing is used, the sample information density is still relevant for judging sample size adequacy, but the instrument’s characteristic curve is more appropriate for the evaluation of targeting.

2.5.7. Polytomous Model Corrections

As noted in 2.4.5, the discussion to this point has progressed under the naive assumption that the thresholds of an item are independent of each other in the same way that responses to an item are independent of other items. The treatment of information and the associated standard estimates are based on Fisher's observation that:

> "The amount of information provided by a combination of two or more independent observations is thus merely the sum of the amounts of information in each piece separately. (1925, p. 710).

However, *independence* of the observations is of fundamental importance to the treatment. If observations are assumed to be independent when in fact they are not, the information in a data set about a parameter will be overstated and the corresponding standard errors will be underestimated. It is clear that, when calibrating item thresholds from subject cases, the various subjects give independent observations. Where there is a time-series, the case to case dependency within a subject is corrected by weighting as discussed in 2.4.6.

However, for the calibration of cases from dichotomised responses to polytomous items the issue of threshold dependency needs to be addressed. The discussion following reviews the independence of item thresholds and sets out the necessary adjustments to correct for the lack of independence when estimating cases.

It can be noted that the threshold information comes from an ordinal judgement of a value against a threshold. However, it is clear that information can only come from thresholds that are actually considered by the respondent. For notation, assume there are k ordinal response categories and t $(= k - 1)$ thresholds separating these categories. Let a response to a threshold be coded as 0 if the response category is below the threshold and 1 otherwise. If the thresholds were independent, there would be 2^t possible response patterns. However, only $t + 1$ of these patterns are compatible with the assumption that the categories are ordinal, so there is a clear dependency between the threshold responses.

A response in the first category will be considered first. The response to the first threshold will be coded as 0, reflecting the judgement that the quantity of the attribute being measured is below this threshold. However, to comply with the ordinal pattern, the remaining thresholds must also be coded as 0. The information provided by the first threshold can be considered independent since there is only one observation. The coding of zero against subsequent thresholds adds no new information[3] to that associated with the judgement associated with the first statement. Accordingly, the set of thresholds summed for information should comprise only the first threshold in this case.

Similarly, with a response in the last category, the response to the last threshold will be coded as 1, reflecting the judgement that the quantity of the attribute being measured is above this threshold. To comply with the ordinal pattern, it follows that all lower thresholds must also be coded as 1. The coding of these lower thresholds as 1 adds no new information to what is known by coding the highest threshold as 1. Accordingly, the set of thresholds summed for information should comprise only the last threshold in this case.

For the more general case in which a response is in a category intermediate between the first and the last, the threshold immediately below the category will be coded as 1 and the threshold immediately above the category will be coded as 0. These will be termed the *lower* and *upper* thresholds respectively. To comply with the ordinal pattern, all thresholds below the lower threshold must also be coded as 1, and all thresholds above the upper threshold must be coded as 0. These thresholds add no new information to what is already known from the judgements associated with the lower and upper thresholds. Accordingly, in this general case,

[3] The logic here is that information comes from the uncertainty removed. If only one outcome is possible, there cannot be any information associated with the outcome.

the set of thresholds summed for information should comprise only the lower and upper thresholds.

An intuitive description of a possible response process is as follows. The process starts with the respondent choosing a candidate response category. This initial choice may be made by guessing, intuition, or even randomly. The respondent then verifies the category by checking against the lower and upper thresholds of the category. If these are consistent with the choice the process concludes and the candidate category becomes the final choice. If the test against either of the bounding thresholds fails, the candidate category is rejected and the category on the other side of the failing threshold becomes the next candidate. If a successful boundary check has already been made for the threshold on the other side of a category to the failing boundary check, the information from the earlier successful boundary check is discarded because the outcome is implicit in the outcome of the failing boundary check. The respondent continues the process by checking the boundary on the other side of the new candidate. Whether this is the process that actually occurs is speculative and would require further study. However, whether or not the response process occurs as described above, the criterion used for this work is that, for valid measurement, the "ruler" should produce the same measurement whether starting somewhere in the middle or starting at either end and working systematically from that point. Regardless of the process, it is clear that a judgement that an attribute is below a threshold entails a judgement that it is also below all higher thresholds. Likewise, a judgement that an attribute is above some threshold entails the judgement that it is also above all lower thresholds. Accordingly, information can only come from the thresholds bounding a category. Where one of the extreme categories is chosen, that is a single threshold. Where an intermediate category is chosen it is the two bounding thresholds of that category. This proposition is supported by the results of the *Monte Carle* simulations presented in Chapter 5.

In summary, the above logic identifies the thresholds for which the associated judgements provide information about case ability. The information that an item provides about the ability of a case can be determined by summing the information

associated with these thresholds. The set of thresholds to be summed for item i with k_i categories and a response in category c_i can be defined as:

$$J_{ic} \stackrel{\text{def}}{=} \begin{cases} \{1\} & \Big| \; c_i = 1 \\ \{c_i - 1, c_i\} & \Big| \; 1 < c_i < k_i \\ \{c_i - 1\} & \Big| \; c_i = k_i \end{cases} \tag{2.127}$$

The total information given by an item i about a case with ability θ for which the response is in category c_i can be defined as:

$$P_j(\theta, i) \stackrel{\text{def}}{=} \frac{1}{1 + e^{-\alpha(\theta - \beta_{ij})}} \tag{2.128}$$

$$I_{ic}(\theta) = \alpha^2 \sum_{j \in J_{ic}} P_j(\theta, i)(1 - P_j(\theta, i)) \tag{2.129}$$

When estimating the ability of a case, the response category is known and the above logic can be used directly. However, to provide the information density and standard error density curves described in 2.5.5, an estimate that is conditioned only on ability is needed. This requires a way of mapping ability to categories. The probability $\check{P}_c(\theta, i)$ that category c_i will be chosen in item i for a case with ability θ, can be expressed as:

$$\check{P}_c(\theta, i) \stackrel{\text{def}}{=} \begin{cases} 1 - P_1(\theta, i) & \Big| \; c_i = 1 \\ P_{c-1}(\theta, i) - P_c(\theta, i) & \Big| \; 1 < c_i < k_i \\ P_c(\theta, i) & \Big| \; c_i = k_i \end{cases} \tag{2.130}$$

An estimate of the information given by an item i about a case with ability θ can be achieved by using a weighted sum of equation 2.130 across response categories with weights corresponding to the probabilities that the respective categories will be chosen:

$$I_i(\theta) = \sum_{c=1}^{k_i} \check{P}_c(\theta, i) I_{ic}(\theta) \tag{2.131}$$

The total test information can then be calculated by summing this item information across all items. The corresponding variances can be calculated by taking the

reciprocals of the item and test information respectively and the associated standard errors by taking the square roots of the variances.

2.5.8. Continuity Correction

The abilities and difficulties measured fall on a continuous underlying metric. However, the use of response categories gives discrete observations. For dichotomous items, the natural scores of cases and items are sufficient statistics. It follows that there are only as many possible output values as there are scores available. Each such value thus necessarily represents a range of values in the underlying continuum. An estimate of this range can be carried out as follows.

Equation 2.110 (p. 74), which gives the expected case score given ability, is strictly monotonic. An inverse, the expected ability given a score $E(\theta|Score)$, can therefore be estimated by bisection. The range of abilities associated with a case with a natural score of s can be defined as between $E\big(\theta|(s-0.5)\big)$ and $E\big(\theta|(s+0.5)\big)$. This range can be treated as approximately uniform in the region of s so the equivalent standard deviation can be defined as:

$$\sigma_s \stackrel{\text{def}}{=} \frac{E\big(\theta|(s+0.5)\big) - E\big(\theta|(s-0.5)\big)}{\sqrt{12}} \tag{2.132}$$

A second, and more subtle, effect of the discrete nature of the scores is what is termed *inflation* in this work. The essential logic of the measurement model as set out is to estimate a continuous odds ratio by discrete counts, as set out in equation 2.14 (p. 34). Because of this discrete nature, however, the relative error in the estimate is greater towards the ends of a scale as one or other of the counts involved approaches zero. In the limit, a count of zero maps to an infinite estimate; for example, the difference between a score of zero and a score of 1 is infinite. This is a necessary consequence of any attempt to map an infinite range of abilities or difficulties onto a finite scale and it leads to the familiar ogival shape of test characteristic curves. Nevertheless, this is not problematic in most practical situations. One side effect of this difference in relative error is that the unconditional likelihood procedure is slightly biased, tending to stretch the range of

the scale. Wright and Stone (1999, p. 132) give an approximate correction factor for this inflation of $(K-1)/K$, where K is the number of items and suggest a similar correction for items based on the number of cases. This correction tends to be progressively inaccurate with smaller numbers of cases and items (Jansen, Van den Wollenberg, & Wierda, 1988). The authors question the assumptions used by Wright and Stone in the derivation and conclude "Generally, however, even within the framework of these two assumptions, the correction factor is of a much more complicated nature" (p. 305). However, the authors do not give an alternative correction, but call for further simulations and research. Since accuracy of the estimates with relatively small samples is important to the current project, an alternative correction is derived below.

From *true score theory*, the relation between the variance of an observed score (X) and the true score (T) is given by the reliability of an instrument:

$$r_{xx} = \frac{T}{X} \tag{2.133}$$

In the presence of measurement error, the reliability (r_{xx}) will be less than 1 and thus the true score will be less than the observed score, given that both are zero centred and expressed in the same units. This can be viewed as inflation, when mapping from true scores to observed, or *regression of the mean* when mapping from observed scores to true scores. An alternate correction for inflation can therefore be applied by using a factor of r_{xx} rather than $((K-1)/K)$, where r_{xx} is an appropriate reliability coefficient. The Rasch Person Separation Reliability and Item Separation Reliability have been found to be useful for this purpose.

Whether or not a correction for inflation is needed depends on the context in which measurement is undertaken. Since the model is only defined up to a linear transformation, inflation is not problematic unless item or case estimates need to be stable across different contexts. This can occur, for example, when comparing results from a small sample size to a larger sample, adding additional items to a small test, or test equating. For the present work, it is important for estimates to remain stable as the unfolding of the time series involves progressively more items.

However, in other contexts, it might be more important to preserve compatibility with estimates produced by other software, which might implement no correction, or the Wright and Stone correction.

Accordingly, the software allows a choice between no correction, the Wright and Stone correction and the reliability correction.

2.6. SUMMARY

This chapter has presented in detail the measurement model used in this project. The conceptual measurement model is shown in Figure 2.8.

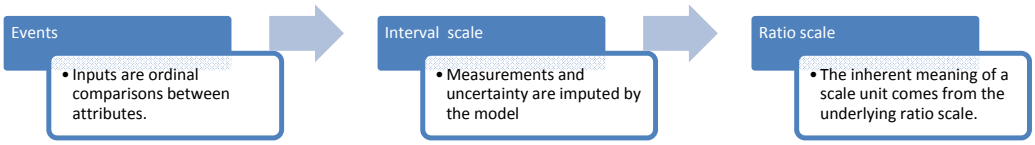


Figure 2.8: Conceptual model of measurement

The model produces linear measurement when the assumptions of the model hold. It also produces formal estimates of uncertainty to accompany each measurement. It places measurements of subjects' abilities and item threshold difficulties on a common metric and allows dichotomous and polytomous items to be freely mixed in a scale and share a common interpretation. The use of a common metric allows meaning of subject ability measures to be constructed by reference to item difficulties and vice-versa.

Formulae for the associated information, variance and standard errors have been presented. These statistics are traditionally considered to be characteristics of the overall measurement instrument. However, it is clear that they vary over the operating range of the instrument. Since a goal of the current project is to determine quality control statistics for each measurement, formulae have also been presented for information, variance and standard errors, conditional on the measurement.

The information density and characteristic curves support the process of targeting and tailoring measurement for a specific purpose. This allows the model to be used

in a process of continuous improvement whereby the results of each measurement can inform the improvement of an instrument for future use. The model is supported by a theoretical framework which is summarised in Table 2.9.

Table 2.9: A theoretical framework for measurement

Concept	Key sources
Epistemology	Kant (1783/2004)
Interval and ratio scales	Stevens (1946)
Quantitative structure	Hölder (1901/1996), Michell (1997)
Conjoint measurement	Luce and Tukey (1964)
Associated uncertainty	R.A. Fisher (1925), Taylor and Kuyatt (1994)
Measurement of perceptions	Weber (1834/1978), Fechner (1860/1966), Stevens (1957)
Measurement models	Rasch (1960/1980), Samejima (1969, 1972), Andrich (1978), Masters (1982), Muraki (1983, 1990)
Proof of existence	Fischer (1987)

However, the assumption has been made throughout this chapter that the assumptions of the measurement model hold. It was noted at the start of the chapter that, at the most fundamental level, measurement is a hypothesis rather than a process. No model can produce accurate objective measurement from inappropriate input data. The next chapter investigates measurement from this perspective of *measurement as hypothesis*.

Chapter 3.

MEASUREMENT AS HYPOTHESIS

The measurement model presented in the previous chapter set out the logic of producing interval measurements from ordinal data, together with a statement of associated measurement uncertainty, expressed as standard errors. However, a measurement model cannot produce valid or useful measurement from arbitrary data. When discussing his *Difference Engine no 1* in his autobiography, Charles Babbage (1864) recalled:

> On two occasions I have been asked, "If you put into the machine wrong figures, will the right answers come out?" In one case a member of the Upper, and in the other a member of the Lower, House put this question. I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question (p. 67).

The quality of output measurements depends both on the adequacy and quality of the input data and on the correctness of the theory underpinning the construction of the measurement instrument. From a scientific perspective (Popper, 1935/2005), measurement should thus be treated as a working hypothesis which one actively seeks to disconfirm. If any assumption is disconfirmed, the measurement hypothesis is rejected and one should have little confidence in any output measurements. If one fails to disconfirm the various assumptions made, the hypothesis remains plausible and one may proceed, with appropriate caution, to use the output measurements.

The focus of this chapter is on the tests that are carried out automatically as part of the measurement process, and on various quality metrics that are used to evaluate and build confidence in the measurements produced. However, these can only give confidence that *something* is being measured, not what that something is. Accordingly, validity cannot be confirmed by the tests and metrics discussed in this chapter, although it may be readily disconfirmed. The tests thus provide a necessary, but not sufficient, condition for overall measurement validity. Further

tests to establish the construct validity of the *self-efficacy* and *challenge* constructs are introduced in Chapter Seven. This split follows Loevinger's suggestion that:

> Evidence for construct validity can be broken down into evidence that the test measures something systematic and evidence for the particular interpretation of what it measures (1957, p. 685).

The evidence that the measurement model is measuring something systematic comprises both formal hypothesis tests and model fit statistics. The hypothesis tests give evidence that the assumptions of the model hold. The fit statistics give assurance that the model appropriately represents the information in the dataset. However, it can be noted that no model will ever fit the data *perfectly;* the key question is whether the model fits the data *sufficiently* to produce useful measurement. Three ideas arise from this perspective. First, it may be possible to quantify the extent to which a measurement is potentially affected by any misfit and to adjust the stated margin of error accordingly. This would allow correct, and hopefully useful, measurement to proceed.

Second, the model can be used in a process of continuous improvement. In each administration of the measurement instrument, poorly fitting items can be investigated and reworded or replaced thus, hopefully, improving quality in future administration. From this standpoint, the investigation of model fit is a never-ending process; each administration is both a test of the theory underpinning the model, and a source of information for on-going improvement. Importantly, since the items are presumably based on a theoretical understanding of the construct being measured (content validity), those that misfit suggest an issue with the theoretical construct or its interpretation. Investigation of these items is thus likely to be fruitful in leading to improved theoretical understanding of the construct. A similar argument can be made for misfit of subjects. However, although the latter can and should be investigated, in practice there is less scope for on-going improvement because, in most situations, new subjects will be used in each new administration.

Third, a measurement exercise can still be useful, even if some measurements fail. Thus, for example, even if some of the items or cases in a set fail, the remaining measurements may still be valid. Accordingly, one might remove from the analysis some of the subjects or items that do not fit the model, allowing improved measurement of those remaining. This should be done with caution, however, because these are not *missing at random* (Little & Rubin, 1987) and thus removal threatens the validity of the measurements. Specifically, removing items could affect content validity, and removing subjects could affect any intended generalisation from the sample. Nevertheless removal is a viable option if these concerns are addressed.

The central assumptions of the measurement model are now briefly summarised. First, the model assumes that there is a continuous latent trait or construct which governs the responses made to items, that subjects possess a quantity of this trait termed *ability* which can be represented as a location on a linear metric corresponding to this trait, that items have a *difficulty* that can be represented as a location on the same metric, and that the responses made by subjects to items depend only on the subjects' abilities and the items' difficulties. Second, it is assumed that the higher response categories represent more of the attribute being measured (i.e. an ordinal response pattern). Third, it is assumed that there is some error in determining the responses, but that this error is random, or at least unrelated to subject or item characteristics. Fourth, it is assumed that, for any item, the probability of success is non-decreasing as subject ability increases and that, for any subject, the probability of success on an item is non-increasing as item difficulty increases. These two conditions define *double monotonicity* (Mokken, 1971). Finally, it is assumed that the response made by a subject to an item is independent of other items and that subjects respond to items independently of other subjects; this is termed *local independence* or conditional independence.

For the hypotheses tested in this chapter, it is assumed that the instrument has been constructed with the above points in mind and thus the approach taken is confirmatory rather than exploratory. That is, it is assumed that the construct "makes sense" conceptually and is believed or intended to have the characteristics

described above. Consequently, the focus is on the formal tests that can be carried out to verify these assumed characteristics, and to identify specific subjects or items for which the characteristics might not hold. The tests are expressed as hypotheses to be tested, and are summarised in Table 3.1. Most of these tests are carried out both at the overall instrument level, and at the level of individual items and subjects. The tests at the subject and item levels are intended to help identify the cause of any failure and thus support a process of continuous improvement. The tests are also supported by a number of statistics and indices that can be used to help judge the impact of the failure of any hypothesis.

Table 3.1: The measurement hypotheses that are tested automatically.

	Hypothesis	Domain
H1	The construct being measured is quantifiable	Theoretical
H2	Response categories are ordinal	Theoretical
H3	The construct is unidimensional	Theoretical
H4	There is no differential item functioning	Theoretical
H5	There is no subject response set	Theoretical
H6	There is local (conditional) independence	Theoretical
H7	The measurement model converges correctly	Technical
H8	Data are adequate for measurement	Technical
H9	There is a common metric for all measurements	Technical
H10	Response patterns are reproducible from measurements	Model Fit
H11	There are no more outliers than expected	Model Fit
H12	Fit statistics accord with theoretical expectations	Model Fit
H13	There is adequate reliability for useful measurement	Usefulness

The first six hypotheses relate to theoretical aspects of the assumptions of the model. First, the assumption that ability and difficulty can be placed on a linear metric that corresponds to a continuous latent trait is tested by the hypothesis that the trait has a *quantifiable* structure (H1). The assumption that higher response categories represent more of the attribute being measured is tested by the hypothesis that the response pattern is *ordinal* (H2). The assumption that responses

depend only on ability and difficulty is tested by the hypotheses that the latent trait is *unidimensional* (H3), that there is no *differential item functioning* (H4), and that there is no *response set* (H5). The assumption that the response made by a subject to an item is independent of other items and that subjects respond to items independently of other subjects is tested by the hypothesis that there is local (or conditional) independence (H6). The next three hypotheses relate to the technical software implementation. These are that the measurement model converges (H7), that there are adequate data for measurement (H8), and that a common metric is achieved in the measurement process (H9). Another three relate to the proposition that the model adequately represents the input data. These are that response patterns are reproducible from the model (H10), that there are no more outliers than expected (H11), and that the various fit statistics accord with theoretical expectations (H12). Finally, the hypothesis that there is sufficient reliability for useful measurement (H13) is introduced. Taken together, these 13 hypotheses provide a comprehensive framework for the evaluation of successful measurement.

The remainder of this chapter is organised as follows. The first 12 sections address each of the hypotheses in turn. Finally, the overall conceptual framework is presented, the impact of the failure of any hypothesis is discussed and the main points of the chapter are summarised.

3.1. QUANTIFIABLE STRUCTURE (HYPOTHESIS 1)

The measurement model assumes that ability and difficulty can be placed on a linear metric that corresponds to a continuous latent trait. This is tested by the first hypothesis:

- H1: The construct being measured is quantifiable

The formal axioms under which a construct is quantifiable by a measurement process are given by Luce and Tukey (1964). Karabatsos (2001) analysed a number of measurement models for compliance with these axioms and his main points are summarised here, using the notation of the measurement model.

Consider a matrix in which columns represent item thresholds (ij) and rows represent subject cases (nt). Assume that the columns are arranged in non-decreasing difficulty order from left to right, that the rows are in non-increasing order of ability from top to bottom and that each cell contains the response probability P_{ntij} as defined in this work. It is easy to verify that for the measurement model used in this work, each row has non-increasing entries moving from left to right and each column has non-increasing entries moving from top to bottom. It follows that the probabilities are doubly monotonic (Mokken, 1971). This is sufficient to construct ordinal measurement models such as ISOP (Scheiblechner, 1995) but is not sufficient for linear measurement. However, if each cell P_{ntij} is replaced by the monotone transformation $log\,(P_{ntij}/(1-P_{ntij}))$, double monotonicity is preserved, but now the difference $\alpha(\theta_1-\beta_1)-\alpha(\theta_1-\beta_2)=\alpha(\beta_2-\beta_1)$ between any pair of columns is identical on each row and similarly the difference between any two rows $\alpha(\theta_1-\beta_1)-\alpha(\theta_2-\beta_1)=\alpha(\theta_1-\theta_2)$ is identical on each column. Thus the model also conforms to the row independence and column independence axioms which, together with double monotonicity, are required for conjoint measurement. This is sufficient to establish linear measurement. It can be noted that such row and column independence holds for the Rasch model (in which $\alpha=1$) and also for the measurement model used in this work (in which α is constant), but does not hold for other models, such as the two parameter logistic model for which α can vary from item to item, or where it may vary from subject to subject.

It follows that the specified measurement model conforms to the formal requirements of the conjoint measurement axioms. It can be argued, therefore, that conformance with the model, as evidenced by model fit statistics, provides sufficient evidence that an attribute is quantifiable. Indeed, Perline, Wright, and Wainer (1979) conclude that the Rasch model is a practical realisation of conjoint measurement and that, in consequence, fit statistics can be used as evidence of conjoint measurement. This perspective is not unreasonable, but Karabatsos (2001) argues that there is an inherent logical weakness in using parameters inferred from data to test whether the data conform to the parameters, and that stronger

evidence would be supplied if the conjoint measurement axioms were tested independently of the model. He then describes an approach to testing the assumptions that any does not use parameters from the measurement model. Following his approach, the hypothesis that the construct is quantifiable is tested in this thesis without reference to the model parameters as follows. First, subject cases and item thresholds are arranged into rank order by scores and grouped into quantiles. The observed subject percentage success rate is recorded for each of these quantiles in a two dimensional table. The number of quantiles used depends on the number of subjects and items. Where the number is 100 or more, deciles are used. Where it is less than 100, a lesser number of categories is used to achieve a balance between the number of units and the number of observations per unit, down to a minimum of 5 (pentiles). To ensure that all cells have non zero entries, adjacent quantiles are merged if there are no responses in any cell, thus resulting in a smaller number of cells in total. The total number of cells actually used is denoted *K* herein. These success rates form the empirical person and item response functions (characteristic curves) of the dataset, which together define a response surface. A Marcov Chain Monte Carlo (MCMC) bootstrap process is then used to estimate the expected distribution of these success rate parameters under the constraints of the conjoint measurement axioms. This process uses 1,000 burn-in cycles, followed by 1,000 samples to determine the confidence limits of each parameter. The observed success rates of the empirical response functions are then tested against this theoretical distribution and any falling outside the 95% confidence limits are marked as failures. Some failures are expected even when the data conforms to the axioms. For example, if there are 100 parameters, then five (5%) of these can be expected, on average, to fall outside 95% confidence limits.

Formally, the hypothesis is tested as follows. An exact binomial test is used to calculate the probability of obtaining the observed number of failures, or more, in a set of *K* parameters drawn from a population in which 5% are failures. If the probability is less than the significance level (usually 0.05), the measurement hypothesis is rejected.

Interpretation of any failures is facilitated by a visualisation of the empirical response surface. An example of such a visualisation produced by the software developed for this thesis is given in Figure 3.1. This figure shows the empirical response surface for the PISA 2009 maths dataset for New Zealand (OECD, 2010). In this hypothesis test, none of the 50 parameter estimates were outside the 95% confidence limits.

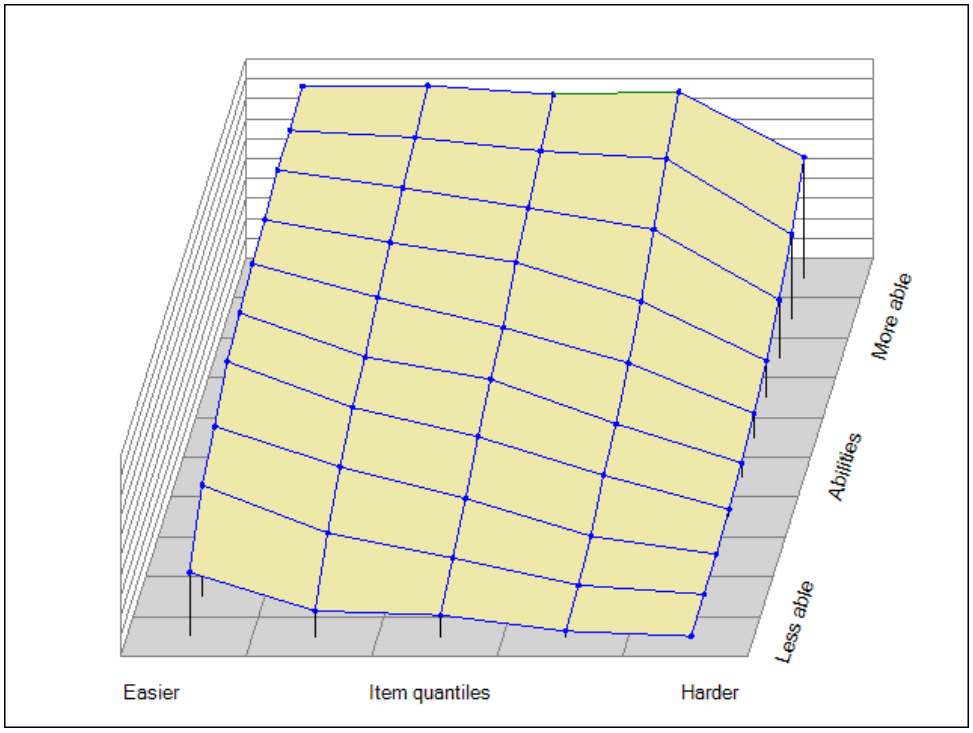


Figure 3.1: Empirical response surface of the NZ PISA 2009 maths dataset

In the figure, the parameter estimates falling within the imputed confidence limits are indicated with a blue dot and those outside with a red dot. Lines are also shown between each estimate. These lines are coloured blue if the slope is empirically strictly monotonic, green if not empirically monotonic but within confidence limits, and red otherwise.

In summary, this section has presented a formal test of the hypothesis that the dataset meets the axioms required for the construct to be quantifiable. Additional evidence in support of the hypothesis that the construct is quantifiable is given by the model fit statistics described later in this chapter.

Rejection of this hypothesis suggests that the underlying assumption of double monotonicity does not hold in the dataset. Accordingly, the measurement process is classified as failed and output measurements should not be used without further investigation. If subsequent investigation is able to attribute the failure to specific items or subjects, these can be removed from the model and the measurement exercise repeated. Identification of the specific sources of the violation can be guided by visual inspection of the empirical response surface and by the various model fit statistics. If visual inspection of the empirical response surface suggests that violation is limited to specific regions, such as the most able students, then it may be possible to proceed to use the measurements, subject to the caution that little faith should be placed in the measurements produced for this region.

3.2. ORDINALITY (HYPOTHESIS 2)

In addition to testing quantifiable structure, the hypothesis test in the previous section gives some assurance of consistent item and subject ordering. However, the test is based on allocated scores and the implicit assumption was made that higher scores are allocated to higher abilities. Given the confirmatory approach taken in this chapter, this is a reasonable assumption, but in a practical setting it is always possible for this assumption to be incorrect on some items, perhaps because of ambiguous wording, or even simple error. This section introduces a formal test of the hypothesis:

- H2: Response categories are ordinal

The inputs to the measurement model comprise ordinal judgements. These ordinal judgements are encoded by assigning observations to categories in an assumed ordinal sequence. Ideally, these categories are ordinal *by construction* so that there is little doubt about the ordinal nature. An example of this from the current project is the sequence: not started, started, completed. In other more general settings, there might be a difference in interpretation between the person defining the categories and those scoring or responding to the item. Variation of interpretation could be ascribed to differing subjective values or conceptions of word meanings. An example might be a classification in the affective domain. Those familiar with

the domain might use the following sequence: receiving, responding, valuing, organising, characterising (Krathwohl, Bloom, & Masia, 1964). This might seem a natural increasing sequence to those who have studied the domain, but others might choose a different sequence.

The hypothesis that the response categories are ordinal is tested as follows. First, for each category in each item, the number of responses, the mean imputed ability of subjects responding in that category, and the standard deviation of these ability estimates are collected. Under the (confirmatory) assumption that most responses are coded correctly, and thus that the estimated subject abilities are not too incorrect, these mean subject abilities associated with the categories should form an increasing sequence within each item when the response categories are indeed ordinal. Accordingly, if these mean abilities form an increasing, or at least non-decreasing, sequence, the ordinal nature of the categories remains plausible. However, if the mean ability in any category is less than that of a lower category, there is possible evidence of disordinality.

This is tested formally by a two-sample t-test of the subject abilities responding in each adjacent category. If this is significant (at the $p < .1$ level), an appropriate warning diagnostic is given, together with the relevant statistics. The statistics include the difference in means, the number of cases involved, the t-statistic and the p level.

The level of $p < .1$ has been chosen, rather than the more usual $p < .05$, to increase the likelihood of identifying even mild disordinality for further investigation. The proposition made here is that any disordinality should be rare in a well-constructed scale and, consequently, even a mild suspicion of disordinality should be carefully investigated. Even if such an investigation does not find disorder in the categories, there is nevertheless a clear indication that there is insufficient separation between the categories. This suggests that perhaps there are too many categories, which could consequently create the illusion that there is more accuracy and precision in the scale than is actually there.

The overall test of the hypothesis of ordinal response categories is defined formally as follows. First, the categories on either side of each threshold of each item are compared. If any disordinality is diagnosed, the threshold is classified as failing. The number of thresholds tested and the number failing the disordinality test are summed across all items. The expected number of failures under the null hypothesis of no disordinality is 10% of N, where N is the total number of thresholds tested. An exact binomial test is used to calculate the probability of obtaining the observed number of failures, or more, in a set of N drawn from a population in which 10% are failures. If this calculated probability is less than the significance level (0.05), the hypothesis of ordinal response categories is rejected. As intimated above, failure of the hypothesis at the overall level should be rare and would give a clear signal for a detailed review of the measurement instrument.

In summary, this section has presented a formal test of the hypothesis that response categories are ordinal. This is tested both at the overall instrument level and at the individual item level. If the hypothesis is rejected at the overall instrument level the measurement process is considered to have failed and output measurements should not be used without further investigation. Occasional failure at the item threshold level should be investigated but is not a cause for major concern because some failures are expected even when the assumption of ordinality holds.

3.3. DIMENSIONALITY (HYPOTHESIS 3)

The metaphor used for the measurement model is a ruler (Figure 3.2) that represents the latent trait measured.

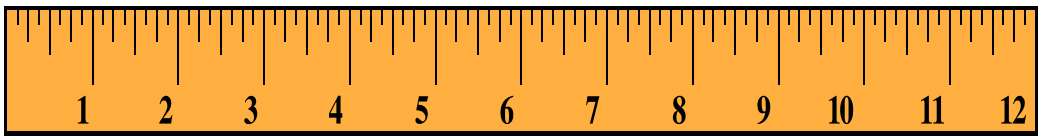


Figure 3.2: A ruler as a metaphor for measurement

Objects, both cases and items, are assigned to locations on this ruler corresponding to their estimated locations on the measured trait. In common use, a ruler is marked in conventional units such as centimetres or inches. No such conventional

units exist, as yet, for psychometric measurement. In the absence of such agreed standard units, the ruler is arbitrarily marked off in logits or some multiple of a logit. Similarly, the origin is arbitrarily chosen relative to properties of cases, items, or some external criterion. Despite this indeterminacy of scale and origin, both the ordering of the objects, and the relative distances between the objects, remain consistent under any linear transformation that changes the scale and origin. Thus, the measurement remains interpretable as long as some frame of reference is available, such as characterising subject abilities with the difficulties of items at about their ability level.

As implied by this metaphor, measurement is conceptually always one dimensional. However, from a practical perspective, it is not possible to construct a set of items that align *perfectly* with an intended latent trait. Thus, all real data are inherently multidimensional to some extent. If multidimensionality is present then the output measurements will be a composite of the various input factors. It follows that multidimensionality is primarily a threat to interpretation, and thus validity. However, it need not be a threat. As Cronbach (1951) points out:

> For a test to be interpretable, however, it is not essential that all items be factorially similar. What is required is that a large proportion of the test variance be attributable to the principal factor running through the test (p. 320).

There are two points of view on the extent to which multidimensionality affects interpretation and validity. These may be broadly labelled as associated with the *factor analysis* tradition and the *measurement* tradition, respectively. In the factor analysis tradition, it is believed that many constructs are inherently multidimensional and that incorporating this perspective in a test allows the full range of a construct to be captured. It is also believed that the same underlying construct can be manifested in different ways. Thus, using a composite measure allows a meaningful overall index to be formed. It also provides a framework in which the relative contribution of individual factors can be investigated. In the measurement tradition, it is argued that such composite measures confound measurement and that it is better to measure each individual factor separately, and

only then investigate any association between the factors, or form a desired composite measure from a known combination of factors. Whether the items and observations contributing to measurement should be one-dimensional or not thus depends on the perspective taken when the construct was defined. To accommodate both perspectives, the software provides an option as to whether a strict interpretation of unidimensionality should be enforced.

For the rest of this section, the assumption is made that the construct is intended to be unidimensional and describes the tests of the hypothesis:

- H3: The construct is unidimensional

Hattie (1985) gives the following definition: "Unidimensionality can be rigorously defined as the existence of one latent trait underlying the set of items" (p. 152). This implies that responses by subjects to items should depend only on the ability of the subject and the difficulty of the item. The approach taken follows the measurement tradition and tests the homogeneity of the scale. This is investigated from two perspectives: item and subject. The traditional item perspective identifies how well items relate to the construct defined by the overall scale. The subject perspective identifies the extent to which there is a shared interpretation of the scale items among subjects. The overall logic in each case is to identify the number of *feasible* components. If this is zero, the sub-hypothesis is rejected because of insufficient scalability. If it is two or more, the sub-hypothesis is rejected because of multidimensionality. If it is exactly one, the sub-hypothesis is supported. To summarise, there are two sub-hypotheses:

- H3A: Homogeneity analysis by item identifies one component.
- H3B: Homogeneity analysis by subject identifies one component.

The overall hypothesis of unidimensionality is accepted if both of these sub-hypotheses are accepted, and is rejected if either of them is rejected. The logic of the homogeneity test is presented in detail for the item sub-hypothesis. The logic of the subject sub-hypothesis follows by symmetry.

Loevinger (1948) defined three coefficients of homogeneity for item-item homogeneity (H_{ij}), item-test homogeneity (H_i) and overall test homogeneity (H). Her item-item homogeneity (H_{ij}) is equivalent to the Pearson *phi* correlation coefficient divided by its maximum possible value for the observed marginal totals. Mokken (1971) presented an equivalent set of coefficients in a formulation that is easier to generalise across arbitrary groupings. The coefficients are conceptually equivalent and indeed the numerical value of H_{ij} is the same for both formulations. Mokken's definitions are presented below:

$$V_{ij} \stackrel{\text{def}}{=} \begin{cases} \pi_i(1-\pi_j);\ \pi_i > \pi_j \\ \pi_j(1-\pi_i);\ \pi_i \leq \pi_j \end{cases}$$

$$H_{ij} = \frac{\left(\pi_{ij} - \pi_i\pi_j\right)}{V_{ij}}$$

$$H_i = \frac{\sum_{j\in g; j\neq i}\left(\pi_{ij} - \pi_i\pi_j\right)}{\sum_{j\in g; j\neq i} V_{ij}}$$

$$H = \frac{\sum_{i\in g}\sum_{j\in g; j\neq i}\left(\pi_{ij} - \pi_i\pi_j\right)}{\sum_{i\in g}\sum_{j\in g; j\neq i} V_{ij}}$$

The notation here is that of a 2 x 2 contingency table, where π_x represents the proportion of observations in which variable x is a success; π_{xy} is the proportion of observations in which both x and y are successes; and g is an arbitrary set of items. If the last is the entire test, H_i and H refer to the test overall; if it is a subset of items, they define equivalent coefficients for the subset.

Mokken and Lewis (1982) provide an accessible summary of the use of these coefficients for a scale building procedure based on homogeneity. Their procedure is briefly summarised here. First, a minimum level *c* of homogeneity is chosen; the authors suggest a value of .3 is appropriate. Next, the item-item homogeneity (H_{ij}) of all item pairs is calculated and the pair with the highest value is identified. If this is below the minimum criterion, scale building is considered to have failed and no scale is produced. If the criterion is met, the two items are selected for the scale and the remaining items placed into a pool of candidate items that may potentially

extend the scale. Next, a recursive process is used to extend the scale by adding suitable items from the pool. In each step, taking g as the set of items already selected, the item test homogeneity (H_i) is calculated for each pool item and the item with the highest value is identified. If this meets the minimum criterion, it is transferred from the pool to the scale and the recursive process continues.

If a scale has been created and some unselected items remain in the pool, the process is repeated on these unselected items to determine if a homogeneous scale can be built corresponding to a second component. If this is successful, the process is repeated to determine a third component, and so on, until either all items have been allocated to a component or no further homogeneous components can be formed.

For each component, an overall coefficient of homogeneity (H) can be calculated. Mokken and Lewis (1982, p. 422) give the following guidelines for interpretation of the *scalability* of the scale based on this coefficient:

- *Not scalable* if H is less than 0.3
- A *weak* scale if H is at least 0.3 but less than 0.4
- A *medium* scale if H is at least 0.4 but less than 0.3
- A *strong* scale if H is at least 0.5

The process described above illustrates how the number of feasible components can be identified. Since the approach taken in this chapter is confirmatory, a slightly modified process is used. In this modified process the defined scale is taken as the starting point and all items for which the item-test homogeneity (H_i) is below 0.3 are removed and formed into the candidate item pool. The above process is then carried out on the pool to determine if additional components can be formed.

Although the discussion above has focussed on items, it is clear that an equivalent set of coefficients can be defined for subjects and an equivalent scale building process can be followed constructed. Thus the homogeneity analysis identifies the number of feasible components for the sub-hypotheses of both items (H3A) and

subjects (H3B). In each case, the sub-hypothesis is rejected if the number of feasible components found is not exactly one.

When the hypothesis of unidimensionality is rejected the factor associated with each item and each subject is identified. Additionally, when the option to apply a strict interpretation of unidimensionality has been chosen, a procedure is carried out to minimise the impact of multidimensionality on the measurements produced. This involves re-running the measurement model with additional dimensionality constraints, whereby subjects are measured using only responses to items associated with the first item factor and items are measured using only responses made by subjects associated with the first sample factor. The effect of this is that the subjects and items associated with the respective first factors will be given the same measurements and standard errors that would have been imputed had the off-target subjects and items been removed from the model. However, retaining the off-target subjects and items allows them to be assigned measurements in the same metric as the on-target objects.

To assist in the interpretation of dimensionality, several conventional Principal Components Analyses, based on correlation matrices, are also carried out. Although these do not form part of the formal hypothesis tests, they can be used to build insights into the structure of the dataset. However, when interpreting the analyses, it is important to distinguish between components and dimensions. Consider, for example, the monotonic *one dimensional* function $y = x + x^3$, which has two components. A Pearson correlation coefficient models this relationship with a linear function and assumes a bivariate normal distribution (which is clearly impossible for this function). A Principal Components Analysis (PCA) is simply a rotation of basis space designed to support analysis of such linear functions. Because the relationship between scores and underlying linear measurements is necessarily non-linear, a number of artefacts will result from the PCA. For example, *horseshoe* or *arch* effects appear when the mean values of items fall in a wide range (Podani & Miklos, 2002). Indeed, it has been long known that there is an interaction between item difficulty levels and the number of components identified by PCA. For example, Ferguson (1941) found:

> In general, the greater the number of degrees of difficulty among the items in a test or among the tests in a battery, the higher the rank of the matrix of inter-correlations; that is differences in difficulty are represented in the factorial configuration as additional factors (p. 323).

It is possible to interpret such factors as real. For example, Guilford (1941) reported:

> A factor analysis of the ten sub-tests of the Seashore test of pitch discrimination [of 0.5 to 30 cycles per second differences] revealed that more than one ability is involved. One factor, which accounted for the greater share of the variances, had loadings that decreased systematically with increasing difficulty [from 30 to 0.5 cps differences]. A second factor had strongest loadings among the more difficult items with frequency differences of 2 to 5 cps. A third had strongest loadings at differences of 5 to 12 cps. No explanation for the three factors is apparent, *but the hypothesis is accepted that they represent distinct abilities* [emphasis added]. In tests so homogeneous as to content and form, where a single common factor might well have been expected, the appearance of additional common factors emphasizes the importance of considering the difficulty level of test items, both in the attempt to interpret new factors and in the practice of testing. The same kind of item may measure different abilities according as it is easy or difficult for the individuals to whom it is applied (p. 67).

However, it is more reasonable (and parsimonious) to interpret the effect as a non-linear relationship. In general, factor analysis works best when there is a narrow range of item difficulties, but measurement tests are more informative when there is a wide range of item difficulties. In a simulation study comparing Rasch fit statistics and PCA for the diagnosis of various levels of multidimensionality, Smith (1996) concludes:

> If one has no idea of the definition of the underlying variable the rating scale is attempting to assess and there is a real possibility that the factors measured by various items are uncorrelated, then the factor analytic approach might be appropriate. However, when there is a reasonable assumption that the items

> are measuring the same thing or, if not, that the two factors might be highly correlated, then the best approach is to use the Rasch item fit method. (p. 39)

Despite these difficulties, Principal Components Analysis is widely used to assess dimensionality. The goal for the present work is to determine the number of feasible components in the dataset, where feasible is defined as significantly greater than that expected from chance variation. With a unidimensional scale, the hope is that such an analysis will show a single dominant component and the remainder at "noise" levels. Since PCA is just a rotation of basis axes, it will in most cases, show as many components as there are variables. Thus, evaluation of significant components is based on the magnitude of the eigenvalues.

There is no consensus in the literature as to how this should be done and several approaches have been proposed. A common criterion is to retain components with an eigenvalue of at least 1. This criterion was suggested by Guttman (1954, p. 153) and popularised by Kaiser (1960). This criterion is often interpreted as meaning that a component explains at least as much variability as an average variable (Bryman & Hardy, 2009, p. 29). A more formal theoretical justification is that this criterion corresponds to non-negative reliability (Kaiser, 1960, p. 145). Cattell (1966) suggested that the Guttman-Kaiser criterion may retain too many factors when there are many variables (and possibly too few when there are few variables) and suggested visual inspection of a "scree" plot to determine a point of discontinuity. Horn (1965) pointed out that the Guttman-Kaiser criterion assumes that there is no measurement error. He demonstrated that the first n/2 eigenvalues of a real observed correlation matrix are inflated by sampling error and least squares bias. Consequently, the criterion for a correlation matrix derived from real data should be greater than 1. He suggested a *parallel analysis* based on generating correlation matrices of the same dimensions from random variables to establish the critical values. Frontier (1976) suggested comparing eigenvalues to those expected under a *broken-stick* distribution. If a stick is broken randomly into k pieces, and the pieces arranged in order of decreasing size, the expected relative size S_i of piece i is:

$$S_i = \frac{1}{k}\sum_{z=0}^{k-i}\frac{1}{k-z}$$

Jackson (1993) compared ten suggested criteria on simulated datasets and concluded that the most promising were the broken stick model and the bootstrapped eigenvalue-eigenvector method. For the broken stick model, he noted:

> The broken stick method correctly assessed the dimensionality of the data matrices. […] This method provided a good combination of simplicity of calculation and accurate evaluation of dimensionality relative to the other statistical methods (p. 2211).

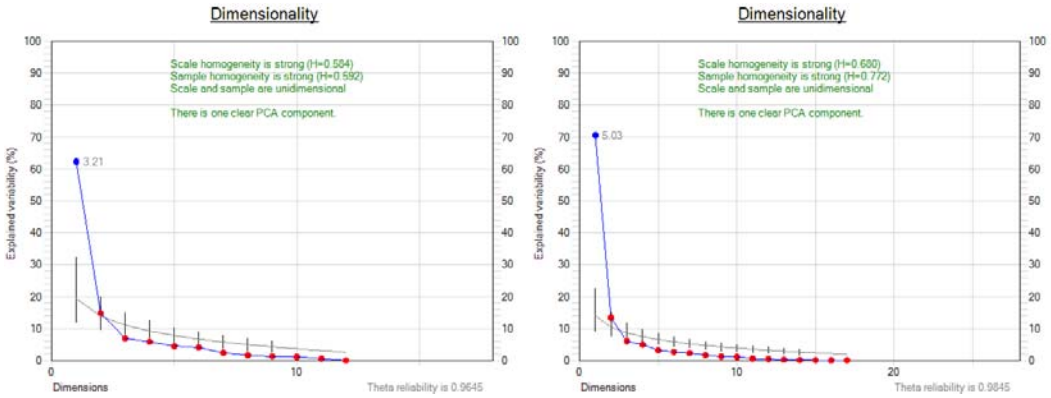
It can be seen that Horn's parallel analysis and the broken-stick approach are essentially equivalent. The shared idea is to estimate the expected distribution of eigenvalues under the assumption of the analysis of random data and then compare the observed eigenvalues to this expected distribution. The approach taken herein is based on the broken-stick model and the process is operationalized as follows. First, 95% confidence intervals for the expected eigenvalues are estimated by a bootstrap process. This involves generating 2000 empirical samples from a broken-stick distribution of the appropriate size. Each empirical sample is constructed by drawing k – 1 (where k is the number of variables) samples from a uniform distribution, using these as cut points to break the stick into k pieces, and then sorting the pieces into descending order. The confidence interval for each eigenvalue is then estimated by identifying the 2.5% and 97.5% percentiles from the sample. Next, a component is classified as *feasible* if the magnitude of the observed eigenvalue is greater than the upper limit of the confidence interval associated with the corresponding theoretical component under random simulation. The number of feasible components is defined as the number of consecutive feasible components, starting with the first component.

Three separate analyses are run. The first analysis is based on the Pearson correlation of scores between pairs of items across subject cases. This establishes the extent to which the items tap into a common underlying construct. The second

analysis uses, for each pair of subjects, the average of the correlation of scores across items between pairs of the cases associated with the subjects. This is used to verify the existence of a dominant dimension in the subjects which, in turn, identifies the extent to which the subjects have a shared understanding of the items. Both aspects are required for validity. The third analysis is based on the Pearson correlation between pairs of items of the residuals. For a single observation, the unexplained portion, or *residual*, Y_{ntij} is defined as:

$$Y_{ntij} \stackrel{\text{def}}{=} X_{ntij} - P_{ntij} \qquad (3.1)$$

Unlike the first two analyses, the hope here is that there will be no feasible factor. The presence of a dominant factor in the residuals need not invalidate the measurement but may indicate that there is more information in the dataset than that related to the intended measurement. If the instrument was intended to be one dimensional, this may pose a threat to validity. The dimensional analysis from an item perspective of the two datasets used in this review is shown in Figure 3.3.



(a) (b)

Figure 3.3. Dimensionality of two sample datasets

This figure shows the "scree" plots (Cattell, 1966) of the variance explained by each component in the Principal Components Analysis of the datasets. Similar graphs can be produced from the Principal Components Analysis of the sample and residuals. The presence of a dominant dimension can be seen in both of these examples. The light grey line shows the broken stick distribution and the error bars the 95% confidence limits for the broken stick estimates. The colour coding reflects the strength of the components. The blue dot denotes a dominant component and red

dots denote non-significant components. The number shown alongside the first component expresses its eigenvalue as a multiple of the corresponding expected broken stick value.

The overall logic of this section can be summarised as follows. The hypothesis that the construct is unidimensional is formally tested by checking that there is exactly one homogeneous component in both the scale (items) and the sample (subjects). If more than one component is detected in either, and the option to enforce strict unidimensionality has been chosen, the software will automatically re-run the measurement model with dimensionality constraints, thus producing unidimensional measurement, aligned with the first factor, even when multidimensionality is present in the dataset. The formal testing is supplemented by several Principal Components Analyses that can be used to gain further insights into the structure of the dataset.

3.4. DIFFERENTIAL ITEM FUNCTIONING (HYPOTHESIS 4)

Differential Item Functioning (DIF) occurs when an item is more difficult for one set of subjects than another, and this difference is not attributable to the abilities of the members of the sets. In an educational setting, this might be a test question in which tacit knowledge was assumed and the extent of this tacit knowledge varied across groups of subjects. For example, a question set in a female sporting context might have a bias towards females; references to a past national event might have a bias against recent immigrants. It is distinguished from differential performance on items, or *item impact*, which is attributable to differing abilities and thus not biased. Zumbo (1999) explains the difference as follows:

> Item impact is evident when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the underlying ability being measured by the item. DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item *after matching on the underlying ability* that the item is intended to measure. (p. 12)

Thus, DIF can be detected by examining differential responses to an item by groups of subjects, while conditioning on ability. The hypothesis to be tested is:

- H4: There is no differential item functioning

The approach taken to address this hypothesis is to analyse the residuals, as defined in equation 3.1 (p. 113), remaining after fitting the model. Because an assumption of the model is that responses depend only on subject abilities and item difficulties, there should be no additional information in the dataset about responses to items over and above that supplied by the modelled abilities. Consequently, the residuals should not be significantly associated with any other characteristic of the subject. Accordingly, a test of DIF can be carried out by analysing the association of residuals with subject characteristics such as gender. An arbitrary number of variables identifying such characteristics may be defined and these are termed *DIF variables* herein.

From equation 2.89 (p. 68), it can be noted that the maximum likelihood estimate of an item's difficulty occurs when the sum across subjects of the observed scores for an item matches the sum of expected scores. Thus, although the residual values of individual subjects' responses will be non-zero, the expected value of the sum of the residuals across subjects is zero. These residuals can be summarised in a $k \times 2$ contingency table, where k represents the number of categories in the DIF variable and there are 2 columns: one containing a count of subject responses with positive residuals and the other a count of responses with negative residuals. In the absence of DIF, there should be no systematic association of these counts with the categories of the DIF variable. Consequently, DIF can be tested by a standard Pearson Chi-Square test of independence.

If significant DIF is detected, the impact is quantified in three ways. First, the *phi* coefficient (Cramér's V) and effect size (φ_C^2) is reported. Second, the relative difficulty of the item for each of the DIF variable categories is estimated and reported. This relative difficulty is expressed in the common metric of the scale. Third, the impact on estimated ability, also expressed in the common metric, is reported for each of the DIF variable categories.

Another manifestation of DIF or item bias is when different subjects interpret an item in a different way. The software uses the factors determined by the dimensionality test discussed earlier for hypothesis 3 to test this. A significant association suggests that subjects associated with different factors may be interpreting the item differently.

The presence of DIF in any of these analyses suggests a possible threat to measurement validity, but it is not necessarily a threat. For example, students with different backgrounds might conceptualise a problem in different, but equally valid, ways. It is important, then, to investigate the cause and determine whether the DIF represents a real threat to validity. In practice, the only remedy available when real DIF is detected is to exclude the item from the scale. The software provides an option to automate this approach to management of DIF. When this option is chosen and DIF is diagnosed, the measurement model is automatically run again with the added constraint that those items diagnosed with DIF are not used to calibrate cases. The effect of this is to reduce the information available for case measurements and thus increase the reported standard errors. However, removing items may represent a threat to content validity, and thus investigation is still required and measurements should be used with caution.

The overall logic of this section can be summarised as follows. The test of the hypothesis that there is no differential item functioning is tested by analysing each item. The hypothesis is rejected if any item is diagnosed with DIF. Diagnosis of DIF for an item involves checking for independence of the residuals across categories associated with DIF variables or sample factors. If an item fails any test, it is diagnosed as exhibiting DIF. All available nominal (category) variables associated with a subject are analysed as potential DIF variables. If DIF is detected, the impact on difficulty and ability estimates is quantified and an option allows the measurement model to be run again automatically excluding the items diagnosed as exhibiting DIF.

3.5. RESPONSE SET (HYPOTHESIS 5)

It is natural to assume that the responses of people to items are based on the meaning of the items to which they respond. However, it has long been known people's responses are also influenced by other factors (Cronbach, 1946). These factors are usually referred to as response sets, response styles, or response biases. Oskamp defined *response sets* as "systematic ways of answering which are not directly related to the question content, but which represent typical behavioural characteristics of the respondents" (1977, p. 37). Paulhus (1991, p. 17) defined a *response bias* as "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (that is, what the items were designed to measure)" and *response styles* as response biases that "an individual displays … consistently across time and situations". The term *response set* is used in the present work to refer to any non-content factor affecting responses.

Baumgartner and Steenkamp identify seven common types of response set (2001, p. 144). Table 3.2 sets out their definitions and the theoretical explanations they give for each. Analysis of response set is problematic for conventional analysis based on scores. For example, Greanleaf (1992) points out that measurement of Extreme Response Set is more accurate if the items are uncorrelated. This conflicts with the goal of measurement which requires highly correlated items. In contrast, the use of a formal measurement model allows simple and comprehensive analysis of response set.

Table 3.2: Common response styles and their theoretical explanations

Response style		Explanations
ARS	Acquiescence Response Style, a tendency to agree with items regardless of content. Also called agreement tendency, yea-saying, or positivity.	Characteristic of stimulation-seeking extroverts who have a tendency to accept statements impulsively. • Due to uncritical endorsement of statements by respondents who are low in cognitive abilities or have low status • More common for items that are ambiguous, vague, or neutral in desirability or for issues about which respondents are uncertain. • Most likely when respondents lack adequate cognitive resources because of distraction, time pressure, and so forth.

Response style		Explanations
DARS	Dis-acquiescence Response Style, a tendency to disagree with items regardless of content. Also called disagreement tendency, nay-saying, or negativity	Characteristic of controlled and reflective introverts who try to avoid external stimulation.
NARS	Net Acquiescence Response Style, a greater tendency to agree with items, rather than disagree, regardless of content. Also called directional bias	See above two definitions
ERS	Extreme Response Style, defined as a tendency to endorse the most extreme response categories regardless of content.	Reflection of rigidity, intolerance of ambiguity, and dogmatism. • Associated with higher levels of anxiety and possibly deviant behaviour. • Characteristic of respondents with less differentiated cognitive structures and poorly developed schemas. • Greater for meaningful stimuli (i.e. stimuli that are important or involving to respondents).
RRRS, MRS	Restricted Response Range Style or Mild Response Style, a tendency to use a narrow range of response categories around the mean response.	
MPRS	Midpoint Responding Style, defined as a tendency to use the middle scale category regardless of content.	Due to evasiveness (desire not to reveal one's true opinion), indecision (uncertainty about one's position), or indifference (lack of interest in an issue).
NCR	Non-contingent Responding, defined as a tendency to respond to items carelessly, randomly, or non-purposefully	Due to lack of motivation to read the instructions and interpret items appropriately

The approach taken in the present work is set out below. From equation 2.53 (p. 49), it can be noted that, for each subject case and for each item, the model gives the probability that the subject will respond in each of the response categories. Summing these probabilities by response category across a subject's responses to items allows the construction of expected counts of observations in the categories under the assumption that the model holds. A standard χ^2 goodness of fit test (Pearson, 1900) can then be used to compare these expectations with the actual

counts observed. The summations and tests are carried out at the individual subject case level, the individual subject level, and the overall instrument level. These tests serve both as formal tests of the fit of the measurement model and as an omnibus test for the presence of response set. If response set is detected, then the specific types of response set can be diagnosed by further analysis of the specific response patterns. Although there are different theoretical explanations for the various response sets in Table 3.2, they may be usefully combined into a more terse classification for analysis as set out below.

Neutral category bias can be defined as the overuse or underuse of the central category. Overuse corresponds to the midpoint responding style (MPRS). Baumgartner and Steenkamp do not identify a style corresponding to underuse, but this can be defined as negative MPRS. Neutral category bias is only defined and tested if there is an odd number of response categories. It can be diagnosed by a standard χ^2 goodness of fit test on the aggregation of the expected and observed counts into two categories: neutral and others.

Extremity bias can be defined as the overuse or underuse of the extreme categories. Overuse corresponds to the extreme response style (ERS); underuse corresponds to the mild response style (MRS), also known as the restricted response range style (RRRS). Extremity bias is only defined and tested when there are at least four response categories. It can be diagnosed by a standard χ^2 goodness of fit test on the aggregation of the expected and observed counts into two categories: one corresponding to the extreme response categories and one corresponding to the others.

Acquiescence bias can be defined as the tendency to agree with items regardless of content. Different specific diagnoses are given, depending on how this is manifested. The distinction made is based on the relative direction and magnitude of positive and negative bias, which need not balance if there is also any neutral category bias. If the positive and negative bias are in opposite directions, for example if agree categories are used more than expected and disagree categories less than expected, the net acquiescence response style (NARS) is postulated and

tested. Otherwise, the acquiescence response style (ARS) is postulated if the magnitude of bias is greater on the agree side and the dis-acquiescence response style (DARS) is postulated if it is not. Each of these categories can be present as both positive and negative manifestations. Thus, for example, a distinction is made between negative ARS (using agree less than expected) and DARS (using disagree more than expected). Acquiescence bias is diagnosed by a standard χ^2 goodness of fit test on the aggregation of the expected and observed counts into two categories: one corresponding to the postulated pattern and one corresponding to its absence. To identify positive endorsement, the assumption is made that categories are defined in order from most disagreement to most agreement and that there are equal numbers of categories relating to agreement and disagreement. When reverse coded items are encountered, an appropriate adjustment is made to preserve meaning.

Where response set is present, and none of the above bias elements are present, *non-contingent responding* (NCR) can be diagnosed.

Where the definition of categories varies across the items in the scale, the overall set of possible response categories is the superset of the defined response categories for each item and the expected and observed counts are modelled at zero for each response category not defined for the item. The overall test of model fit and response set thus remains valid. However, some limitations apply to the response set tests in this situation. The neutral category bias test is carried out only on the subset of items that has a neutral category defined. The extremity bias test is carried out only on the subset of items that has at least four categories defined. The acquiescence bias test is only valid if the concept of acquiescence is valid across items. Thus it is possible to detect acquiescence bias when there is a mix of different response definitions, such as dichotomous yes/no items with response category definitions that indicate strength of agreement in three or more levels. It is also possible to diagnose response set for dichotomous items if some are reverse coded. However, acquiescence bias cannot be safely detected when such items are mixed with others for which the concept of agreement is meaningless.

One major type of response set that cannot be detected directly within the framework set out here is *desirable response set*. Desirable response set, or socially desirable responding (Paulhus, 2002) describes the tendency of respondents to reply in a manner that will be viewed favorably by others. There are two main elements to this response set: self-deception and impression management (Paulhus, 1984). Detection of desirable responding thus requires items that tap into these concepts, rather than analysis of general items. However, if such items are included in the instrument respondents may be classified according to the level of socially desirable responding, and an analysis of differential item functioning, as discussed in the last section, may be carried out to quantify the impact of the bias.

The hypothesis tested in this section is:

- H5: There is no subject response set

The hypothesis is tested both at the individual subject case level and at the overall level. In each instance, a standard χ^2 goodness of fit test (Pearson, 1900) is used to compare the actual counts observed with the expectations from the model. The hypothesis is rejected at the individual subject case level if a significant departure from model expectation is found. The hypothesis is rejected at the overall level if a significant departure is found and there is at least one subject case with diagnosed response set. The significance level used in both cases is $p < .05$.

If the hypothesis is rejected at the individual subject case level, additional tests are carried out to identify and categorise the diagnosed response set as discussed above.

Three measures of effect size are also given for when response set is diagnosed. The first is a simple count by category of differences between observations and expectation. The second is a discordance index expressing these deviations as a percentage. The third is the estimated measurement bias attributable to the corresponding response set, expressed in the measurement metric. This is defined quantitatively as the ability estimate for the actual observed response pattern, less the ability estimate under the assumption that there is no response set.

$$\text{bias} \stackrel{\text{def}}{=} \hat{\theta}_{observed} - \hat{\theta}_{model} \quad (3.2)$$

The estimate of ability $\hat{\theta}$ used is the maximum likelihood estimate of ability for a specified response distribution. For the *observed* estimate, this is the empirical response distribution observed, where the probability of a response in a category is 1 for observed responses and 0 otherwise. For the *model* estimate, the response distribution imputed by the measurement model is used.

The presence of response set signals a possible threat to measurement validity. Although bias as defined above can give an indication of the magnitude of the impact of response set on subject ability estimates, the effect on item difficulty estimated is unclear. It is important, then, to investigate the cause and determine whether the response set represents a real threat to validity. In practice, the only safe remedy available when real response set is detected is to exclude the subject from the scale. The software provides an option to automate this approach to management of response set. When this option is chosen and response set is diagnosed, the measurement model is automatically run again with the added constraint that those subject cases diagnosed with response set are not used to calibrate items. The effect of this is to reduce the information available for item measurements and thus increase the reported standard errors. However, removing subjects may represent a threat to generalisation from the sample, and thus investigation is still required and measurements should be used with caution.

To summarise, the central hypothesis tested in this section is that there is no subject response set. A standard Pearson χ^2 goodness of fit test is used to investigate this. The hypothesis is tested both at the overall level and at the individual subject level. If the hypothesis is rejected at the subject level, additional tests are carried out to diagnose and categorise the response set. The effect size and estimated measurement bias expressed in the measurement metric attributable to the response set are reported. An option is provided to run the model again when response set is diagnosed, excluding subjects with response set from item measurement estimates.

3.6. LOCAL INDEPENDENCE (HYPOTHESIS 6)

From the relationship to the *Naïve Bayes Classifier*, as discussed in 2.4.1, it can be seen that the underpinning assumption of the model is that of conditional independence. Conditional independence is usually labelled *local independence* (Lord, 1980) in the measurement literature and these terms are treated as equivalent herein. Mathematically, the assumption made is that the joint probability of two variables is equal to the product of their marginal probabilities, which only holds when the variables are independent.

$$P(A \cap B) = P(A)P(B) \tag{3.3}$$

Equivalently, joint probability can be expressed as the product of a conditional probability and a marginal probability:

$$P(A \cap B) \stackrel{\text{def}}{=} P(A|B)P(B) \tag{3.4}$$

It follows that the assumption of independence implies that:

$$P(A|B) = P(A) \tag{3.5}$$

Local item (or subject) dependence can be seen as an additional dimension (Wang & Wilson, 2005). This perspective, termed *trait dependence* (Marais & Andrich, 2008) has already been addressed under the hypothesis of unidimensionality (H3) described in section 3.3 and accordingly, this aspect of local dependence is not pursued further in this section. However, local dependence can also be manifested as a dependency between items or subjects. This perspective is termed *response dependence* (Marais & Andrich, 2008). First, responses by a subject to an item should not be influenced by the subject's response to any other item. In an educational context, this could occur when a test includes several items that are related to a common problem. For example, this commonly occurs when a scenario is presented and a number of questions are presented that relate to the scenario. Alternatively, there might be *item-chaining* in which the answer to an item depends on a correct answer to a previous question. Yen (1993) points out that item chaining is found in some mathematics performance assessments when, after

providing an answer, students are asked to explain their reasoning. From a confirmatory perspective, it might be expected that most obvious forms of item chaining have been replaced by polytomous items, but subtle forms, such as hints about a question's answer in the stem or background information of another question, can be difficult to spot when setting tests.

Second, responses to an item by a subject should not be influenced by any other subject's response to that item. In an educational context, group or team work clearly represents a challenge to this assumption of independence. From the confirmatory perspective taken here, it is assumed that subject responses are intended to be independent of other subjects and that group aspects of work are measured separately. The hypothesis to be tested is:

- H6: There is local (conditional) independence

The approach taken is to test for local dependence both among items and among subjects. If dependence is found for either of these, the overall hypothesis is rejected. If neither shows dependence, the overall hypothesis is accepted. The logic of the item dependence test is set out in detail herein; the logic of the subject dependence test follows by symmetry.

Item dependence is investigated by investigating similar patterns across residuals of pairs of items. According to Andrich and Kreiner (2010), evidence of local dependence between two items can be obtained by calculating the correlation between the residuals of the observed and expected responses. Yen's Q3 statistic (1984) is the Pearson correlation between items of the residuals across subjects from the measurement model and has been shown to provide an effective compromise between maximum power and minimal false positive rates (Kim, De Ayala, Ferdous, & Nering, 2011). The test is operationalized by calculating the Pearson correlation of residuals between each pair of items and retaining any significant correlations found in a correlation matrix. Correlations are considered significant at the $p < .05$ level subject to a *Bonferroni* correction (Abdi, 2007) for multiple comparisons. An item is deemed to exhibit local dependence if a significant correlation is found with any other item.

When local dependence is diagnosed, an attempt is made to evaluate the impact on measurement and, as a software option, to apply the necessary corrections to the model. The impact on estimated standard errors is clear. From an information theoretic perspective (Fisher, 1925), local dependence clearly overstates available information, which in turn leads to underestimation of standard errors and overstatement of reliability. The impact on measurement is less clear since most studies have been carried out in the context of the three parameter logistic model. For example, local dependence has been found to have a significant effect on the estimation of all item parameters of the three parameter logistic model (Chen & Wang, 2007). However, the largest effects were found for the discrimination and guessing parameters, both of which depend strongly on the (overstated) Fisher information. The effect on item difficulty estimates can be interpreted as a side effect of these because of interactions between the item parameters in this model. Indeed, in a re-analysis of large scale test using a model that explicitly accounts for local dependence, Wainer and Wang (2000) conclude:

> We have seen that while conditional dependence seems to have almost no effect on the estimation of item difficulty it tends to yield an over-estimate of the guessing parameter and a bias on the estimation of discrimination (Wainer & Wang, 2000, p. 217)

If the software option is chosen, a correction is applied automatically. Since the measurement model used does not include guessing or discrimination parameters an information correction should be sufficient to adjust for local dependency. The approach taken is to estimate the proportion of local dependence in the residual metric and to re-estimate the model with a reduced information weight applied to items (or subjects) with diagnosed dependency.

To estimate the proportion of local dependence, a regression equation is constructed for each item diagnosed with dependency. In this equation, the item's residual is the criterion and the residuals of the items with which it is significantly correlated are predictors. A standard linear regression gives the proportion of variance explained by the predictors and the proportion unique to the criterion

item. From equation 2.113 (p. 80), information is proportional to variance in the score. It can be seen that the common variance among the residuals is used multiple times across the criterion and set of predictors and it is this multiple use that results in the overstatement of information due to local dependence. Let k be the number of predictors and C the proportion of common variance. Then the information overstatement can be corrected by multiplying the information by an information correction factor:

$$ICF = 1 - C\frac{k}{k+1} \tag{3.6}$$

To summarise the main points in this section, the hypothesis is that there is local (conditional) independence. The hypothesis is rejected if any significant local dependence is found among subjects or items. Local dependence is diagnosed by examining the correlation of residuals between pairs of subjects and items respectively. If local dependence is diagnosed, the dependent subjects or items are reported, together with an estimated effect size. If local dependence is detected, the measurement model may be automatically run again with an information correction that removes the overstatement of information associated with the dependence, thus correcting the stated standard errors.

3.7. CONVERGENCE (HYPOTHESIS 7)

The hypothesis tested in this section is:

- H7: The measurement model converges correctly

The measurement process uses several iterative processes that progressively refine estimates until sufficient accuracy is achieved. Sufficient accuracy is defined as accurate to six decimal places. In order to protect against an endless process, the software sets a limit to the number of iterations permitted for each of these processes. For the Newton-Raphson iterations, this limit is set at 100 iterations. Convergence is typically rapid with this algorithm, usually needing only a few iterations, but some starting values and constraints are potentially problematic. If convergence fails in this process, the software falls back to a standard bisection

algorithm for root finding. Although slower, the bisection algorithm is guaranteed to succeed. The limit is 500 iterations elsewhere. If any process fails to converge within the specified limit, the corresponding measurement is marked as immeasurable, reported, and removed from the model. If the overall measurement process fails to converge, measurement is considered to have failed and all measurements are rejected.

The number of iterations of the overall model is reported as a quality metric. No failure to converge has been observed in any of the datasets studied. The standard model usually converges in 15 iterations or less. The 2PL model, which is used for diagnostic purposes, typically requires many more iterations – sometimes 100 or more.

To summarise, the hypothesis tested in this section is that the measurement model converges correctly. The hypothesis is tested both at the overall level and at the level of individual measurements. The outcome of the test at the overall instrument level is either rejection of the convergence hypothesis or acceptance. The outcome of the tests at the individual measurement level is either acceptance or rejection of the convergence hypothesis for that measurement. If the hypothesis is rejected at the overall level, measurement is considered to have failed and all output measurements are rejected.

3.8. ADEQUACY (HYPOTHESIS 8)

The hypothesis addressed in this section is:

- H8: Data are adequate for measurement

Two tests of this hypothesis are carried out and the hypothesis is rejected if either test fails. The first test establishes that there is sufficient variability in the dataset to establish measurement. The second test establishes that the observed response patterns are likely to be systematic rather than random.

The first test of sufficient variability in the data set is identified by the response trimming process. Establishing relative abilities and difficulties relies on counts of

observable events to estimate probabilities or odds-ratios. This is only tractable if, for each measurement, there is at least one event supporting a lower boundary and at least one event supporting an upper boundary. Consequently, it is not possible to establish the relative difficulty of an item threshold that has not been exceeded by any case (subject) or that has been exceeded by all cases. Such item thresholds are marked as impossible to measure and removed from the model. Similarly, it is not possible to establish the relative ability of any case that is below all item thresholds, or above all thresholds. Such cases are marked as impossible to measure and removed from the model.

This process is recursive. Removing cases may make further items immeasurable and removing items may make further cases immeasurable. Items and cases are thus removed progressively until all those remaining in the model are measureable. The percentage of responses trimmed in this process is reported as a quality metric. If there is insufficient variability in the dataset then, potentially, all cases and items could be removed, leading to failure of the measurement model and rejection of the adequacy hypothesis. This is however unlikely in practice and has not been encountered in any of the real datasets studied in this project. Encountering individual cases or item thresholds that cannot be measured is more commonplace and leads to rejection of those measurements rather than the overall model. These measurements are likely to be at the extreme ends of the operating range.

Even when measurement of individual cases or thresholds is rejected, it is possible to provide some useful information about the value. Formally, if no upper bound can be set, then the estimated value should be positive infinity and likewise the estimated standard error should be infinity. However, in this case it may still be possible to set a lower bound. Likewise, when no lower bound can be set, the estimated value is minus infinity and the standard error is infinity. Again, it may be possible to set an upper bound. The only case for which no useful information can be given is when neither a lower bound nor an upper bound can be set. This corresponds to subjects who have made no responses or items with no responses from subjects.

In summary, the outcome of the response trimming process at the instrument level is either measurement failure or a report of the number of responses trimmed. Each individual measurement is also marked as measurable or immeasurable.

The second test establishes that the observed response patterns are systematic rather than random. The test examines the reproducibility of the dataset using an index that is also used for hypothesis 10. The overall logic is that the ability to reproduce observations from measurements is evidence of systematic responding. The measurement model uses sets of dichotomous observations to impute the parameters (case abilities and threshold difficulties) which form the measurements. If these measurements are sufficient statistics, then it may be possible to reproduce or reconstruct the set of observations from these statistics to some extent. In such a reconstruction, an observation may be defined as concordant (C) if it accords with the reconstruction and discordant (D) otherwise. Guttman (1944, p. 140) defined a *coefficient of reproducibility* as:

$$\mathrm{C}_{ntij} \stackrel{\text{def}}{=} \begin{bmatrix} X_{ntij} & |\theta_{nt} \geq \beta_{ij} \\ 1 - X_{ntij} & |\theta_{nt} < \beta_{ij} \end{bmatrix} \tag{3.7}$$

$$\mathrm{D}_{ntij} \stackrel{\text{def}}{=} 1 - \mathrm{C}_{ntij} \tag{3.8}$$

$$\mathrm{C} \stackrel{\text{def}}{=} \sum_{ntij} \mathrm{C}_{ntij}\,, \quad \mathrm{D} \stackrel{\text{def}}{=} \sum_{ntij} \mathrm{D}_{ntij} \tag{3.9}$$

$$\mathrm{COR} \stackrel{\text{def}}{=} \frac{C}{C + D} \tag{3.10}$$

This *coefficient of reproducibility* thus represents the proportion of responses that can be successfully reconstructed from the measurements. From this formulation, one can test whether the measurement model is extracting useful information from the dataset or whether the imputed measurements might simply be artefacts of the randomness of the sample. Formally, one can ask whether C is significantly greater than D, or equivalently the probability of obtaining C or more concordant observations in a set of $N = C + D$ observations that is drawn randomly from a population in which concordance and discordance are equally probable. This is tested by a standard binomial test that determines the probability of obtaining by

chance the observed degree of concordance from the sample. If this is not less than the significance level ($p < .05$), the hypothesis of adequacy is rejected and measurement is considered to have failed.

In summary, this section has addressed the hypothesis that data are adequate for measurement. Adequacy requires both sufficient variability in the dataset and evidence of systematic responding. This is tested both at the overall instrument level, and at the level of each individual measurement. Failure at the individual level results in rejection and reporting of the individual measurement. Failure at the overall level results in rejection of the adequacy hypothesis and rejection of all output measurements.

3.9. CONNECTIVITY (HYPOTHESIS 9)

The hypothesis described in this section is:

- H9: There is a common metric for all measurements

A key measurement goal is that both item difficulties and subject abilities are placed on a single common metric. The model handles sparse data in a natural manner; it does not require all subjects to respond to all items. However, when the data are too sparse, it might not be possible to establish a common metric. Conceptually, two items can be placed on the same metric if multiple subjects respond to both items; the various responses allow an estimate to be made of the relative difficulty of the items. Likewise, two subjects can be placed on the same metric if both respond, at least, to a common subset of items; an estimate can then be made of the relative ability of the subjects.

With a time series, multiple estimates are produced for each subject's ability at different points of time and all of these estimates should be placed on the common metric. The term case is used in this work to denote an estimate of subject ability at a point in time. With this terminology, cases connect items, and items connect cases. Cases and items can therefore be conceptualised as a graph that has cases as nodes and items as edges, or alternatively as a graph with items as nodes and cases as edges.

A graph is connected if there is a path through edges from any node to any other node. If the graph is connected then all cases and items can be placed on the same metric. However, if the graph is disconnected, then there may be two or more disconnected groups and, although the cases and items in each group may be placed on a metric common to the group, the metrics for each such group may differ.

In this project, a connectivity analysis is carried out to verify that a common metric is being used. In this analysis, items are treated as nodes and cases as edges. If the graph is disconnected, the number of disconnected groups is reported and the measurement model is deemed to have failed. If the graph is connected, the minimum number of cases connecting items is calculated; this minimum connectivity is then reported as a quality metric of the scale.

In general, connectivity might be expected to be a problem with a time series, since subjects might interact with relatively few items on any particular occasion. However, the weighting approach used in this project maintains connectivity in this situation. It does this by associating an information weight with each graph edge. Where a subject responds to an item with the same response over m successive cases, an edge is created for each of these m cases, each with an information weight of $1/\mathrm{m}$.

In summary, the hypothesis addressed in this section is that there is a common metric for all measurements. This is tested by analysing a connectivity graph to determine that each item is connected directly or indirectly by case responses to every other item. Where this is found to hold, the hypothesis is accepted and the minimum level of connectivity (cases) is reported. If it does not hold, the hypothesis is rejected and the number of disconnected groups is reported. When the hypothesis is rejected, measurement is considered to have failed and all output measurements are rejected.

3.10. REPRODUCIBILITY (HYPOTHESIS 10)

The hypothesis addressed in this section is:

- H10: Response patterns are reproducible from measurements

A goal of any measurement model is that it adequately represents the input data. The test of adequacy (hypothesis 8) used Guttman's *coefficient of reproducibility* to assess whether there was sufficient information in the dataset about the objects being measured to proceed with measurement. The test in this section uses this same index to identify how well the measurement model captures this information. If the model were deterministic, one would expect higher values of the coefficient of reproducibility to indicate better fit. However, the measurement model is inherently stochastic. Accordingly, it is more appropriate to investigate how close the observed reproducibility is to that expected under the assumptions of the model. The expected values of concordance and discordance under these assumptions are:

$$E(C_{ntij}) \stackrel{\text{def}}{=} \begin{bmatrix} P_{ntij} & |\theta_{nt} \geq \beta_{ij} \\ 1 - P_{ntij} & |\theta_{nt} < \beta_{ij} \end{bmatrix} \tag{3.11}$$

$$E\left(D_{ntij}\right) \stackrel{\text{def}}{=} 1 - E\left(C_{ntij}\right) \tag{3.12}$$

$$E(C) \stackrel{\text{def}}{=} \sum_{ntij} E\left(C_{ntij}\right), \qquad E(D) \stackrel{\text{def}}{=} \sum_{ntij} E\left(D_{ntij}\right) \tag{3.13}$$

$$E(\text{COR}) \stackrel{\text{def}}{=} \frac{E(C)}{E(C) + E(D)} \tag{3.14}$$

A test of model fit is carried out by comparing the observed counts of observed concordance (C) and discordance (D) with these expected values using a standard χ^2 goodness of fit test (Pearson, 1900). The test is carried out at the overall level, and for each item, and each subject case. The test is considered to have failed if the probability is less than the significance level ($p < .05$). There are three possible outcomes from this test:

- The reproducibility is as expected (at $p < .05$)

- The reproducibility is significantly greater than expected
- The reproducibility is significantly less than expected

Where the reproducibility is as expected, the hypothesis is accepted. Where the reproducibility is greater than expected, there is less randomness in the dataset than predicted by the model. This indicates that there is local dependence among subjects or items. However, such local dependence is detected and corrected in the test of local dependence (hypothesis 6) and thus no additional reporting or correction is needed. Accordingly, the hypothesis is also accepted in this situation. In contrast, misfit of the model is indicated when reproducibility is significantly less than expected. Consequently, the hypothesis is rejected in this situation and appropriate diagnosis and correction is required.

Different rules are applied for diagnosis and reporting at the individual subject case or item level, and at the overall level. All misfits are reported at the individual level, but the overall test is sensitive and a modified rule is used. It can be noted that no model will fit data perfectly and that as the number of observations increases, progressively smaller departures from the model will be identified as statistically significant. Martin-Löf (1974) expresses the problem as follows:

> In statistical practice, we are faced with the following dilemma. When the number of observations is small, that is, when we have little information about the random phenomenon that we are studying, we easily get a positive result: this or that model fits the data satisfactorily, whereas with large sets of data our results are purely negative: no matter what model we try, we are sure to find significant deviations which force us to reject it. (p. 3)

To avoid reporting very small, but statistically significant variations, misfit is only diagnosed at the overall level if misfit is also diagnosed for at least one individual subject case or item. This also avoids the possible confusion that may arise if overall misfit is reported, but none is found on more detailed investigation.

If misfit is diagnosed, correction is applied at the individual case or item level. Where reproducibility is less than expected, there is less information in the dataset

than expected and thus standard errors may be underestimated. As discussed in the introduction to this chapter, where it is possible to quantify the extent to which a measurement is potentially affected by any misfit of the model, the stated margin of error may be adjusted accordingly. This is achieved by using a *standard error inflator*.

Although the measurement model used makes no assumptions about the distribution of the objects attributes being measured, an implicit assumption is conventionally made, when a standard error is interpreted, that measurement errors are distributed normally. Indeed, the *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results* (Taylor & Kuyatt, 1994) recommend:

> Convert a quoted uncertainty that defines a "confidence interval" having a stated level of confidence (see subsection 5.5), such as 95 or 99 percent, to a standard uncertainty by treating the quoted uncertainty as if a normal distribution had been used to calculate it (unless otherwise indicated) and dividing it by the appropriate factor for such a distribution. These factors are 1.960 and 2.576 for the two levels of confidence given. (p. 3)

This assumption of normality is made in the present work when representing confidence intervals visually as error bars. Such an assumption seems plausible *a priori* because observed errors may often be considered as the sum of many independent factors and thus will tend to a normal distribution as the number of factors increases. However, the assumption may be in doubt when there is misfit to the model. Nevertheless, it can be noted, from *Chebychev's inequality* that no more than $1/k^2$ of the possible values will fall outside k standard deviations, whatever the distribution. For example, no more than 5% of the values can fall outside 4.47 ($\sqrt{20}$) standard deviations. Moreover, under the mild assumptions that the distribution has finite variance and is unimodal, the *Vysochanskij–Petunin inequality* determines that no more than 5% can fall outside ± 3 standard deviations. It is therefore possible to construct a meaningful confidence interval even if the assumption of a normal distribution of errors is not met. However, rather than

expecting different interpretations of the standard error when misfit is detected, the approach taken in this thesis is to adjust the standard error so that a conventional interpretation of standard errors and confidence intervals produces appropriate inferences.

This adjustment is achieved by multiplying each estimated standard error by a *standard error inflator*. Where the model fits, this inflator is 1; where misfit is detected, it will be greater than 1. There are several tests of model fit and, where applicable, a different standard error inflator is associated with each test when misfit is detected; the specific standard error inflators used are introduced in the description of the relevant tests. If the model fails more than one test of fit, the largest inflator is used. However, the inflator is not unbounded but is constrained to a maximum value determined by Chebychev's inequality. For example, with a confidence level of 95% the maximum value is 2.28 ($4.47/1.96$); at 99% it is 3.88 ($10/2.576$). Where an empirical test of distribution determines the distribution to be unimodal, the inflator is further constrained according to the *Vysochanskij–Petunin* inequality.

The use of a standard error inflator allows useful interpretation to proceed even when some misfit or bias is detected. However, the inflator is overly conservative in the sense that it is based on the maximum possible distortion that could occur. Accordingly, appropriate diagnostics are also given to allow the reason for misfit to be investigated so that the instrument can be improved for future administrations. It can be noted that misfit to the model is most likely to occur at the extreme ends of the operating range (Wright & Stone, 1999). Conversely, in an educational context, critical decision points should be substantively within the operating range of a well-designed measurement instrument. From this perspective, overstating the margin of error has minimal practical impact in an educational context and is preferable to rejecting the measurement.

It may be noted from equation 2.114 (p. 81) that information is inversely related to variance and is also proportional to the number of observations. Where there is redundancy, or the information is otherwise untrusted, an information correction,

and thus a standard error correction, can be derived if the proportion of true, or trusted, observations can be assessed. Accordingly, if π_T represents the proportion of observations that are trusted and π_U represents the proportion that is untrusted, the general form of the standard error inflator is:

$$SE\ inflator \cong \sqrt{\frac{1}{\pi_T}}\ \ or\ \sqrt{\frac{1}{1-\pi_U}} \qquad (3.15)$$

For the goodness of fit test, the trusted proportion is defined as:

$$\pi_T = \frac{\sum Min(O, E)}{\sum O} \qquad (3.16)$$

Here, O represents the number of observations in a category, E represents the number expected, and the summation is taken over all the categories.

To summarise, the hypothesis tested in this section is that response patterns are reproducible from measurements. The test verifies that the model reproduces responses as well as expected by theory and thus captures the information in the dataset appropriately. The hypothesis is tested both at the overall level and at the level of each individual subject and item measurement. If misfit is detected, the imputed standard error of individual measurements is adjusted so that the implicit confidence limits associated with the measurement are likely to encompass the true measurement.

3.11. OUTLIERS (HYPOTHESIS 11)

The hypothesis tested in this section is:

- H11: There are no more outliers than expected

In an educational context, outliers occur when an examinee with low estimated ability unexpectedly succeeds on a difficult item, or when an examinee with high estimated ability unexpectedly does not succeed on a relatively easy item. Where the number of such outliers is significantly higher than expected, there may be reason to doubt whether the ability and difficulty estimates truly capture the real

values. An observation can be defined as an outlier if the probability of its occurrence under the modelled assumptions is less than a specified significance level (i.e. p < π). For conciseness, the following definitions will be made.

$$\delta \stackrel{\text{def}}{=} \alpha\left(\theta_{nt} - \beta_{ij}\right) \tag{3.17}$$

$$\gamma \stackrel{\text{def}}{=} \log_e\left(\frac{1-\pi}{\pi}\right) \tag{3.18}$$

With this notation, then from equation 2.75 (p. 66), the likelihood of an individual dichotomous observation can be expressed as:

$$\mathcal{L}_{ntij} = \frac{e^{\delta X_{ntij}}}{1 + e^{\delta}} \tag{3.19}$$

A dichotomous observation X_{ntij} can then be classified as an outlier when:

$$Outlier \stackrel{\text{def}}{=} \mathcal{L}_{ntij} < \pi \tag{3.20}$$

$$\frac{e^{\delta X_{ntij}}}{1 + e^{\delta}} < \pi \tag{3.21}$$

$$e^{\delta X_{ntij}} < \pi + \pi e^{\delta} \tag{3.22}$$

Considering, first, the situation where the observation is 0,

$$1 < \pi + \pi e^{\delta} \tag{3.23}$$

$$1 - \pi < \pi e^{\delta} \tag{3.24}$$

$$e^{\delta} > \frac{1-\pi}{\pi} \tag{3.25}$$

$$\delta > \log_e\left(\frac{1-\pi}{\pi}\right) \tag{3.26}$$

$$\delta > \gamma \tag{3.27}$$

In the situation where the observation is 1,

$$e^{\delta} < \pi + \pi e^{\delta} \tag{3.28}$$

$$e^{\delta} - \pi e^{\delta} < \pi \tag{3.29}$$

$$e^{\delta}(1 - \pi) < \pi \tag{3.30}$$

$$e^{\delta} < \frac{\pi}{1 - \pi} \tag{3.31}$$

$$\delta < \log_e\left(\frac{\pi}{1 - \pi}\right) \tag{3.32}$$

$$\delta < -\gamma \tag{3.33}$$

The likely range of δ is thus from $-\gamma$ to γ and it can be considered to represent an outlier if it is outside this range. Substituting the reference model notation, a dichotomous observation can thus be considered an *outlier* under the conditions below.

$$\gamma < \alpha(\theta_{nt} - \beta_{ij}) \qquad |X_{ntij} = 0 \tag{3.34}$$

$$\alpha(\theta_{nt} - \beta_{ij}) < -\gamma \;\; |X_{ntij} = 1$$

The first of these conditions corresponds to unexpected lack of success on an easy item; the second corresponds to unexpected success on a difficult item. From the perspective of the model, some outliers are to be expected. For example, if outliers are defined at the $p < .05$ significance level, one would expect that, if the model fits, then on average about 5% of the observations would be outliers. If there are too many outliers, or too few, then there is evidence of model misfit. The question then is what variation in the number of outliers detected can be considered normal, and when is the number so low or so high as to be considered significant as evidence of misfit. This is tested with a standard binomial test.

There is a problem, however, with this approach. The value of γ in equations 3.27 (p. 137) and 3.33 (p. 138) is approximately 2.94 logits for the conventional significance level of $p < .05$. However, a well-targeted test aims for a close match between subject abilities and item difficulties; typically, this is within 2 logits. It is therefore difficult to identify misfits or outliers in a well-targeted test though this

approach. This produces what Wright and Stone (1999) term an *efficiency-fit paradox*:

> 1. Responses to items which provide maximum information allow minimum misfit detection.
>
> 2. Responses to items which allow maximum misfit detection provide minimum information. (p. 145)

Nevertheless, the software carries out the outlier analysis described above for both subject case and item threshold estimates. However, because of the paradox identified above, it uses a one-tailed test rather than the more conventional two-tailed test. Accordingly, it will flag as a misfit any estimate for which the number of outliers is unexpectedly high. A consequence of the tension between a well-targeted test and misfit detection is that this is considered a useful but insufficient test of model fit. The test is carried out both at the overall instrument level and for each item and subject case. Failure at the overall instrument level results in rejection of the hypothesis. Failure at the case or item level results in reporting of the individual case or item misfit but does not cause rejection of the overall hypothesis.

Outliers have relatively little impact on estimates since the information provided by an observation falls off as the distance between ability and difficulty increases. The contribution of an observation at the threshold distance (2.94 logits) used to identify outliers is less than 20% of the contribution when there is a match between ability and difficulty. Similarly, the impact on standard error estimates is relatively low. Nevertheless, a standard error inflator is applied when a case or item has more outliers than expected. For this inflator, excess outliers are treated as untrusted observations. Let π_U be the proportion of untrusted observations; then the corresponding inflator is:

$$Inflator \cong \sqrt{\frac{1}{1-\pi_U}} \tag{3.35}$$

Because of the relatively low information weight of outliers, this inflator is conservative, resulting in a slight overstatement of the standard error. On the other hand, only outliers are considered in this correction which could lead to understatement. However, there are several standard error inflators related to misfit and the largest of these is used. In practice, it is expected that the other error inflators will be larger when misfit is present because they are based on the full range of observations rather than only on outliers. Accordingly, it is deemed appropriate to take a conservative position on this inflator.

In summary, the hypothesis addressed in this section is that there are no more outliers than expected. This is tested at the overall level and at the level of each individual subject and item measurement. Failure at the overall level results in rejection of the hypothesis. When the hypothesis is rejected, measurement is considered to have failed and all output measurements are rejected. Failure at the individual level results in the adjustment of the imputed standard error for that measurement, but measurement is not considered to have failed.

3.12. FIT STATISTICS (HYPOTHESIS 12)

An alternative approach, based on analysis of residuals is given by Wright & Stone (1999). An observation in a dataset can be viewed as comprising two components: a portion explained by the model, and an unexplained portion. For a single observation, the unexplained portion, or *residual* (Y_{ntij}), is defined in equation 3.1 (p. 113). This can be standardised by dividing by the modelled standard error to give a *standardised residual* (Z_{ntij}).

$$Z_{ntij} \stackrel{\text{def}}{=} \frac{Y_{ntij}}{\sqrt{P_{ntij}\left(1 - P_{ntij}\right)}} \tag{3.36}$$

This standardized residual has a logistic distribution with an expected mean of 0 and a variance of 1 (Wright & Stone, 1999). In the asymptotic case, the squared residual (Z^2_{ntij}) can therefore be expected to follow a χ^2 distribution, with degrees of freedom given by the number of observations. To aid interpretation, two statistics, based on the residuals are derived: outfit and infit. The *Outfit* statistic is

defined (Wright & Stone, 1999, p. 53) for subject case estimates (U_{nt}) and item threshold estimates (U_{ij}) as:

$$U_{nt} \stackrel{\text{def}}{=} \frac{\sum_{ij}\left[Z_{ntij}^2\right]}{N_1} \tag{3.37}$$

$$U_{ij} \stackrel{\text{def}}{=} \frac{\sum_{nt}\left[Z_{ntij}^2\right]}{N_2} \tag{3.38}$$

In these equations, N_1 represents the number of item responses summed for the case and N_2 the number of case responses summed for the item threshold. The *Infit* statistic is defined (Wright & Stone, 1999, p. 53) for subject case estimates (V_{nt}) and item threshold estimates (V_{ij}) as:

$$V_{nt} \stackrel{\text{def}}{=} \frac{\sum_{ij}\left[Y_{ntij}^2\right]}{\sum_{ij}\left[P_{ntij}\left(1-P_{ntij}\right)\right]} \tag{3.39}$$

$$V_{ij} \stackrel{\text{def}}{=} \frac{\sum_{nt}\left[Y_{ntij}^2\right]}{\sum_{nt}\left[P_{ntij}\left(1-P_{ntij}\right)\right]} \tag{3.40}$$

Here, Y_{ntij} is the residual as defined in equation 3.1 (p. 113). Linacre and Wright (1994) explain the difference between infit and outfit as follows:

> *Outfit* is dominated by unexpected outlying, off-target, low information responses and so is outlier-sensitive. *Infit* is dominated by unexpected inlying patterns among informative, on-target observations and so is inlier-sensitive. (p. 350)

Both *infit* and *outfit* statistics can be interpreted in a similar way. Each statistic has an expected value of 1.0, and a range from 0 to infinity. Values greater than 1.0 indicate that the data are less predictable than the model expects; values less than 1.0 indicate the data are more predictable than expected. The values also have a direct interpretation. A value of 1.3 indicates that there is 30% more randomness or "noise" in the data than modelled; a value of 0.8 indicates a 20% deficiency in predicted randomness. Broad guidelines for interpretation given by Wright and Linacre (1994) are shown in Table 3.3.

Table 3.3: Interpretation of mean square statistics

Value	Interpretation
> 2.0	Distorts or degrades the measurement system
1.5 - 2.0	Unproductive for construction of measurement, but not degrading
0.5 - 1.5	Productive for measurement
< 0.5	Less productive for measurement, but not degrading.

Table 3.4 sets out the diagnosis categories used, together with suggested corrective action synthesised from the advice of Wright and Linacre (1994). The interpretation of diagnostics is the same for both cases and items. However, in practice, there is little that can be done about cases, other than choosing a different sample in future when measurement is used in a research context. Accordingly, the guidance set out in the table focuses on what changes can made to items to improve the quality of the instrument in future administrations. Wright and Linacre (1994) also recommend investigating and removing items with high mean-squares before looking at items with low mean-squares because large mean-squares are likely to a partial cause of low mean-squares on other items.

Table 3.4: Diagnostic categories for combinations of infit and outfit statistics.

Category	Condition	Suggested corrective action
Immeasurable	Any	Investigate the cause. If the item is too easy or too difficult, consider replacing it.
Degrading	Either infit or outfit exceeds 2.0	Consider removing, rewording or replacing the item.
Noisy	Infit is 0.5 to 1.5, Outfit is 1.5 to 2.0	Possible guessing or carelessness – retain the item unless improvement is clear.
Unproductive	Infit is 1.5 to 2, outfit is 0.5 to 2.0	Low discrimination – retain the item unless a better item is available.
Productive	Infit is 0.5 to 1.5, outfit is 0.5 to 1.5	Useful for measurement – retain the item.
Over fitting	Either infit or outfit below 0.5	Inefficient because of limited new information but still useful for measurement - replace with a more efficient item when developing a new test, otherwise retain the item.

The *infit* and *outfit* statistics can identify the presence of excessive randomness (noise) in the data. Such misfit could lead to overstatement of the accuracy of reported measurements. Wright (1995) suggests the following standard error inflator to give an improved and more conservative estimate:

$$Max\left(1.0, \sqrt{infit}\right) \tag{3.41}$$

The infit statistic also provides the basis for a formal test of the hypothesis addressed in this section:

- H12 Fit statistics accord with theoretical expectations

Two tests of fit are carried out. The first test is implemented using the approach of Wright and Stone (1999). This involves converting the infit statistic to a t-statistic and then carrying out a standard t test. The test is carried out for each item and for each subject case. The item or case is flagged as having misfit if the t-test is significant at the $p < .05$ level. The second test is a standard χ^2 goodness of fit test of observations against model expectations. For items this test compares the number of observed responses in each category with the number expected under the assumptions of the model, as defined by equation 2.53 (p. 49). For subjects, the superset of categories across all items is used. This allows the category structure to vary from item to item. Counts of the number of observed responses and the number expected under the assumptions of the model are aggregated across all items and the observed and expected responses are compared. Each item or case is flagged as having misfit if the goodness of fit test is significant at the $p < .05$ level. Each item and subject case must pass both tests to be considered as fitting the model. Although both tests address different aspects of fit, there is likely to be considerable overlap between the tests and thus a family-wise error correction is not appropriate.

The test of the overall hypothesis of fit is based on these individual tests of item and subject case fit. It is to be expected that on average, 5% of cases or items will be diagnosed as having misfit, even when the assumptions of the model hold. The logic of the overall test is that no more items and cases are diagnosed with misfit

than would be expected as false positives when the assumptions of the model hold. This is implemented by two binomial tests comparing the number of misfits diagnosed for cases and items against the expectation of 5%. The hypothesis is rejected if either test is significant at a confidence level of $1 - \sqrt{(1 - .05)}$ which is the Šidák correction (Abdi, 2007) for multiple comparisons and provides an overall 5% error rate.

In summary, this section has addressed the test of the hypothesis that the statistics of model fit accord with theoretical expectations. This is tested both at the overall level and for each individual subject case and item. Failure at the overall level results in rejection of the hypothesis. Failure at the individual level results in the adjustment of the imputed standard error for that measurement, but measurement is not considered to have failed. Investigation of the cause of any failure is supported by statistics and diagnostics at the individual subject and item level.

3.13. RELIABILITY (HYPOTHESIS 13)

From a traditional perspective (Lord & Novick, 1968), an observed score is viewed as comprising two components: a true score and an error score. These are assumed to be uncorrelated and in consequence the variance of the observed score (σ_X^2), the true score (σ_T^2), and the error score (σ_E^2) are related as follows:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{3.42}$$

The reliability (ρ_{xx}) of an instrument is defined as the ratio of the variance of the true score and the observed score as follows:

$$\rho_{xx} \stackrel{\text{def}}{=} \frac{var(truescore)}{var(observed\ score)} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \tag{3.43}$$

Estimates of instrument reliability take three main forms. With a *test-retest* approach, a test is administered to the same sample at two points in time and the consistency of the test over time is assessed. With a *parallel forms* approach, two equivalent forms of the test are created and administered to a sample at the same time; the consistency between the two forms is assessed. With an *internal*

consistency approach, a single instrument is administered to a single sample at a point in time and the consistency of responses across items is assessed. This last is the approach used in this project. A widely used measure of internal consistency is known as *Cronbach's alpha*, although it was originally derived by Louis Guttman (1945) who termed it λ_3. Coefficient alpha is defined (Cronbach, 1951, p. 299) as:

$$\alpha \stackrel{\text{def}}{=} \frac{n}{n-1}\left(1-\frac{\sum_i V_i}{V_t}\right) \tag{3.44}$$

Here, n is the number of items; V_i is the variance of scores for item i; and V_t the variance of the observed total scores. Cronbach also gives an equivalent definition (1951, p. 305) as:

$$\alpha \stackrel{\text{def}}{=} \frac{n}{n-1}\left(\frac{\sum_i \sum_j C_{ij}}{V_t}\right);\ \forall i,j; i \neq j \tag{3.45}$$

Here, C_{ij} is the covariance between items i and j. Thus coefficient alpha can be seen as the ratio of inter-item covariance to total variance. The factor $n/(n-1)$ corrects for the proportion of variance in any item which is due to the same elements as the covariance. Cronbach (1951, p. 331) gives six important meanings for coefficient alpha. It can be interpreted as:

- The mean of all possible split-half coefficients.
- The correlation expected between two random samples of items from a pool like those in the given test.
- A lower bound for the coefficient of precision (the instantaneous accuracy of the test with these particular items),
- A lower bound for coefficients of equivalence obtained by simultaneous administration of two tests having matched items.
- An estimate and a lower bound of the proportion of test variance attributable to common factors among the items.
- An upper bound to the concentration in the test of the first factor among the items.

The standard calculation of Cronbach's alpha is sensitive to missing data and is unsuitable for sparse datasets. An alternate calculation of coefficient alpha, following the Spearman-Brown prediction formula, is:

$$\alpha = \frac{n\bar{r}}{1 + (n-1)\bar{r}} \tag{3.46}$$

Here n is the number of items and $\bar{r}$ is the average item-item correlation. This form is more tractable for sparse datasets (Lopez, 2007) and is used in the current project when data are sparse. Another conceptualisation of reliability is Armor's theta (Armor, 1974). This can be expressed as:

$$\rho_{xx} = \left(\frac{n}{n-1}\right)\left(\frac{E-1}{E}\right) \tag{3.47}$$

Here, E is the largest eigenvalue of the correlation matrix. Li and Wainer (1997) give an abstract treatment of reliability that unifies these conceptions of reliability.

$$\rho_{xx} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum w_i^2 s_i^2}{s_w^2}\right) \tag{3.48}$$

Here, w_i is a weight applied to measurement i and s_w^2 is the variance of the weighted sum. If C is the covariance matrix, D the diagonal matrix of variances, w the weight vector, and w' its transpose, this can be restated as:

$$\rho_{xx} = \left(\frac{n}{n-1}\right)\left(1 - \frac{w'Dw}{w'Cw}\right) \tag{3.49}$$

Defining R as the correlation matrix and v as $D^{1/2}w$, this can be defined as:

$$\rho_{xx} = \left(\frac{n}{n-1}\right)\left(1 - \frac{v'v}{v'Rv}\right) \tag{3.50}$$

This formulation makes it clear that reliability is a function of the correlation matrix with weights (w) standardised by multiplying by the standard deviation ($D^{1/2}$). The various reliability coefficients result from different weights. Where the weights standardise the measurements, Cronbach's alpha becomes the Spearman-Brown formulation. Thus, Spearman Brown reliability can be thought of as Cronbach's

alpha calculated on the sum of z-scores. Where v is the first eigenvector of R, it becomes Armor's theta.

The foregoing discussion has focussed on the reliability of the measurement model as an instrument for measuring subjects. As in the previous discussions of the characteristic curve and Fisher Information, it is also possible to investigate the reliability of the sample as a device for measuring items. The corresponding formulation for coefficient alpha is again:

$$\alpha = \frac{n\bar{r}}{1 + (n-1)\bar{r}} \tag{3.51}$$

Here n is the number of cases and $\bar{r}$ is the average case-case correlation. Likewise, Armor's theta can be defined for the components of a Principal Components Analysis of the sample.

However, although these reliability coefficients give an indication of the overall reliability of the instrument and sample, they do not indicate the reliability of an individual measurement. By appealing to the definition of reliability in equation 3.43 (p. 144), a measure of instrument reliability (ρ_{xx}), conditional on ability (θ), can be defined as:

$$\bar{\theta} \stackrel{\text{def}}{=} \frac{\sum_{nt} \theta_{nt}}{\mathrm{N}} \tag{3.52}$$

$$Var(cases) \stackrel{\text{def}}{=} \frac{\sum_{nt} (\theta_{nt} - \bar{\theta})^2}{\mathrm{N} - 1} \tag{3.53}$$

$$E(\rho_{xx}|\theta) = 1 - \frac{E(V|\theta)}{Var(cases)} \tag{3.54}$$

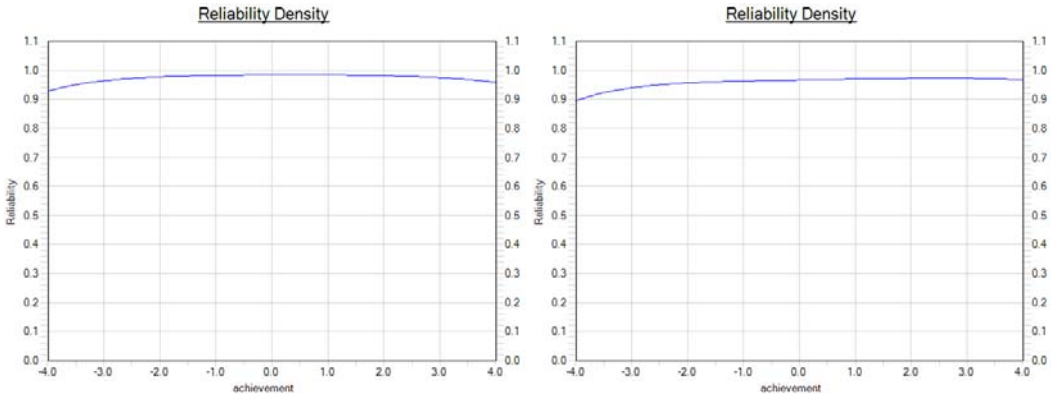
Here $E(V|\theta)$ is the expected variance of a measurement for a given ability as defined in equation 2.121 (p. 82) and N is the number of cases. Similarly, a measure of sample reliability (ρ_{yy}), conditional on difficulty (β) can be defined as:

$$\bar{\beta} \stackrel{\text{def}}{=} \frac{\sum_{ij} \beta_{nij}}{\mathrm{n}} \tag{3.55}$$

$$Var(thresholds) \stackrel{\text{def}}{=} \frac{\sum_{nt}\left(\beta_{nt} - \bar{\beta}\right)^2}{\mathrm{n} - 1} \tag{3.56}$$

$$E\left(\rho_{yy}|\beta\right) = 1 - \frac{E(V|\beta)}{Var(thresholds)} \tag{3.57}$$

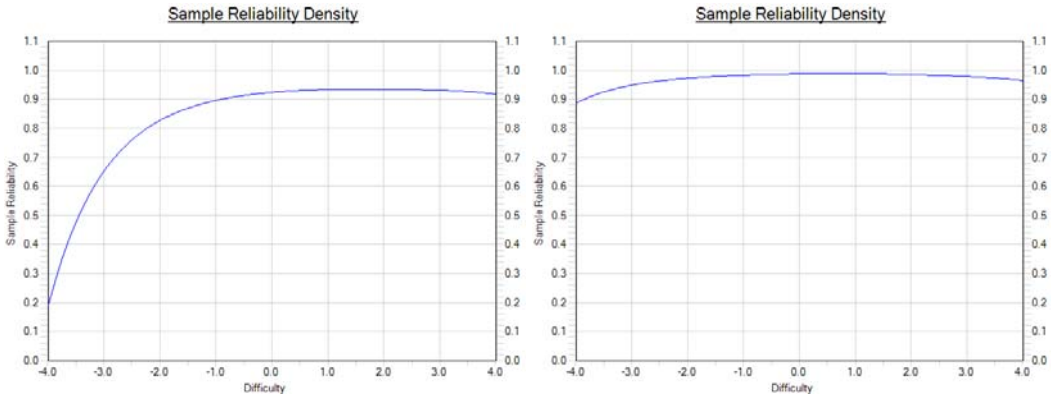
Here $E(V|\beta)$ is the expected variance of a measurement for a given difficulty, defined as the inverse of the conditional Fisher Information given in equation 2.124 (p. 83); n is the total number of thresholds. Equations 3.54 and 3.57 define the reliability of the instrument and sample conditional on ability and item difficulty respectively. The instrument reliability density graphs corresponding to the two datasets used in chapter three are shown in Figure 3.4.



(a) (b)

Figure 3.4: Instrument reliability density graphs for the sample datasets.

It can be seen that the reliability is not constant but falls off towards each end of the scale. This is because reliability is related to Fisher information which also falls off towards the end of the scale. The corresponding sample reliability densities are shown in Figure 3.5. It is noticeable that the reliability falls off rapidly on the left hand side of graph a. This is because the sample had few subjects in this ability region.



(a) (b)

Figure 3.5: Sample reliability density graphs for the two datasets

Each of the reliability formulations discussed above is calculated both for the instrument and for the sample. Cronbach's alpha and Armor's theta are based on raw scores which are non-linear. It is useful to have equivalent overall reliability indices based on a linear metric. The Rasch person separation reliability (PSR) and item separation reliability (ISR) serve this purpose. The Rasch PSR is defined (Wright & Stone, 1999, p. 153) as:

$$PSR = 1 - \frac{MSPE}{SVAR} \tag{3.58}$$

Here $MSPE$ is the mean square person error and $SVAR$ is the sample variance of the ability estimated. By comparison with equation 3.53 (p. 147), it can be seen that this is an equivalent formulation with the mean square person error substituted for the conditional variance at a scale location. The equivalent item separation reliability (ISR) is defined (Wright & Stone, 1999, p. 155) as:

$$ISR = 1 - \frac{MSIE}{IVAR} \tag{3.59}$$

Here $MSIE$ is the mean square item error and $IVAR$ is the variance of the difficulties estimated. By comparison with equation 3.57 (p. 148), it can be seen that this is an equivalent formulation with the mean square item error substituted for the conditional variance at a scale location.

The Rasch PSR and ISR are reliability indices that are conceptually equivalent to Cronbach's alpha and Armor's theta but based on the linear measurement metric.

However, equations 3.58 and 3.59 suggest that the essential datum is the relationship between measurement error and sample variability. With this motivation, a separation index (G) can be defined as the ratio of the unbiased estimate of the sample standard deviation to the root mean square measurement error of the sample (Wright & Stone, 1999). If M represents the root mean square measurement error and V represents the observed sample variance, then the unbiased sample standard deviation is $\sqrt{V - M^2}$. Thus G can be defined as follows:

$$G^2 = \frac{V - M^2}{M^2} = \frac{V}{M^2} - 1 \tag{3.60}$$

$$1 + G^2 = \frac{V}{M^2}$$

Moreover, any formula for reliability can be conceptualised as:

$$R = 1 - \frac{M^2}{V} \tag{3.61}$$

Thus

$$R = 1 - \frac{1}{1 + G^2} \quad = \frac{\mathrm{G}^2}{1 + \mathrm{G}^2} \tag{3.62}$$

Rearranging this equation gives the following relationship:

$$G^2 = \frac{R}{1 - R} \tag{3.63}$$

From equations 3.62 and 3.63 it can be seen that any reliability coefficient can be readily translated into a separation index and vice versa. An associated conception is the number of statistically distinct strata identifiable with the instrument. This is related to the separation index G as follows (Wright & Masters, 1982):

$$H = (4G + 1)/3 \tag{3.64}$$

$$G = (3H - 1)/4$$

From the above, it can be seen that a reliability index can be related directly to the number of statistically distinct strata that can be discerned by an instrument with that reliability. This allows an interpretation of a reliability index in terms of *fitness for purpose*. This perspective accords with everyday conceptions of measurement

usefulness and reliability. For instance, an instrument that measures weight to an accuracy of ±1 kilogram might be useful and reliable for measuring objects that range between 100 and 500 kilograms in weight but would be of little use for those that range between 800 and 1200 grams.

For example, if an instrument is required to distinguish between 3 strata in a sample, then reliability, whether measured by Cronbach's alpha, Armor's theta or the Rasch PSR, should be at least 0.8.

The formal hypothesis tested in this section is:

- H13. There is adequate reliability for useful measurement

The interpretation of useful measurement depends on the measurement purpose. From the foregoing discussion, it is clear that the number of distinct strata required will depend on that purpose and thus, a general test of the hypothesis is not possible. However, a partial test is possible. When the measurement error is larger than the standard deviation of the objects being measured, it is clear that the instrument cannot be fit for purpose. This situation corresponds to a calculated reliability index that is negative. Accordingly, the hypothesis will be rejected when this is detected. In practice, this test is insufficient and a judgement will need to be made as to whether there is sufficient reliability for the intended purpose. This will depend on the number of statistically distinct strata required. Once this has been defined, equations 3.62 (p. 150) and 3.64 (p. 150) show how the minimum value required for a reliability index can be determined.

In summary, the hypothesis tested in this section is that there is adequate reliability for useful measurement. Several reliability indices have been described. Cronbach's alpha and Armor's theta are conventional reliability indices, but these are based on scores rather than linear measurements. Additional reliability indices for the instrument and sample based on linear measurements have been introduced. The reliability indices, conditional on ability or difficulty indicate how reliability varies across the operating range of the instrument. The Person Separation Reliability index measures how well the items can discriminate among subject cases. The Item

Separation Reliability index measures how well the sample can discriminate among items. All reliability indices can be interpreted in terms of the ratio of sample variability to measurement error or, equivalently, the number of statistically distinct strata that can be discerned. This last perspective allows reliability to be interpreted in terms of fitness for purpose. A general test of the hypothesis cannot thus be carried out unless the purpose is known. However, formulae have been presented that allow the required level of reliability to be determined once the purpose is known. An extension of the test that incorporates purpose is introduced in Chapter Four. Nevertheless, a partial test is always possible, and the hypothesis will be rejected if there cannot be fitness for purpose. This occurs when reliability is zero or negative, or equivalently when measurement error exceeds the sample variability.

3.14. CHAPTER SUMMARY

The quality of output measurements is dependent both on the adequacy and quality of the input data and on the correctness of the theory underpinning the construction of the measurement instrument. A measurement exercise is thus a working hypothesis which one actively seeks to disconfirm. This chapter has set out the formal tests that are carried out to support or disconfirm the measurement hypotheses. These hypotheses provide a comprehensive framework for evaluation of the success of the measurement exercise.

Hypothesis 1 tests whether the dataset conforms to the axioms required for the construct to be quantifiable. Hypothesis 2 tests whether response categories are ordinal. Occasional failure at the item threshold level should be investigated but is not a cause for major concern because some failures are expected even when the assumption of ordinality holds. However, the overall hypothesis of ordinality is rejected if there are more failures than expected by chance. Taken together, hypotheses 1 and 2 establish whether measurement is conceptually possible with the input data set. If either hypothesis is rejected the measurement process is deemed to have failed and output measurements should not be used without further investigation.

Hypothesis 3 tests whether the construct is unidimensional across both subjects and items. The test is supplemented by several Principal Components Analyses that can be used to explore the structure of the dataset. Whether unidimensionality is required depends on the perspective taken when the construct was defined. Accordingly, rejection of the hypothesis need not cause rejection of output measurements. Hypothesis 4 tests whether there is any differential item functioning. The hypothesis is rejected if any item is diagnosed with DIF. When DIF is detected, the impact on difficulty and ability estimates is quantified. Hypothesis 5 tests whether there is any subject response set. The hypothesis is rejected at the overall level if any subject is diagnosed with response set. The effect size, the estimated measurement bias, and a characterisation of the set, are reported for each subject diagnosed with response set. Hypothesis 6 tests whether there is local (conditional) independence. The hypothesis is rejected if any significant local dependence is found among subjects or items. If local dependence is diagnosed, the dependent subjects or items are reported, together with an estimated effect size. Taken together, hypotheses 3, 4, 5 and 6 establish whether subjects are responding to items in the intended manner. Failure of any of these hypotheses need not invalidate the output measurements, but judgement is required as to the impact of any rejected hypothesis. The software also has an option to manage failure of each hypothesis. However, the management options effectively ignore the untrusted information in the dataset. This may pose a threat to content validity or the ability to generalise from the sample. Together with the first two hypotheses, these hypotheses give a cumulative assurance that the dataset has suitable input data for measurement.

Hypothesis 7 tests whether the measurement model converges correctly. The hypothesis is tested both at the overall level and at the level of individual measurements. The outcome of the test at the overall instrument level is either rejection of the convergence hypothesis or acceptance. The outcome of the tests at the individual measurement level is either acceptance or rejection of the convergence hypothesis for that measurement. Hypothesis 8 tests whether data are adequate for measurement. Adequacy requires both sufficient variability in the

dataset and evidence of systematic responding. This is tested both at the overall instrument level, and at the level of each individual measurement. Failure at the individual level results in rejection and reporting of the individual measurement. Hypothesis 9 tests whether all measurements can be placed on a common metric. This is tested by analysing a connectivity graph to determine that each item is connected directly or indirectly by case responses to every other item. Where this does not hold, the hypothesis is rejected and the number of disconnected groups is reported. Hypotheses 7, 8 and 9 verify the technical success of the measurement process. If any of these hypotheses is rejected, measurement is considered to have failed and all output measurements are rejected. Together with the previous six hypotheses, this group gives the cumulative assurance that measurement has been achieved.

Hypothesis 10 tests whether response patterns are reproducible from measurements. The test verifies that the model reproduces responses as well as expected by theory and thus captures the information in the dataset appropriately. Hypothesis 11 tests whether there are more outliers than expected. Hypothesis 12 tests whether the statistics of model fit accord with theoretical expectations. Taken together, hypotheses 10, 11 and 12 establish that the model adequately represents the input data. These hypotheses are tested both at the overall level and for each individual subject case and item. Failure at the overall level results in rejection of the hypothesis. Failure at the individual level results in the adjustment of the imputed standard error for that measurement, but the measurement process is not considered to have failed. Investigation of the cause of any failure is supported by statistics and diagnostics at the individual subject and item level. Together with the previous hypotheses, these three hypotheses give the cumulative as assurance that the intended measurement has been achieved with known accuracy.

Hypothesis 13 tests whether there is adequate reliability for useful measurement. It establishes that the accuracy achieved is sufficient to discriminate items and subjects appropriately. Only a partial test of this hypothesis is possible since a complete test is not possible unless the specific measurement purpose is known. Accordingly, a judgement is required as to whether the required reliability has been

achieved. Formulae have been presented that allow the required level of reliability to be determined once the purpose is known. An extension to this test that incorporates purpose is introduced in Chapter Four. Together with the previous hypotheses, this hypothesis gives the cumulative assurance that measurements are fit for purpose. The hypotheses described above and the associated inferences and cumulative assurances are summarised in Figure 3.6.

Within this framework, it is to be expected that, in any practical measurement application, some hypotheses will be rejected from time to time. When this happens, appropriate judgement is needed as to whether output measurements should be used. For some of the hypotheses, the judgement is clear. For example, it can be noted that some of these tests should result in total rejection of the measurement model. In particular, hypotheses H1 (quantifiable structure), H2 (ordinality), H7 (convergence), H8 (adequacy) and H9 (connectivity) are all required for measurement. Failure of any of these should result in rejection of all measurements. However, there is less clarity for the remaining hypotheses.

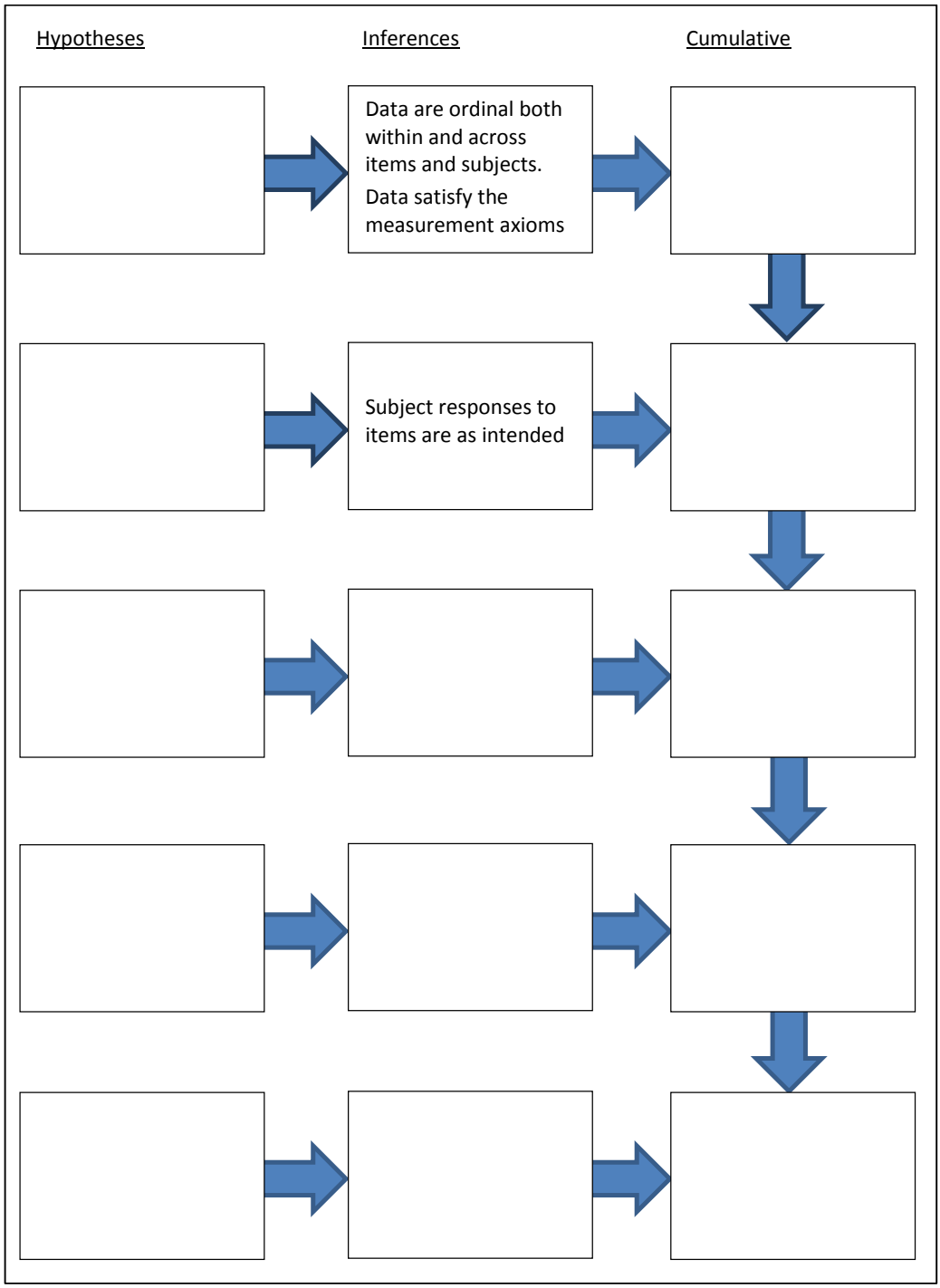


Figure 3.6: Inferences and cumulative assurance from the hypotheses

Hypotheses H3 (unidimensionality), H4 (no DIF), H5 (no response set) and H6 (no local dependence) require judgement as to whether any issue detected is critical and should result in rejection of all measurements, or whether it is minor and measurements can be used with appropriate caution. Moreover, although several

automatic corrections are available when tests in this group fail, the corrections are not a panacea.

Content validity may be threatened where subject cases are measured based only on items that are associated with the first factor (H3), and that exhibit no DIF (H4). Similarly, generalizability may be threatened when item measurements are based only on subjects that are associated with the first sample factor (H3) and that exhibit no response set (H5). The correction for local dependence (H6) does not affect validity, but does reduce available information and thus may increase standard errors. Judgement is thus still required as to whether to apply the automatic correction options.

Hypotheses H10 (reproducibility), H11 (outliers) and H12 (fit statistics) address how well the input data is captured by the model and attempt to remedy any detected misfit by adjustment of the standard error. Moreover, an attempt has been made to estimate the impact of any violations. Fit statistics can also be expected to show some degree of misfit. It has been noted that no model will ever fit the data perfectly and therefore some violation of the model's assumptions is to be expected in any real dataset. It follows that judgement is required as to whether the degree of misfit is minor and may be safely ignored, or whether the misfit identifies a conceptual problem with the measurement theory or with the implementation of the theoretical construct being measured. Hypothesis 13 also requires judgement as to the number of strata required for the measurement and the consequent level of reliability required.

Moreover, it should be noted that the nature of significance testing means that some reported failures will unavoidably be *false positives*. This means, for example, that when testing at the item or case level with the conventional significance level of $p < .05$, around 5% of items or cases reported as misfits are likely to be false positives and consequently need not be of major concern, although they should nevertheless be investigated. A conservative approach has been taken to use family-wise corrections only at aggregate levels since it is believed to be better to flag these individual items or cases for further investigation, rather than allowing

them to be masked by a higher level of aggregation. It follows from the above points that judgement is needed on the impact which any violations or misfit might have on the usefulness of measurement.

To summarise these points, this chapter has presented a comprehensive framework for the evaluation of the success of the measurement exercise. An attempt has been made to automate as much of the testing as possible and to provide appropriate corrective action when issues are diagnosed. Nevertheless, substantial judgement is still required in the interpretation of any issues raised. In practice, evaluation of the measurement hypothesis is a never-ending task and, in each run, an educator will need to prioritise any issues raised and make judgements as to which must be addressed before the measurements can be used and which should be addressed in future administrations. This is potentially a substantial burden without additional support.

Many, but not all, of the tests described in this chapter are available in commercial software. However, even with the tests that are available, no single available software package implements all of them. Moreover, many of these packages require substantial statistical expertise. It is not considered realistic to expect an educator to use several different packages to implement the tests described or to require an educator to have the required level of statistical expertise. Because of these issues, it is believed that, to provide a practical measurement tool for an educator, custom software is needed that brings all the necessary tests together into a single software package and provides appropriate support and guidance to support the judgements that the educator needs to make. This needs what is termed an *expert system* in computer science. The next chapter describes the expert system software developed for this project.

Chapter 4.
SOFTWARE DEVELOPMENT

> *Science is knowledge which we understand so well that we can teach it to a computer; and if we don't fully understand something, it is an art to deal with it. (Knuth, 1974, p. 668)*

The quotation above is taken from an article in which Don Knuth gave his reasons for choosing *The Art of Programming* as the title for his seminal work, rather than the *Science* of programming. Although computer science has made substantial progress since that article was published, few computing professionals would disagree that successful software still requires more than technical correctness. Knuth used the term art (in the original sense: *skill*, from the Latin *ars, artis*) to denote this additional requirement: what is now usually termed *human factors* in the computing literature. An underlying theme of this chapter is the need to pay attention to these human factors.

Chapter Two presented a theoretical framework for measurement and defined a formal measurement model. The last chapter provided a comprehensive framework for testing the measurement hypotheses. However, the complexity of the model and testing framework might limit its acceptance by the educators who are the target users of the model. The concern here echoes the self-efficacy theme of the present work: if the measurement model is seen as complex and too hard to understand and use, then educators and learners may shy away from its use. Implementation of the model should therefore attempt to shield educators and learners from much of this complexity in order to provide a practical and accessible tool. This chapter describes the key features of the development of a computer software implementation of the model which is designed to shield educators and learners from much of this complexity and, thus, provide a practical tool for educators and students. Achieving this requires both the technical correctness provided by a computer science approach and the attention to human factors implied by Knuth's characterisation of programming as Art.

The model was implemented in the context of a general purpose data collection and analysis software package written by the author; a brief overview of this package is given in Appendix A. This chapter focuses on the aspects of the package that relate to the measurement model and the associated human factors. The development followed a *Design Science* methodology (Hevner, March, Park, & Ram, 2004) and used an iterative approach, supported by a "beta" group of educators who trialled the tool and gave feedback over a period of two years.

However, despite the iterative nature of the development process, it is generally recommended (Peffers, et al., 2006) that communication of the process is approached in a linear manner. Accordingly, the organisation of the rest of this chapter is as follows. The first section discusses the issues motivating the need for custom software, reviews the requirements of the software and identifies the key issues that need to be addressed. The second section sets out the objectives of a solution arising from this problem definition. The third section summarises the design and development of key aspects of the software implementation. The fourth section demonstrates and evaluates the efficacy of the software implementation to solve the identified objectives; some parts of the evaluation are introduced in this section and presented in detail in Chapters Five and Seven. Finally, the role of the software in making the implementation of a formal objective measurement model accessible to educators, effective and practical is summarised.

4.1. PROBLEM IDENTIFICATION AND MOTIVATION

With the wide availability of desktop computers, many sophisticated statistical packages (e.g. SPSS, SAS, STATA, R, and MATLAB) are now available to researchers and practitioners. Broadly speaking, these packages are designed for expert use and are highly capable and effective when used appropriately. However, a number of scholars have commented that this proliferation of tools has led to misuse of statistical procedures. For example, Ross (1985) notes that the ready availability of sophisticated procedures in packages such as SPSS can lead to inappropriate use by those who may be unfamiliar with the logic and subtleties of the procedures. Such criticism generally leads to a call for more consultation with statisticians, better

understanding by users of statistical procedures, and better training in the software packages. For example, Nelson and Rawlings (1983) comment as follows.

> Widespread use of computers, together with the ready availability of software packages have also resulted in a number of misuses of statistics. Agronomists should recognise their need for statistical assistance in planning experiments and in analysing and interpreting experimental data. (p. 105)

Maus and Endresen (1979) go further in their criticism:

> A solution should be sought along the following lines: (a) educate users in the proper use of statistics; (b) abolish statistical packages altogether and revert to special programs for each survey and test ... (p. 128)

From a software development perspective, however, this may be looking at the problem the wrong way. Practitioners in many domains are faced with making numerous day-to-day decisions. It is reasonable to expect that the quality of these decisions would be enhanced if they were supported by appropriate objective evidence. However, it is not realistic to expect all practitioners to be expert statisticians, or to consult with such experts on every day-to-day decision. In computer science, fitness for purpose of computer software requires not only appropriate functionality, but an alignment between the operation of the software and user capabilities. What is needed, then, is a software package that provides objective evidence in a form that can be readily understood and interpreted by practitioners, and which protects them, wherever possible, from the risk of misinterpretation. Inevitably, this will lead to some loss of flexibility. The essential requirement of such a package is to replace the need for specialist judgement and interpretation by a set of rules or heuristics. This will help protect against any misjudgement or misinterpretation by the user but, conversely, means that the software does not take advantage of any expert knowledge the practitioner may possess. Nevertheless, this loss of flexibility can be mitigated by an approach in which the software implements, by default, a standard interpretation and set of rules, whilst also allowing overrides and the use of custom procedures.

The dominant paradigm for the statistical packages mentioned above is based on an *imperative* model in which the user issues commands to the software. With this paradigm, a dataset is collected and then the analyst specifies, and carries out, a number of procedures to test and analyse the data. Typically, several procedures are carried out in sequence, and judgements of the results of each are made by the analyst before proceeding to subsequent tests. The inherent weakness of such imperative models is the requirement for the analyst to have substantial expertise in specifying the procedures and making these judgements. The strength of the approach is that the analyst is free to devise whatever tests or procedures he or she wishes. An alternative approach is to use a *declarative* paradigm. With a declarative paradigm, the required output is defined and the software then automatically runs and interprets a series of appropriate tests and procedures to produce the requested output. In computer science, this latter paradigm is termed an *expert system*: a software program that uses heuristics and encoded domain knowledge to support the processing and interpretation of available information.

In summary, a declarative paradigm focuses on *what* should be done, whereas an imperative paradigm focuses on *how* it should be done. A declarative paradigm thus allows a user to work at a higher level of abstraction, but requires the encoding of domain knowledge and heuristics to determine the specific processes required to accomplish the objective. To achieve the goal of shielding educators and learners from much of the inherent complexity of the measurement model, the first required element of a solution is thus to:

- use a declarative paradigm

Moreover, determining the specific measurement objective requires the educator to specify the purpose clearly. From a technical perspective, specifying the purpose includes setting acceptable bounds on statistics such as reliability indices, but these statistical terms may not resonate with all educators. On the other hand, educators can be expected to be familiar with concepts such as pass marks, pass rates and the number of test items. Although the connection is rarely made between these and statistical concepts like reliability, there is a direct relationship. Accordingly, a

promising approach is to allow educators to specify requirements using concepts that are familiar to them and for the software to carry out the necessary translation into statistical terminology. To specify purpose and objectives, the second element required of the solution is thus to:

- use concepts and terms familiar to an educator

It is also important that the implementation should function effectively in a typical educational setting. In this regard, a distinction can be made between a measurement approach and other, more general, quantitative techniques. Many such quantitative statistical procedures are concerned with making inferences about a general population from a representative sample. Such inference typically requires a relatively large sample size and appropriate sample selection. However, educators often work with class sizes that are relatively small for statistical purposes, and they may have little say in the selection of students for their class. Thus, educators have to work with the actual students in their class: what is often termed a *convenience sample*. In contrast to a general quantitative approach, which has a focus on inference *from* a sample to a population, a measurement approach focuses on estimating values *within* the current sample. Loosely speaking, this distinction corresponds to the difference between an educator concluding "all students find topic X difficult" and "my students find topic X difficult". With a measurement approach, useful estimates can often be produced with as few as 20 students or items. Nevertheless, although the demands for measurement *per se* are less stringent than those for population inference, using a measurement approach does not preclude inference to a population. Indeed, many of the techniques widely used for population inference are based on the *general linear model* which depends, in part, on the assumption that inputs are measured at interval level. Consequently, achieving linear measurement provides a more solid basis for population inference than the use of raw scores. Thus, population inference is never weaker, and may be stronger, when using linear measurements rather than raw scores. To allow a practical solution in typical educational environments, without compromising inference to a more general population, the third required element of a solution is thus to:

- use a linear measurement approach

The natural outputs of any measurement exercise are numerical. This applies, not only to the measurements and standard errors produced, but to the outputs of the statistical tests and to the various fit statistics. Accordingly, the primary output of the software should be numeric, and often presented in tabular form. However, the level of comfort with such numerical output is likely to vary across educators. To facilitate interpretation of the output by educators with different preferences, the numeric outputs should be supplemented by representations in both visual and verbal forms. Visual representation can convey a great deal of information in a compact format, and verbal representation, including narrative forms, can support and guide the appropriate interpretation of outputs. In particular, attaching appropriate qualitative labels can enhance the interpretation of many statistics, and a narrative form is naturally suited to giving practical guidance on the actions to take when issues are diagnosed. The fourth required element of the solution is thus to:

- use multiple representations of output information

Although not all educators are expected to be comfortable with the complexities and subtleties of the statistical procedures used, they are nevertheless likely to have substantial knowledge of their subject domain and students, and they can also be expected to have considerable life experience. The use of appropriate metaphors and analogies can help the software take advantage of this knowledge and experience to build confidence, intuition and understanding. The fifth required element of the solution is thus to

- use appropriate metaphors and analogies

Finally, in addition to the measurements and standard errors of the subjects and items measured, there are 13 hypotheses, and numerous statistics and indices, which are used to describe and diagnose the success of the measurement exercise. Each of these has its own specific requirements and subtleties. To avoid the danger of this complexity becoming overwhelming for a practitioner, it is necessary to

provide a unifying framework for the diverse concepts and constructs. The sixth required element of a solution is thus to:

- use a unifying conceptual framework

To summarise the main points of the discussion above, fitness for purpose requires not only an appropriate technical implementation of the model, but an alignment between the operational demands of the software and the capabilities and preferences of the educators who are the intended users. A declarative paradigm allows the user to work at a higher level of abstraction than the more common procedural paradigm, but requires the software to encode domain knowledge and heuristics. This paradigm requires educators to identify the measurement purpose, which can be achieved more readily if it can be specified in terms that are likely to be familiar to them. Moreover, the software needs to operate in a context which may have a small class size and a convenience sample. This suggests the use of a linear measurement approach rather than an approach based on population inference, although population inference is still possible. Clarity and interpretation of outputs can be enhanced by using visual and verbal representations to supplement numerical outputs. Understanding can also be enhanced by appropriate use of metaphors and analogies and by the use of a unifying conceptual framework.

In short, an *expert system* which uses a declarative paradigm, concepts and terms familiar to educators, a linear measurement approach, multiple representations of output information, appropriate metaphors and analogies, and a unifying conceptual framework, is likely to provide a solution which can be used with confidence by educators and will operate effectively in a typical educational setting.

4.2. OBJECTIVES OF A SOLUTION

Three broad research questions/objectives were set out in the introductory chapter of this thesis:

- Is it possible to develop objective measures of challenge and self-efficacy from self-report data? (Objective 1)

- How can these measurements be communicated clearly to educators and learners? (Objective 2)

- Develop a practical computer software implementation that will communicate the measurements in real time (Objective 3)

It is clear from these objectives that the software should provide a complete and correct implementation of the measurement model and hypothesis tests. The model and tests have been described in detail in the last two chapters and, accordingly, are not further elaborated in this section. However, from the discussion in the previous section, it is possible to elaborate on the need for *clear communication* identified in objective two and on the need for the implementation to be *practical as* identified in objective three. From these considerations, the human factors can be addressed by adding the following objectives to the technical requirements:

- The software must be *usable* by educational practitioners
- Output must be *interpretable* by educators and learners.
- The software should provide *useful* measurement in realistic educational settings.

The objective that the tool be *usable* by educational practitioners suggests the development of an expert system. This, in turn, requires the development of appropriate inference rules and heuristics. It also suggests the use of a declarative, rather than an imperative, approach. The objective that output from the tool be *interpretable* by educators and learners suggests supplementing numerical outputs with verbal and visual representations. It also suggests the use of appropriate metaphors and analogies. Both usability and interpretability suggest making a connection between concepts that are likely to be familiar to educators and the outputs and parameters of the model and statistical terminology. Moreover, the use of a unifying conceptual framework will help educators manage the complexity of the measurement exercise. The objective that the tool should provide *useful* measurement in realistic educational settings leads to adopting a linear measurement approach that is suitable for small class sizes and convenience

samples. Taken together, these additional requirements should make the implementation of a formal objective measurement model accessible to educators, effective, and practical.

In summary, the aim is to develop an expert system which implements the measurement model and hypothesis tests as described earlier and uses:

- a declarative paradigm,
- concepts and terms familiar to educators,
- a linear measurement approach,
- multiple representations of output information,
- appropriate metaphors and analogies, and
- a unifying conceptual framework

4.3. DESIGN AND DEVELOPMENT

Much of the development of modern computer software may be considered routine professional practice, and this aspect is therefore not elaborated herein. However, an *expert system* also requires the use of heuristics and encoded domain knowledge, which are not part of routine practice. There is also a need for the software to use a unifying conceptual framework, appropriate metaphors, and output representations, none of which are routine. The focus in this section is on these non-routine aspects of the development and, in particular, on those relating to human factors and the representational model.

Two unifying elements form the heart of the conceptual framework. The first is the use of an *information theoretic* approach. Thus, for example, the time-series model, the treatment of polytomous items, and the corrections applied when violations of the measurement hypotheses are found, are all based on consideration of the proportion of available information that can be trusted to give true information about the imputed model parameters. The second element is the use of *reliability* indices as measures of fitness for purpose. As discussed under hypothesis 13 in the last chapter, any reliability index can be conceptualised as defining the number of statistically distinct strata identified by the measurement exercise. However, the

terminology used, such as reliability coefficients, may not resonate with all educators. Likewise, concepts, such as equating information to the variance (uncertainty) removed by an observation (Shannon, 1948), may not resonate with all educators. Further, measuring information in bits, while natural in computer science, is not common practice in education. A central part of the development is thus the mapping of these terms and concepts to others, such as the pass mark, the expected pass rate, the number of students and the number and nature of the items used, all of which are likely to be more familiar to educators.

There is also a need to align the measurement model with established educational practice. The specified measurement model works naturally at the item threshold level. However, it is more convenient for an educator to work at the item level. For example, it is more natural for an educator to think in terms of easy and difficult items than in terms of easy and difficult cut points or thresholds within items. Thus, the first level of mapping needed is to develop appropriate representations at the item level that are equivalent to the item threshold measures and statistics. The first subsection below discusses how such item level measurements and statistics can be derived from the associated threshold level numbers.

Likewise, there is a need for appropriate representations at the subject level. The measurement model can work directly at the subject level. However, when a time-series is used, multiple observations (cases) are associated with each subject and there is a need to deal with these at the aggregate subject level. The second subsection introduces the subject level representations.

The use of subject level and item level representations allows an educator to focus mainly on student and item information, with more detailed threshold and case information available if there is a need to drill down to a greater depth in analysis. Ultimately, however, an educator needs to decide how much trust should be placed in the output measurements produced by the measurement exercise. As detailed in the last chapter, there are many hypothesis tests associated with the measurement exercise and a unifying conceptual framework for testing the hypotheses was presented. This framework set out the relationships between the various tests,

leading ultimately to assurance of fitness for purpose. In practice, however, some violation of hypotheses is to be expected and judgement is required as to the impact of any such violation on the output measurements. Moreover, the details of each test vary and it is not realistic to expect all educators to be familiar with all the subtleties of each test. What is needed is a unifying overall framework for the presentation and analysis of the results. The third subsection introduces the use of reliability as a unifying concept for specifying purpose and describes the *traffic light metaphor*, which is used to communicate the outcomes of the measurement process. This metaphor allows an educator to specify purpose by choosing from a number of profiles (which correspond to the number of statistical strata required), and enables fitness for purpose to be evaluated automatically against this choice of profile, and communicated clearly.

However, evaluation of the success, or otherwise, of a measurement exercise is not enough: educators need to know what they should do to remedy any failure, and more important, what steps they should take before the exercise to ensure that success is likely. The fourth subsection introduces the conceptual model used to address this. Building on the central role of reliability in determining fitness for purpose, it maps the measurement and statistical terminology to terms that are likely to be familiar to an educator. This mapping enables guidance to be given to educators on the number of subjects and items, and the mix and nature of items, that are required to achieve measurement success. Thus, an educator can make the necessary arrangements at the planning stage of a measurement exercise, rather than waiting for analysis of results. Moreover, from a perspective of continuous improvement, the mapping helps an educator understand what changes can be made to improve quality in future administrations. The provision of an evaluative report, in narrative form (see Appendix B), is particularly useful for this last purpose.

Any expert system requires the use of heuristics and encoded domain knowledge. Inevitably, some of these heuristics may be only approximate, and encoded domain knowledge may not apply exactly to every situation. There is thus a risk that an educator might be misled by some of the detailed evaluation, and that expert

opinion on the specific situation would provide better guidance. Consequently, it might be considered safer, from a software development perspective, to defer such guidance to the user or an associated expert advisor. However, this misses the point of an expert system and such risk should not be an excuse for inaction. Accordingly, the approach taken is to provide reasonable estimates at the planning stage of what might be expected from the measurement exercise, and then to rely on a process of continuous improvement to refine and enhance the instrument and administration. Furthermore, the goal in evaluation is to provide a broad evaluation of the measurement outcome, while also providing sufficient conventional statistics to enable supplementary expert guidance where this is appropriate. The fifth subsection summarises the main limitations of the expert systems approach.

The sixth subsection describes calibration drift, an issue that emerged during the development process. Finally, the seventh subsection summarises the key features of the design and its implementation.

4.3.1. Item Level Measures and Statistics

The measurement model works at the item threshold level. However, it is often more convenient for educators to work with items. Thus, for example, it is more natural for an educator to consider adding an item to a test, or removing one than adding a threshold to an item, or removing one. This subsection describes how equivalent item parameters and statistics can be imputed from the associated threshold figures. The calculations are illustrated with real data taken from a previous study. In that study, an ordinal *Likert* scale was used with categories: N (Never or almost never), R (Rarely), S (Sometimes), O (Often), and A (Always, or almost always). The representational scale used was centred at zero, measured in logits (i.e. the parameter α was 1), and ranged from -4 to +4. Since five categories were used, there were four item thresholds. The modelled difficulties of these thresholds were: -3.22, -1.27, 0.60, and 2.40 logits. The associated standard errors were: 0.51, 0.23, 0.17, and 0.20. Figure 4.1 shows a plot of the probability that a response would be in a category above each of the four thresholds, conditional on subject ability. These curves have the ogival shape that is characteristic of the

logistic function and intersect the 50% probability level on the Y-axis at the modelled threshold difficulties.

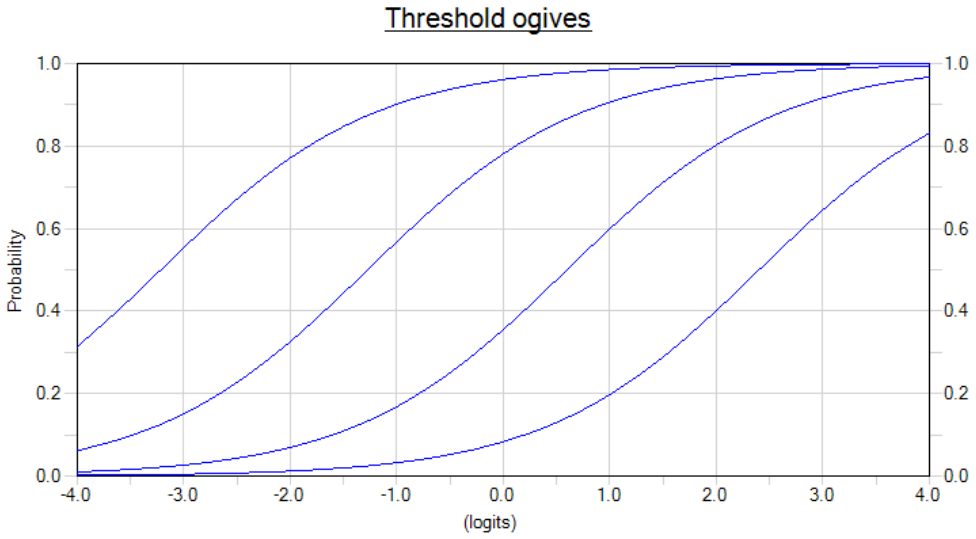


Figure 4.1: Probability of exceeding the item thresholds conditional on imputed case ability

From equation 2.111 (p. 74), the expected proportionate score for an item (*i*) with k categories and thresholds (β_{ij}), conditional on subject ability (θ), is given by:

$$E(score|\theta) = \sum_{j=1}^{k-1} \frac{1/(k-1)}{1 + e^{-\alpha(\theta-\beta_{ij})}} \tag{4.1}$$

Consistent with the interpretation of item thresholds, the difficulty of the item can be defined as the location at which the expected proportional score is 50%; that is, a subject at this ability level is equally likely to score above this level as below. For the item shown, this was at -0.35 logits. A more compact representation of Figure 4.1 is shown in Figure 4.2. This simply shows the categories separated by the thresholds at their modelled locations and the imputed item difficulty as a red mark.

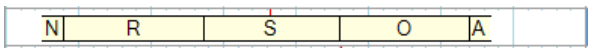


Figure 4.2: Compact representation of item thresholds and item difficulty.

This compact representation encodes the essential information from the probability curves shown in Figure 4.1. The vertical bars separating the categories identify the

scale locations of the thresholds or cut points separating the categories. The open boxes at the end highlight the fact that the extreme categories extend to infinity in both directions. The width of the categories gives an indication of the range of abilities over which the category provides useful information. A narrow category range suggests that too many categories may be being used, giving the illusion of more accuracy than is really there; a wide range suggests that adding a threshold or cut point to subdivide the category could provide additional useful information. For the item used to illustrate these calculations, there were five categories (k = 5) and with the values given above, the expected score, conditional on ability is:

$$E(score|\theta) = \frac{1/4}{1 + e^{-(\theta+3.22)}} + \frac{1/4}{1 + e^{-(\theta+1.27)}} + \frac{1/4}{1 + e^{-(\theta-0.60)}} + \frac{1/4}{1 + e^{-(\theta-2.40)}} \quad (4.2)$$

A plot of this function is given in Figure 4.3. From this figure, it can be seen that this *characteristic curve* is reasonably close to linear over a wider range than for the individual thresholds. This is because the logistic function has maximum information and slope at its midpoint; summing a set of functions with different difficulties spreads this information over a wider range. The imputed item difficulty for this item was -0.35 and it can be seen that the curve crosses the 50% point on the Y-axis at this location.

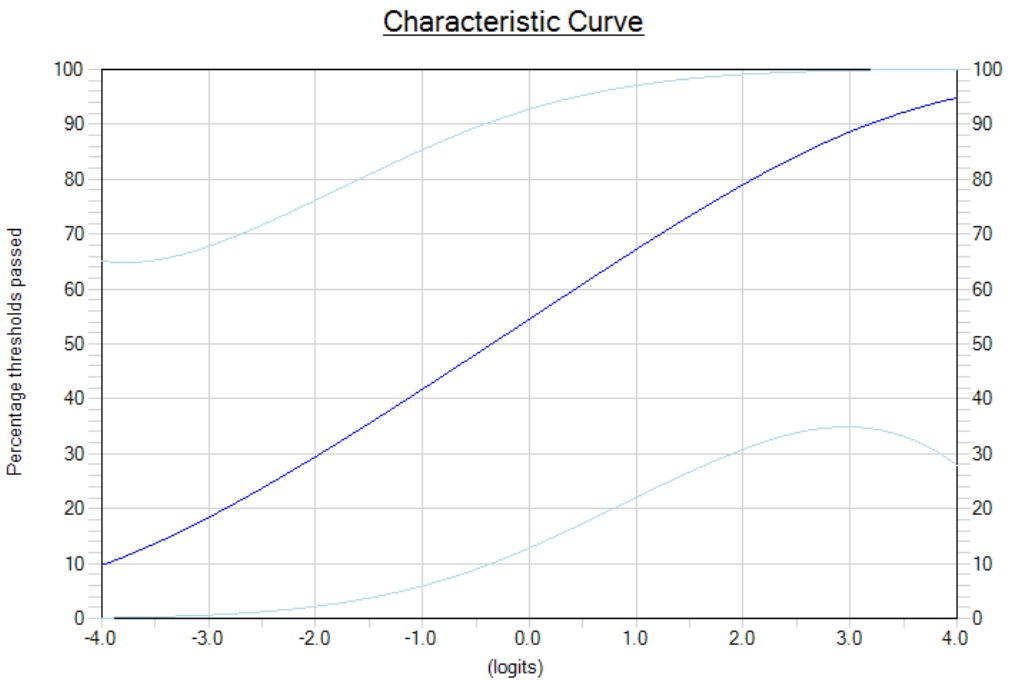


Figure 4.3: Sample item characteristic curve

The standard error associated with this estimate can be imputed by summing the information the sample gives about the item difficulty location and then calculating the associated standard error in accordance with equation 2.126 (p. 83); for this item it was 0.18.

Following equation 2.53 (p. 49), it is also possible to derive from the item threshold difficulties the probability that a response will be made in any given category, conditional on ability. A plot of this is given in Figure 4.4. This figure also shows the maximum likelihood estimates of ability (θ), given the category chosen. These estimates are at: -4, -2.2, -0.34, 1.50, and 4.0, with associated probabilities 0.69, 0.45, 0.44, 0.42 and 0.83, respectively. Thus, for example, a student with an ability estimated at -2 logits on the scale would be modelled as having approximately 44% chance of responding with R, 25% chance of responding with S, 22% chance of responding with N, 5% chance of responding with O and 2% chance of responding with A.

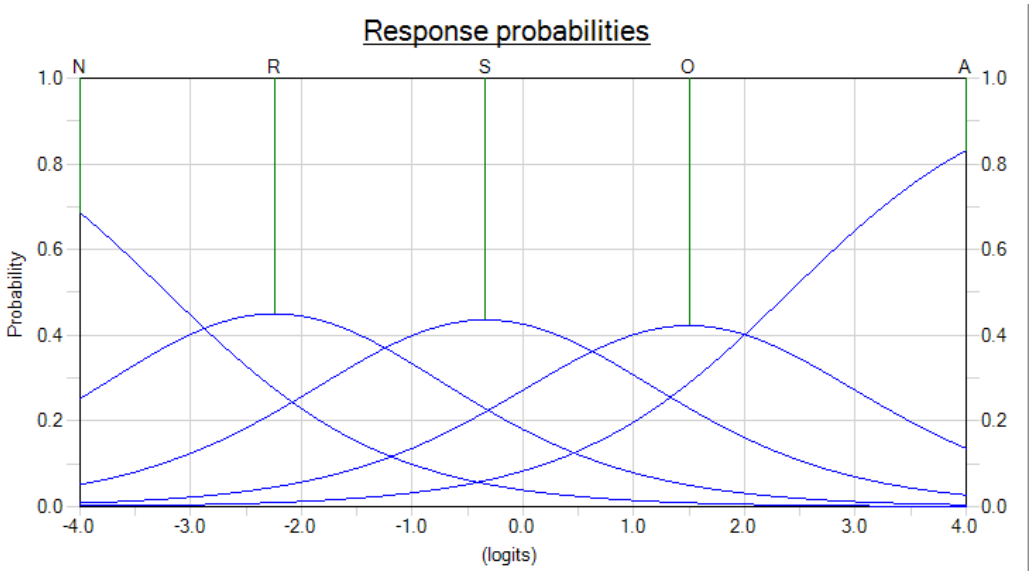


Figure 4.4: Response probabilities for each response category, given a subject case location.

In this diagram, there is a regular progression across the ability scale, with each successive category becoming the most likely response category for a range of abilities. However, this need not be the case. Where a category represents a narrow range of abilities, it might not be the modal category over any range; even at the maximum likelihood ability estimate, it might be more likely that the respondent would choose either of the adjacent categories than the category itself. This is not necessarily a problem, but it suggests that few respondents have chosen this category and, thus, that it may not be contributing much useful information for measurement purposes.

Item level equivalents of the other threshold statistics are set out below. For an item with m categories, indexed by k, definitions are given below for the expected implicit score (E_{nti}), the observed score (X_{nti}), the residual (Y_{nti}), the implicit score associated with each category (S_{ik}), the probability ($\check{P}_{ntik}$) that a response of a subject will be in category k, the modelled variance (V_{nti}) and the standardised residual (Z_{nti}).

$$E_{nti} \stackrel{\text{def}}{=} \sum_{j} P_{ntij} \tag{4.3}$$

$$X_{nti} \stackrel{\text{def}}{=} \sum_j X_{ntij} \tag{4.4}$$

$$Y_{nti} \stackrel{\text{def}}{=} X_{nti} - E_{nti} \tag{4.5}$$

$$S_{ik} \stackrel{\text{def}}{=} k - 1 \tag{4.6}$$

$$\check{P}_{ntik} \stackrel{\text{def}}{=} \begin{bmatrix} 1 - P_{ntik} & k = 1 \\ P_{nti(k-1)} - P_{ntik} & 1 < k < m \\ P_{nti(k-1)} & k = m \end{bmatrix} \tag{4.7}$$

$$V_{nti} \stackrel{\text{def}}{=} \sum_k \left[\check{P}_{ntik}(S_{ik} - E_{nti})^2\right] \tag{4.8}$$

$$Z_{nti} \stackrel{\text{def}}{=} \frac{Y_{nti}}{\sqrt{V_{nti}}} \tag{4.9}$$

From these definitions and the equation (3.38, p. 141), originally given by Wright and Stone (1999, p. 53), *infit* for an item i can be defined as:

$$Infit(i) \stackrel{\text{def}}{=} \frac{\sum_{nt}[Y_{nti}^2]}{\sum_{nt}[V_{nti}]} \tag{4.10}$$

Similarly, from the equation (3.40, p. 141) given for thresholds, where N is the number of case responses summed, *outfit* for an item can be defined as:

$$Outfit(i) \stackrel{\text{def}}{=} \frac{\sum_{nt}[Z_{nti}^2]}{N} \tag{4.11}$$

The interpretation of these statistics is the same as those given for item thresholds. Finally, an item is marked as measurable if it has any measurable thresholds; it is not necessary for all thresholds to be measurable for an item to produce useful measurement. The normal reason for an item being immeasurable is, thus, that all case responses were below all thresholds (i.e. score = 0%) or above all thresholds (i.e. score = 100%). This is consistent with the definition applied to item thresholds.

The above discussion has introduced a set of item level statistics that are conceptually equivalent to those defined at the threshold level. Two verbal labels are also used to aid interpretation of some of these statistics. First, a verbal label is

used to summarise the infit and outfit statistics. The category *immeasurable* has been discussed above. All remaining categories can be used for measurement since the reported standard error is inflated to accommodate the degree of misfit identified. The category *degrading* represents a substantial degree of error inflation and thus limits the quality of the measurement. The remaining categories (Noisy, Unproductive, Productive, and Over-fitting) are all useful for measurement. The definition and interpretation of these categories was given in Table 3.4 (p. 142).

Second, a verbal label is used to indicate whether the case passed the tests of model fit. The label OK is used to indicate that the item passed all tests. Other labels are used to indicate the specific reason for failure. These labels are:

- *No*, when the item is immeasurable.
- *Outliers*, when there were more outliers than expected,
- *Misfit* when the standardised residual (Z_{STD}) statistic (3.36, p. 140) is outside the 95% confidence range.
- *No Fit* when the observed response pattern across subjects does not match that predicted by the model within 95% confidence limits.
- *Weak* when there is insufficient information in subjects' responses to the item to be confident about the measurement. (see hypothesis 8 in section 3.8)
- *Infit* when the infit statistic is larger than the outfit statistic and is greater than 2 (i.e. Degrading).
- *Outfit* when the outfit statistic is not less than the infit statistic and is greater than 2 (i.e. Degrading).

Three additional diagnoses are reported at the item level. These are not based on item threshold data and are specified at the item level because conceptualisation at this level is clearer. First, based on the dimensionality analysis (hypothesis 3) set out in section 3.3, the scale factor with which the item is most strongly associated is reported. When multidimensionality is diagnosed for the scale, this can help identify the items that are associated with that factor. Second, where local dependence is diagnosed by the test of Hypotheses 6 (see section 3.6), the degree

of dependence is reported together with the other items that are involved in the dependency. Third, differential item functioning, where diagnosed by the tests of hypothesis 4 (see section 3.4), is reported, together with the sample groups for which it was diagnosed. These groups include any diagnosed factors in the sample (representing differing interpretations of the item) and any available categorical variables classifying the sample, such as gender, age, etc.

These statistics, labels, and diagnoses allow an educator to work at the item level with the measurements and concepts, rather than at the threshold level. They are sufficient for normal use and interpretation of the measurement exercise. However, to enable additional expert interpretation, some conventional statistics are also calculated:

- *Item-Total Correlation*. This is the correlation across cases of the score for each case on the item with the overall score for the case, excluding the current item. The item-total correlation (r_{IT}) is based on scores rather than measurement. Nevertheless, it is widely used and can give a broad indication of concordance of the item with the scale construct.
- *Loading on principal component*. This is the factor loading of the item on the dominant component established by the Principal Components Analysis (PCA). Like the item-total correlation, the loading on the principal component (r_{PC}) is based on an analysis of scores, so it, too, is less than ideal. However, like the item-total correlation, it can give a broad indication of concordance of the item with the construct.
- *IRT correlation*. This is the correlation across cases of the maximum likelihood ability estimate (see Figure 4.4) for the response category of the case's response with the overall ability estimate for the case. This correlation (r_{IRT}) is based on imputed abilities, rather than scores, and, thus, is in the measurement metric of the instrument.
- *Loevinger's item-rest homogeneity* (H_r) which measures the internal consistency of the item with the rest of the items in the measurement scale.
- *Cronbach's alpha if deleted*. This shows a revised calculation of Cronbach's alpha with the item excluded from the scale, if this would be higher than

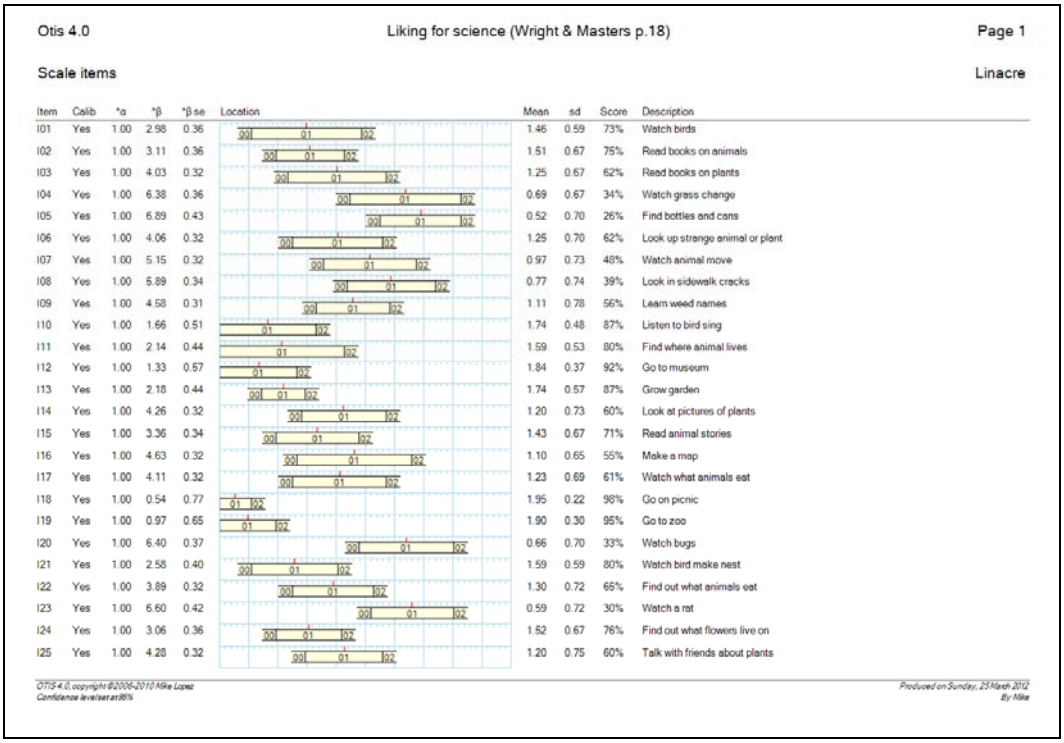
Cronbach's alpha with the item included. Like the item-total correlation, and the loading on the principal component (r_{PC}), this is based on an analysis of scores, so it, too, is less than ideal. However, Cronbach's alpha is widely used as a measure of scale consistency.

All these statistics can be broadly interpreted in a similar way; they give an indication of how well the item fits with the overall construct being measured. Items with negative correlations may degrade measurement, and this may be an indication that the item has been miscoded. Low values suggest that the item is contributing little to the overall measurement. The calculation of what Cronbach's alpha would be, if the item were deleted, is sometimes used as part of a scale building exercise, although it may be preferable to use an approach to scale building which is more theoretically based, such as factor analysis for exploratory measurement, or empirically based, such as Mokken's procedure (Mokken & Lewis, 1982) for general measurement.

This subsection has presented the derivation of item level measures and statistics from the corresponding threshold level figures. It has also introduced a number of qualitative labels that can be used to guide interpretation, and a number of diagnostics that are easier to conceptualise at the item level than the threshold level. Although these are sufficient for measurement evaluation, several conventional statistics have also been set out for those who are familiar with these statistics and their interpretation. The item level representations allow an educator to focus on the item level. More detailed threshold information and conventional statistics are available if there is a need to drill down to a greater depth in analysis.

Two main reports are produced by the software at item level. An example of the first is shown in Figure 4.5. The dataset used to create the report was the *Liking for Science* dataset which is described in more detail later in this chapter. The report sets out the overall measurements produced. The main measures are shown first: the scale unit (α), the estimated item difficulty ($*\beta$), and the standard error ($*\beta$ se). The mapping of the categories and thresholds to the scale metric is then shown graphically using the compact representation introduced above. This is followed by

three conventional measures: the mean and standard deviation of the natural score across subjects, and the average score.


Otis 4.0 Liking for science (Wright & Masters p.18) Page 1

Scale items Linacre

Item	Calib	*α	*β	*β se	Location	Mean	sd	Score	Description
I01	Yes	1.00	2.98	0.36		1.46	0.59	73%	Watch birds
I02	Yes	1.00	3.11	0.36		1.51	0.67	75%	Read books on animals
I03	Yes	1.00	4.03	0.32		1.25	0.67	62%	Read books on plants
I04	Yes	1.00	6.38	0.36		0.69	0.67	34%	Watch grass change
I05	Yes	1.00	6.89	0.43		0.52	0.70	26%	Find bottles and cans
I06	Yes	1.00	4.06	0.32		1.25	0.70	62%	Look up strange animal or plant
I07	Yes	1.00	5.15	0.32		0.97	0.73	48%	Watch animal move
I08	Yes	1.00	5.89	0.34		0.77	0.74	39%	Look in sidewalk cracks
I09	Yes	1.00	4.58	0.31		1.11	0.78	56%	Learn weed names
I10	Yes	1.00	1.66	0.51		1.74	0.48	87%	Listen to bird sing
I11	Yes	1.00	2.14	0.44		1.59	0.53	80%	Find where animal lives
I12	Yes	1.00	1.33	0.57		1.84	0.37	92%	Go to museum
I13	Yes	1.00	2.18	0.44		1.74	0.57	87%	Grow garden
I14	Yes	1.00	4.26	0.32		1.20	0.73	60%	Look at pictures of plants
I15	Yes	1.00	3.36	0.34		1.43	0.67	71%	Read animal stories
I16	Yes	1.00	4.63	0.32		1.10	0.65	55%	Make a map
I17	Yes	1.00	4.11	0.32		1.23	0.69	61%	Watch what animals eat
I18	Yes	1.00	0.54	0.77		1.95	0.22	98%	Go on picnic
I19	Yes	1.00	0.97	0.65		1.90	0.30	95%	Go to zoo
I20	Yes	1.00	6.40	0.37		0.66	0.70	33%	Watch bugs
I21	Yes	1.00	2.58	0.40		1.59	0.59	80%	Watch bird make nest
I22	Yes	1.00	3.89	0.32		1.30	0.72	65%	Find out what animals eat
I23	Yes	1.00	6.60	0.42		0.59	0.72	30%	Watch a rat
I24	Yes	1.00	3.06	0.36		1.52	0.67	76%	Find out what flowers live on
I25	Yes	1.00	4.28	0.32		1.20	0.75	60%	Talk with friends about plants

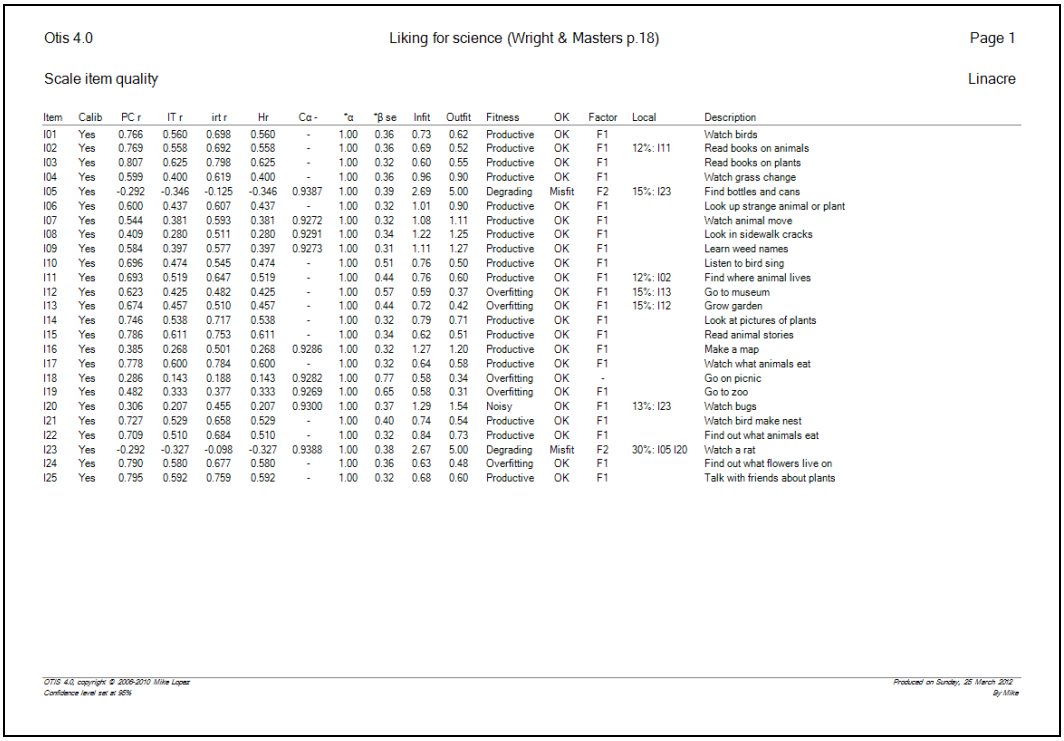
OTIS 4.0, copyright ©2006-2010 Mike Lopez Produced on Sunday, 25 March 2012
Confidence level set at 95% By: Mike

Figure 4.5: An item level report from the software

From this report, it can be seen that items 18 and 19 are relatively easy to endorse; the highest category (02) is the most likely response for a wide range of scale values. Similarly, items 5 and 23 are relatively hard to endorse; only those respondents in the upper half of the scale are likely to choose any category other than the lowest (00).

A separate report is used to show the various quality control statistics. An example of this report from the same dataset is shown in Figure 4.6. Of particular note are items 5 (find bottles and cans) and 23 (watch a rat) which are diagnosed as inadequate on multiple measures: the four correlational measures are all negative; Cronbach's alpha would be improved if the item were removed; the fitness category is labelled as degrading; and model fit is categorised as misfit. It is therefore questionable whether these items really contribute to the notion of a "Liking for science". It can also be noticed that these items load on a second factor, which further supports the idea that they measure something other than a liking for

science. The remaining items all appear to contribute usefully to overall measurement. Items 2 and 11 show mutual local dependence, presumably attributable to an interest in animals over and above that associated with a liking for science. Likewise, items 5, 20 and 23 (find bottles and cans, watch bugs and watch a rat) show mutual local dependence, as do items 12 and 13 (go to museum and grow a garden). Although these items are not degrading, they nevertheless point to ways in which the measure could be improved for future administration.



Otis 4.0 Liking for science (Wright & Masters p.18) Page 1

Scale item quality Linacre

Item	Calib	PC r	IT r	irt r	Hr	Cα -	*α	*β se	Infit	Outfit	Fitness	OK	Factor	Local	Description
I01	Yes	0.766	0.560	0.698	0.560	-	1.00	0.36	0.73	0.62	Productive	OK	F1		Watch birds
I02	Yes	0.769	0.558	0.692	0.558	-	1.00	0.36	0.69	0.52	Productive	OK	F1	12%: I11	Read books on animals
I03	Yes	0.807	0.625	0.798	0.625	-	1.00	0.32	0.60	0.55	Productive	OK	F1		Read books on plants
I04	Yes	0.599	0.400	0.619	0.400	-	1.00	0.36	0.96	0.90	Productive	OK	F1		Watch grass change
I05	Yes	-0.292	-0.346	-0.125	-0.346	0.9387	1.00	0.39	2.69	5.00	Degrading	Misfit	F2	15%: I23	Find bottles and cans
I06	Yes	0.600	0.437	0.607	0.437	-	1.00	0.32	1.01	0.90	Productive	OK	F1		Look up strange animal or plant
I07	Yes	0.544	0.381	0.593	0.381	0.9272	1.00	0.32	1.08	1.11	Productive	OK	F1		Watch animal move
I08	Yes	0.409	0.280	0.511	0.280	0.9291	1.00	0.34	1.22	1.25	Productive	OK	F1		Look in sidewalk cracks
I09	Yes	0.584	0.397	0.577	0.397	0.9273	1.00	0.31	1.11	1.27	Productive	OK	F1		Learn weed names
I10	Yes	0.696	0.474	0.545	0.474	-	1.00	0.51	0.76	0.50	Productive	OK	F1		Listen to bird sing
I11	Yes	0.693	0.519	0.647	0.519	-	1.00	0.44	0.76	0.60	Productive	OK	F1	12%: I02	Find where animal lives
I12	Yes	0.623	0.425	0.482	0.425	-	1.00	0.57	0.59	0.37	Overfitting	OK	F1	15%: I13	Go to museum
I13	Yes	0.674	0.457	0.510	0.457	-	1.00	0.44	0.72	0.42	Overfitting	OK	F1	15%: I12	Grow garden
I14	Yes	0.746	0.538	0.717	0.538	-	1.00	0.32	0.79	0.71	Productive	OK	F1		Look at pictures of plants
I15	Yes	0.786	0.611	0.753	0.611	-	1.00	0.34	0.62	0.51	Productive	OK	F1		Read animal stories
I16	Yes	0.385	0.268	0.501	0.268	0.9286	1.00	0.32	1.27	1.20	Productive	OK	F1		Make a map
I17	Yes	0.778	0.600	0.784	0.600	-	1.00	0.32	0.64	0.58	Productive	OK	F1		Watch what animals eat
I18	Yes	0.286	0.143	0.188	0.143	0.9282	1.00	0.77	0.58	0.34	Overfitting	OK	-		Go on picnic
I19	Yes	0.482	0.333	0.377	0.333	0.9269	1.00	0.65	0.58	0.31	Overfitting	OK	F1		Go to zoo
I20	Yes	0.306	0.207	0.455	0.207	0.9300	1.00	0.37	1.29	1.54	Noisy	OK	F1	13%: I23	Watch bugs
I21	Yes	0.727	0.529	0.658	0.529	-	1.00	0.40	0.74	0.54	Productive	OK	F1		Watch bird make nest
I22	Yes	0.709	0.510	0.684	0.510	-	1.00	0.32	0.84	0.73	Productive	OK	F1		Find out what animals eat
I23	Yes	-0.292	-0.327	-0.098	-0.327	0.9388	1.00	0.38	2.67	5.00	Degrading	Misfit	F2	30%: I05 I20	Watch a rat
I24	Yes	0.790	0.580	0.677	0.580	-	1.00	0.36	0.63	0.48	Overfitting	OK	F1		Find out what flowers live on
I25	Yes	0.795	0.592	0.759	0.592	-	1.00	0.32	0.68	0.60	Productive	OK	F1		Talk with friends about plants

OTIS 4.0, copyright © 2008-2010 Mike Lopez
Confidence level set at 95%

Produced on Sunday, 25 March 2012
By Mike

Figure 4.6: Example item quality control statistics

From an expert system perspective, the essential feature of the quality control report shown in Figure 4.6 is the summary of the detailed statistics in two qualitative labels. Ideally, all items should be classified as *productive* for measurement and *OK* for model fit. However, some deviation from this is still useful for measurement in practice. In particular, all classifications other than *degrading* represent departures from the ideal that are, nevertheless, still useful for measurement. The second label gives more detail on the specific reason for inadequate model fit and the remaining statistics give the necessary underlying statistics to support expert interpretation. In broad terms, the guidance for educators is to remove any *degrading* items from the scale before using the results

and to consider replacing or rewording items with other diagnosed issues in future administrations.

4.3.2. Subject Level Measures and Statistics

As discussed in the last section, a polytomous item has several thresholds. When a time-series is used, a similar relationship exists between a subject and the associated cases. These cases characterise the set of observations relating to the subject at a particular point in time. However, while interpretation is more natural for an educator at the item level than at the threshold level, the opposite is true for subjects and cases. Indeed, the main reason for the use of a time series is a focus on the changes in the case estimates over time. Accordingly, in this sub-section, the representations at the subject case level are presented first; in the absence of a time series, these are all that is required. This is then followed by a discussion of some special considerations that apply to a time-series.

At the subject case level, the basic measures and statistics were presented in chapter two. However, as with items, two verbal labels are used to aid interpretation. First, a verbal label is used to summarise the infit and outfit statistics. The definition and interpretation of the categories used was given in Table 3.4 (p. 142). The classification *immeasurable* identifies cases for which measurement has failed; this category is usually associated with those judged above all thresholds (all responses in the highest response category) or below all thresholds (all in the lowest category). All remaining classifications can be used for measurement since the reported standard error is inflated to accommodate the degree of misfit identified. The classification *degrading* represents substantial error inflation and thus limits the quality of the measurement. The remaining classifications (noisy, unproductive, productive, and over-fitting) are all useful for measurement.

A second verbal label is used to indicate whether the case passed the tests of model fit. The label OK is used to indicate that the case passed all tests. Other wording is used to indicate the specific reason for failure. The labels are the same as those used for items:

- *No*, when the case is immeasurable.
- *Outliers*, when there were more outliers than expected,
- *Misfit* when the standardised residual (Z_{STD}) statistic (3.36, p. 140) is outside the 95% confidence range.
- *No Fit* when the observed response pattern across subjects does not match that predicted by the model within 95% confidence limits.
- *Weak* when there is insufficient information in the subject's responses to be confident about the measurement (see hypothesis 8 in section 3.8)
- *Infit* when the infit statistic is larger than the outfit statistic and is greater than 2 (i.e. Degrading).
- *Outfit* when the outfit statistic not less than the infit statistic and greater than 2 (i.e. Degrading).

Three additional diagnostics are reported. First, based on the dimensionality analysis (hypothesis 3) set out in section 3.3, the scale factor with which the case is most strongly associated is reported. When multidimensionality is diagnosed for the scale, this can help identify the subjects who are associated with that factor; each factor suggests a shared interpretation of the items in the scale. Second, where local dependence is diagnosed by the test of Hypotheses 6 (see section 3.6), the degree of dependence is reported together with the other subjects that are involved in the dependency. Local dependence among subjects suggests a commonality of understanding or misunderstanding. This typically occurs when students discuss and develop their understandings with other members of a peer group over a period of time, and the understandings of the peer group are not aligned with those of the class overall. Third, response set, where diagnosed by the tests of hypothesis 4 (see section 3.4), is reported.

A sample of the representation, taken from the Liking for Science dataset, used later in this chapter, is shown in Figure 4.7; a larger version of this figure is shown in Appendix C. The report sets out the overall measurements produced. The main measures are shown first: the estimated subject ability (θ), and the standard error (See). These are then shown graphically, with error bars denoting the 95%

confidence interval; the colour coding used is explained in the next sub-section. This is followed by the subject's raw score, expressed as a percentage, and a number of diagnostics and statistics.

Otis 4.0 Liking for science (Wright & Masters p.18) Page 1

Scale cases Linacre

Case	Calib	Theta	See	Location	Score	Infit	Outfit	Fit	OK	Factor	Local	Set	Algorithm	KID
1	Yes	4.580	0.44		60%	1.01	0.92	Productive	OK	F1		-	Newton	Marc R
2	No	10.000	1.18		100%	0.00	0.00	Immeasurable	No	-		-	Newton	Lawrence R
3	Yes	5.168	0.42		68%	0.70	0.45	Overfitting	OK	F1		-	Newton	Toby R
4	Yes	4.071	0.42		54%	0.79	0.67	Productive	OK	F1		-	Newton	Michael R
5	Yes	2.888	0.42		38%	1.10	1.08	Productive	OK	F1		-	Newton	Rebecca R
6	Yes	3.458	0.43		48%	1.53	5.00	Degrading	Outfit	F1		-	Newton	TR Cat R
7	Yes	7.791	0.71		88%	1.71	2.94	Degrading	Outfit	F1		-	Newton	Benjamin W
8	Yes	5.038	0.46		66%	1.02	0.86	Productive	OK	F1		-	Newton	Ross L
9	Yes	3.556	0.43		48%	1.49	3.11	Degrading	Outfit	F1		-	Newton	Matthew S
10	Yes	4.172	0.42		56%	0.91	0.70	Productive	OK	F1		-	Newton	Daniel H
11	Yes	5.433	0.47		70%	1.02	0.92	Productive	OK	F1		-	Newton	Paul H
12	Yes	2.354	0.46		34%	1.81	5.00	Degrading	Outfit	-		-	Newton	Daniel L
13	Yes	3.658	0.43		48%	1.03	0.74	Productive	OK	F1		-	Newton	Benjamin L
14	Yes	3.399	0.46		46%	1.61	3.10	Degrading	Outfit	F1		-	Newton	Nancy H
15	Yes	3.945	0.40		54%	1.21	2.30	Degrading	Misfit	F1	55%: Newell S, Raymond N	-	Newton	Stephie D
16	Yes	7.080	0.55		86%	0.56	0.29	Overfitting	OK	F1	29%: Alan L	-	Newton	Marge B
17	Yes	8.701	0.81		94%	1.13	0.65	Productive	OK	F1	71%: Billy M, James R, Ruth P, Zippy A	-	Newton	Gail S
18	Yes	6.628	0.54		82%	0.97	0.62	Productive	OK	F1	33%: Boris B	-	Newton	Richard E
19	Yes	4.915	0.45		64%	0.94	0.76	Productive	OK	F1		-	Newton	Hames F
20	Yes	3.813	0.44		52%	1.04	1.09	Productive	OK	F1		-	Newton	Rob A
21	Yes	4.201	0.40		56%	0.46	0.31	Overfitting	OK	F1		-	Newton	Norm E
22	Yes	5.328	0.42		70%	0.53	0.30	Overfitting	OK	F1		-	Newton	Kathleen H
23	Yes	6.269	0.47		80%	0.64	0.31	Overfitting	OK	F1		-	Newton	Jeff V
24	Yes	4.837	0.44		64%	0.94	0.78	Productive	OK	F1		-	Newton	Elizabeth T
25	Yes	6.928	0.56		84%	0.87	0.93	Productive	OK	F1		-	Newton	Karen S

OTIS 4.0, copyright © 2006-2010 Mike Lopez Produced on Sunday, 25 March 2012
Confidence level set at 95% By Mike

Figure 4.7: Sample measurement representation at the case level

As with items, the essential feature, from an expert system perspective, is the summary of the detailed statistics in two qualitative labels. Ideally, all cases should be classified as *productive* for measurement and *OK* for model fit. Nevertheless, some deviation from this is still useful for measurement in practice; all classifications other than *degrading* represent departures from the ideal that are still useful for measurement. The second label gives more detail on the specific reason for inadequate model fit and the remaining statistics give the necessary underlying statistics to support expert interpretation. As with items, the broad guidance for educators is to remove any *degrading* cases from the scale before using the results, if accurate item classification is needed.

As discussed in the introduction to this sub-section, the main focus with a time-series is on the individual case estimates. However, it is also useful for an educator and learner to develop a sense of how these estimates develop over time.

Figure 4.8 shows a graphical depiction of case ability estimates for a subject; this graph was taken from one of the studies used for the empirical evaluation presented in chapter seven.

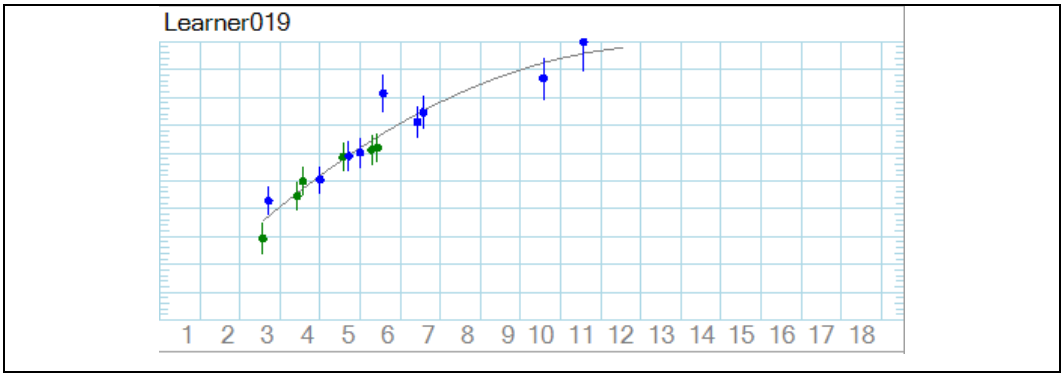


Figure 4.8: A graphical depiction of a time-series.

In this graph, the horizontal axis represents time, expressed in weeks, and the vertical axis represents estimated ability. Each case estimate, together with error bars corresponding to a 95% confidence interval, is shown at the time of the observation. The light grey line shows a best-fit second degree polynomial to the observations. It can be noticed that this line is within the 95% confidence limits of most of the estimates, but there is an outlier estimate in week 6.

The use and analysis of a time-series will be reviewed in more detail in Chapter Seven; the intent in this subsection is to illustrate the use of a graphical approach to provide a compact representation of a substantial amount of data.

4.3.3. Evaluating Results and the Traffic Light Metaphor

Although the diagnostic labelling and graphical representations introduced in the last two subsections allow an educator to work at the subject and item level, and to interpret results at this level, there is nevertheless a substantial volume of data for an educator to absorb and interpret. A central goal of the software implementation is that outputs should be usable, and interpretable, by educators and learners. There are many hypothesis tests and a unifying approach is required to help manage this complexity. At the most basic level, a hypothesis is either rejected or not. However, in practice, there may also be a "grey" area in which the outcome is marginal. Moreover, judgement is required as to the impact of any failure on the

measurements produced. Essentially, there is a need to determine how much confidence should be placed in any measurements produced, and to express this simply and in terms familiar to an educator. To achieve this, interpretation in this project is supported by a "traffic light" metaphor with the basic meaning of: do not proceed (Red); proceed with caution (Amber); or proceed safely (Green). From the perspective of a process of continuous improvement, however, the determination that an output is safe (Green) may be only marginal. The metaphor has therefore been extended with an additional category (Blue) to indicate those outputs that are not marginal. Both green and blue indications thus suggest that it is safe to proceed, but a green indication suggests that safety is marginal and that it may be possible to improve the estimate. This distinction is supported by the mnemonic "Green is Good, Blue is Better".

These quality categories used are summarised in Table 4.1.

Table 4.1: The "Traffic light" metaphor

Category	Colour	Action	Interpretation
Failed	Red	Do not proceed	The measurement process has failed; the resulting measurement(s) should not be used.
Poor	Amber	Proceed with caution.	Confidence in the measurement is low.
Good	Green	Proceed safely.	Quality is acceptable but could be improved.
Excellent	Blue	Proceed safely.	No need for improvement is identified.

Note that, for accessibility reasons, colour is used sparingly in the software, and it is never used as the sole indication of an issue, but rather to highlight issues that are indicated elsewhere with qualitative labels.

This metaphor is used both at the overall instrument level and also at the individual subject and item level, where it highlights those items and subjects that are problematic for measurement and should be reviewed. For some of the tests, the quality category to be assigned follows directly from the logic of the hypothesis test. For example, failure to achieve a common metric (H9) results in the classification of all outputs as failed. For others, however, the category is based on

the value of some statistic. Where there is a continuous statistic governing quality, three cut points are used to map the underlying continuum to the categories. The first cut point (reject) is set at the minimum level consistent with acceptance of the associated hypothesis; measurement will be rejected for statistics below this value. The second cut point (fit) is set at the minimum level deemed *fit for purpose*. The third cut point (safe) is set, arbitrarily, at a level where a 10% loss of information would correspond to the fit for purpose level. The relationship between cut points and categories is summarised in Figure 4.9.

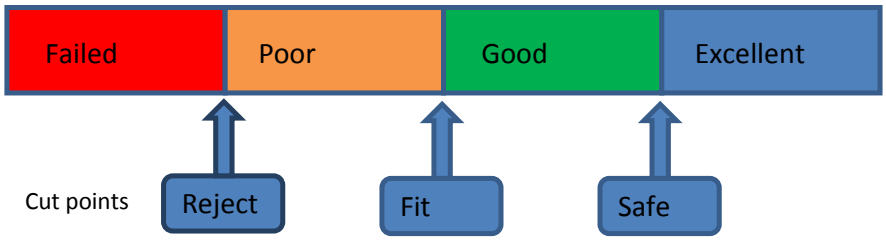


Figure 4.9: Cut points for the traffic light metaphor

The quality category applies both at the overall instrument level, and to individual measurements. At each level, several tests are carried out and each test limits the maximum quality category that will be assigned. Moreover, the quality category at the overall instrument level sets an upper limit to the quality category assigned to each individual measurement. The assigned category thus represents a minimum of the imputed quality categories across all relevant tests.

Five of the hypotheses introduced in the last chapter are critical for measurement. These are: that the construct is quantifiable (H1); that response patterns are ordinal (H2); that the measurement model converges correctly (H7); that data are adequate for measurement (H8); and that a common metric has been achieved for all measurements (H9). If any of these hypotheses fail, it is clear that the failure should lead to rejection of the measurements and, accordingly, the quality category is set to indicate failure.

Four of the hypotheses establish whether subjects are responding to items in the intended manner. These are: that the construct is unidimensional (H3); that there is

no differential item functioning (H4); that there is no subject response set (H5); and that there is local independence (H6). Failure of any of these hypotheses need not invalidate the output measurements, but judgement is required as to the impact of any rejected hypothesis. However, the software also has an option to manage failure of each of these hypotheses. If this option is taken, no further action is required with regard to the quality category because the management process will reduce the information used from the dataset, and the quality category will be automatically adjusted if the resultant loss of accuracy affects the quality. However, if the option is not taken, the maximum quality category is limited to Good (Green), to signal the possibility of improvement.

Three of the hypotheses establish that the model adequately represents the input data. These are: that response patterns are reproducible from measurements (H10); that there are no more outliers than expected (H11); and that fit statistics accord with theoretical expectations (H12). Where cases or items are found to misfit, the standard error is adjusted appropriately and, where a standard error inflator is required, the maximum quality category is limited to Good (Green), thus signalling the possibility of improvement. No further action is required with regard to the quality category because the quality category will be automatically adjusted if the resultant loss of accuracy brings the quality below the fit for purpose levels.

For the remaining hypothesis, that there is adequate reliability for useful measurement (H13), the failed quality category will be assigned if the reliability is not greater than zero, but the remaining categories depend on the reliability achieved and the measurement purpose.

It can be seen that a reliability coefficient may be used both to evaluate fitness for purpose and to provide a unifying framework for the various hypotheses. However, interpreting a reliability coefficient may not be intuitive for all educators. To achieve a more intuitive measure, the evaluation of measurement success can be based on an equivalent conceptualisation in terms of accuracy of measurement. This equivalence has long been known; for example, Anastasi and Urbina note that “the standard error of measurement and the reliability coefficient are obviously

alternative ways of expressing test reliability" (1997, p. 108). In turn, accuracy of measurement can be standardised and expressed as a *dimensionless* quantity by dividing the standard error of measurement by the standard deviation of the sample. This enables the reliability coefficient to be reconceptualised simply in terms of the standardised standard error of estimate (σ_Z^2).

$$\mathrm{r}_{tt} \stackrel{\text{def}}{=} 1 - \sigma_Z^2 \tag{4.12}$$

Moreover, accuracy of measurement can be expressed directly in terms of the number of statistically distinct strata required (H). This conceptualisation allows the educator simply to specify the number of strata required and the software can then carry out the necessary calculations to translate this into the corresponding reliability coefficients. The fit for purpose reliability cut point (R_{fit}) can be calculated from the number of strata (H) using equations 3.62 (p. 150), 3.83 (p. 150) and 3.64 (p. 150) as follows:

$$\mathrm{G} \stackrel{\text{def}}{=} (3H - 1)/4 \tag{4.13}$$

$$R_{fit} = \frac{G^2}{1 + G^2}$$

The corresponding safe level (R_{safe}) can be calculated as:

$$R_{safe} = 0.1 + 0.9\, R_{fit} \tag{4.14}$$

Since the assumptions and heuristics depend on the measurement purpose, the software supports a number of profiles, one for each purpose, where each profile identifies the number of statistically distinct strata required. Clearly, the minimum number of strata for useful measurement is two; less than this cannot usefully separate the objects being measured. A more practical minimum is three. Drawing on the concept from the medical and health field, this profile is named *triage*. Strictly, the term triage (from the French verb *trier*, to separate) refers to any classification into strata. However, it is used here, loosely, to mean classification into three strata. Feedback from the group of educators who trialled the software suggests that this was easily understood. Moreover, the group believed it to be the most useful profile for general use. Accordingly, this was made the default profile in

the software. None of the group believed that the minimal separation profile would be of use to them, but all believed that it should be included for completeness. The remaining profiles were based on consensus among the group. The profiles provided by the software, the corresponding number of strata, and the associated reliability cut points are summarised in Table 4.2.

Table 4.2: The number of strata and reliability cut points for selected purposes

		Reliability cut points	
Purpose	Strata	Fit	Safe
No separation required	0	0.000	0.000
Minimal separation	2	0.610	0.649
Triage	3	0.800	0.820
Basic grading or classification	5	0.925	0.932
Detailed grading or classification	10	0.981	0.983

These strata and cut points apply both to items and subjects. From an expert systems perspective, the essential idea is that an educator can express purpose by choosing one of these profiles, and the software can then carry out the necessary translation into statistical terms, evaluate success in statistical terms, and then report back to the educator simply and directly: do not proceed, proceed with caution, or proceed safely.

This section has introduced the traffic light metaphor as a means of communicating the success of the measurement exercise, together with an evaluation of the quality of the measurements achieved. It has also introduced the reliability coefficient as a unifying framework for the evaluation of fitness for purpose. Finally, a conceptual model has been presented that allows an educator to specify purpose in terms of the number of statistically distinct strata required, and for the software to carry out the necessary translation into statistical measures.

4.3.4. Planning for Reliability

The previous section described how purpose can be specified in terms of required statistical strata, and success communicated in terms of the traffic light metaphor. The evaluation, however, is retrospective and an educator may wish to know in advance what should be done to achieve the target reliability. This depends on several factors and the intent in this section is to disentangle these in order to provide practical guidance. The key factors are the pass mark, the expected pass rate, the number of students, and the number and nature of the items used. The conceptual relationships between these are summarised in Figure 4.10.

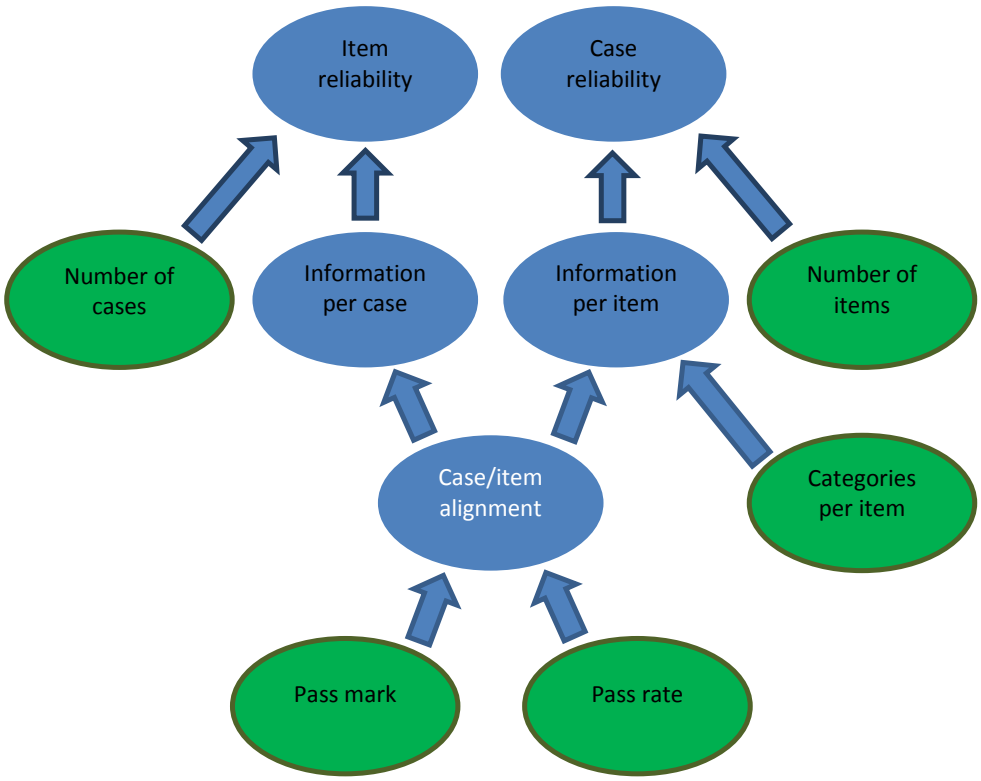


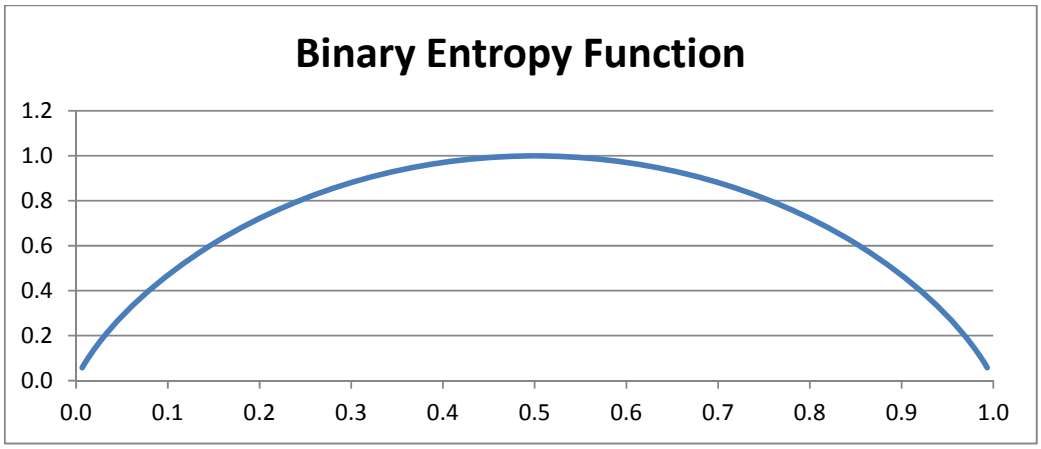
Figure 4.10: Interrelation of statistical terms and educational concepts

In this figure, concepts that are likely to be familiar to educators are shaded in green, and those that relate to technical or statistical terminology are shaded in blue.

Working from the top of this diagram, item reliability expresses the quality of items estimates, and, likewise, case reliability expresses that of case estimates. As discussed in the last section, it can be noted that reliability depends on the accuracy

of measurement, or equivalently on the total information available. Thus, item reliability depends on the total information supplied by the cases, and case reliability depends on the total information supplied by the items. For the next row of the diagram, it is useful to express the total case information as the product of the number of cases and the average information supplied by each case. Similarly, the total information supplied by items can be expressed as the product of the total number of items and the average information supplied by each item.

From a technical perspective, the use of an average is slightly conservative: the information contributed by an observation is higher when the objects are close on the underlying metric than when they are further apart. The actual information contributed follows the binary entropy function (Figure 4.11).



Note: The y axis shows the entropy (H_b) of a Bernoulli trial for each probability of success identified on the x axis.

Figure 4.11: The binary entropy function

From the *inverted-U* shape of this function, it is clear that a line connecting any two points on the curve will fall below the curve, and thus, that an average will underestimate the actual information. Nevertheless, from an expert systems perspective, taking a conservative position is useful, and the decomposition is readily understood by both educators and learners. This allows the number of cases (sample size) and number of items (test size) to be separated from consideration of the information provided by each case and item.

The information provided for an item estimate is the sum across cases of the information provided by the observations, and, likewise, the information provided for a case estimate is the sum across items of the information provided by the observations. Both estimates thus depend on the information provided by the individual observations. The information provided by an individual observation depends on the alignment between the associated case ability and the relevant item threshold difficulties. It is at a maximum where these are close on the metric, and falls off increasingly as the distance between these increases. Thus, the information derived from summing these observations depends on the case/item alignment, as shown in the diagram.

However, there is an asymmetry between cases and items in the information provided. Whereas the information provided by a case about an item threshold depends only on the case ability, the information provided by an item about a case also depends on the number of categories available for classification. Precise determination of the amount of information provided is complex and depends on the distribution of responses among the categories and the abilities of the subjects in each category. However, reasonable estimates have been established by simulation, as described in Chapter Five, and are built-in to the software as heuristics.

The alignment between case ability and item difficulty is specified naturally in logits, but this is not a natural unit for an educator. However, the distance can be specified equivalently in terms of pass rate and pass mark. Let π_M represent the pass mark: the proportion of items on which a candidate must succeed to pass. Let π_R denote the pass rate: the proportion of candidates who are expected to succeed. Let δ_M represent the distance in scale units between the mean threshold ability and the cut point used for the passing criterion. Let δ_R denote the distance in scale units between the mean case ability and the notional ability cut point separating those who pass from those who do not. Let α denote the scale unit used in the model. The distances can then be estimated as follows.

$$\delta_M \cong \frac{1}{\alpha}\log_e\left(\frac{\pi_M}{1-\pi_M}\right) \tag{4.15}$$

$$\delta_R \cong \frac{1}{\alpha}\log_e\left(\frac{\pi_R}{1-\pi_R}\right) \tag{4.16}$$

The distance in scale units between the mean threshold difficulty and the mean case ability is thus $\delta_M + \delta_R$.

$$\delta \cong \frac{1}{\alpha}\log_e\left(\frac{\pi_M\pi_R}{(1-\pi_M)(1-\pi_R)}\right) \tag{4.17}$$

The conceptual framework presented here has set out how the statistical and information theoretic terminology can be mapped to terms and concepts that are more natural to an educator: the number of subjects, the number of items, the number of categories in each item, pass marks, and pass rates. However, more specific guidance can be given on the first three of these. For purposes of exposition, the concept of a *perfectly targeted dichotomous response* is introduced. Although such a response can never be achieved in practice, it nevertheless serves as a baseline against which other responses can be evaluated, and allows the calculation of the minimum number of items or subjects required to achieve the target reliability. From equation 2.113 (p. 80), it can be noted that, where p is the probability of success, the information contributed by a response to a subject or item threshold estimate is $p(1-p)$. Moreover, this will be at a maximum when the subject's ability matches the item threshold difficulty. It follows that a perfectly targeted item contributes 0.25 units of information to the subject and item estimates. To achieve a reliability of r, the total number of units of information required is $1/(1-r)$. Consequently, the number of ideal dichotomous items or subjects required is:

$$N_{ideal} = \frac{1}{0.25(1-r)} = \frac{4}{(1-r)} \tag{4.18}$$

However, perfect targeting can only occur when all items are equally difficult, subjects have the same ability, and the ability matches the difficulty. In practice, subjects will vary in ability and items in difficulty. Indeed, there would be no point

in attempting individual measurement if this were not the case. The effect of this is that most observations will contribute less than the ideal information and thus more subjects and items will be needed. As a practical example, where the distance between ability and threshold difficulty is one logit, the information contribution is about 80% of what it would be if the ability and difficulty were matched.

For the profiles provided by the software, the corresponding number of strata, and the minimum number of cases, dichotomous items and polytomous items required to achieve the safe fit for purpose reliability are summarised in Table 4.3.

Table 4.3: Required number of ideal cases and items for selected purposes

			Items by number of categories		
Purpose	Strata	Cases	Two	Three	Five
No separation required	0	n/a	n/a	n/a	n/a
Minimal separation	2	11	11	9	8
Triage	3	22	22	19	15
Basic grading	5	59	59	49	39
Detailed grading	10	235	235	196	157

The values for cases and dichotomous items follow directly from theory as discussed above. The values for three and five category items are based on the simulations described in Chapter Five.

The discussion above has set out the conceptual framework used to map the information theoretic and statistical concepts and terms to those more familiar to educators. Appropriate formulae have been given to support the various mappings. To make this more accessible in everyday use, a “ready reckoner” has been incorporated into the software. A snapshot of the appropriate screen is shown in Figure 4.12.

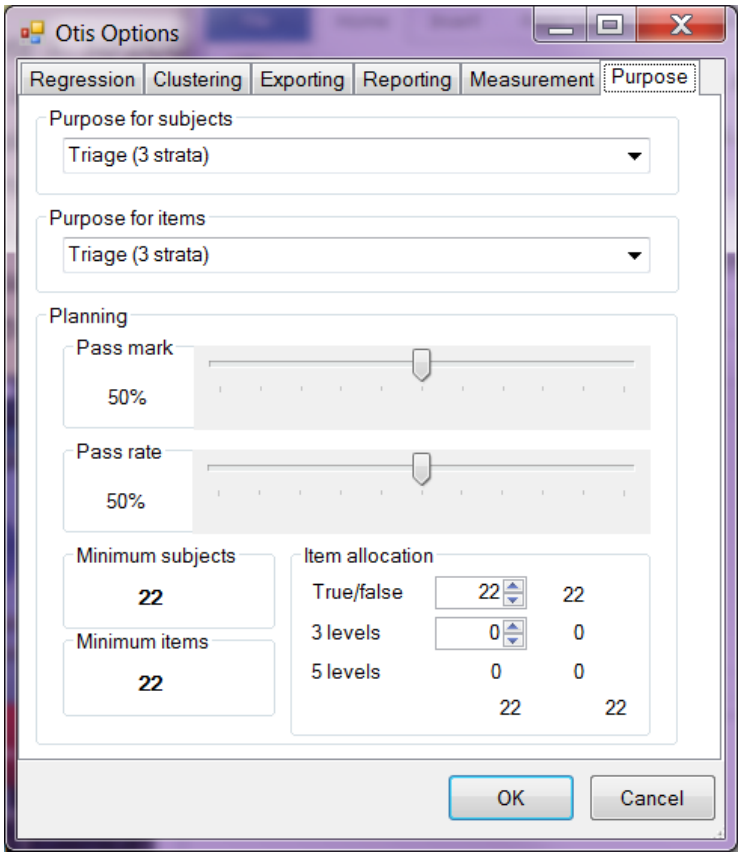


Figure 4.12: The ready reckoner for planning measurement

To evaluate measurement success, all that is needed is to choose appropriate profiles from the two drop down lists at the top of the screen. The rest of the screen is designed to support planning of the measurement exercise. The two sliders allow the pass mark and pass rate to be set; all other figures are recalculated automatically as these are changed. By default, all items are treated as dichotomous. The up and down arrows, at the lower right of the screen, control the allocation of items between dichotomous, three category, and five category items. The left set of numbers in the item allocation section indicate the number of items, the numbers to the right the maximum natural score associated with them, under the assumption that true/false questions are scored {0,1}, three category items {0,1,2}, and five category items {0,1,2,3,4}.

4.3.5. Limitations

It has been assumed throughout the development that the measurement exercise is being carried out within a context of continuous improvement. Moreover, the

position taken is broadly confirmatory. For example, it is assumed that the educator knows how to write effective test questions. However, it is also assumed that the educator may be less familiar with the statistical and information theoretic requirements of reliable measurement. Thus, the broad goal of the expert system is to support the educator through the process of planning, carrying out, and improving measurement. The intent of the software is also to automate, wherever possible, the necessary tasks and to support appropriate interpretation of the outputs.

At the planning stage, the focus is on the number and nature of the items required to measure student ability to the required accuracy, and on the number of students required to classify item difficulty. Inevitably, the guidance given will be approximate rather than exact. The provision of a 10% information loss for the safe level used in planning is arbitrary. Nevertheless, it is considered better to give approximate guidance than no guidance. Moreover, the guidance does help set a realistic expectation for the measurement exercise. Thus, for example, with ten dichotomous items, the maximum possible reliability is 0.6, even with perfectly written items and perfect targeting; this is below the level needed for minimum separation. For a more practical objective, such as triage in the sense used in this project, the minimum reliability is 0.8. This requires at least 20 dichotomous items and 20 students, even if both are ideal in a measurement sense, and the pass rate and pass mark is 50%. The guidance given in this case would be a minimum of 22 for both students and items. In practice, it would be expected that an educator would typically round this up to 25 or 30 to provide an additional safety margin. An important consideration here is the effect of pass rate. If this is 80%, then 35 students and items, rather than 22, would be required to achieve the same accuracy. A similar consideration applies to the pass mark. In sum, the intent is simply to give a realistic starting point for the measurement exercise, not to predict the actual outcome.

When the planned measurement task is carried out, evaluation of success will be carried out automatically. Naturally, it is to be expected that, in a measurement sense, the performance of some items and subjects will be less than ideal. The

essential requirements, from an educator's perspective, are to establish the degree to which the output measurements can be trusted, and to identify what corrective action needs to be taken to remedy any issues found. These requirements are supported by the traffic light metaphor and the narrative diagnostic report (Appendix B). However, there are some limitations to this approach. First, due to the statistical nature of many of the tests, there will inevitably be some false positives among the diagnostics. Second, the critical values used in some of the tests, such as the maximum of two for infit and outfit statistics, although grounded in literature, are essentially arbitrary. The arbitrary nature of such encoded domain knowledge is intrinsic to an expert systems approach. Nevertheless, such encoded domain knowledge is useful to a practitioner, and exposing the underlying statistics allows supplementary expert guidance where appropriate.

Whether the outputs from the measurement exercise are considered acceptable, or not, there is a need to consider what should be done to improve the measurement instrument for future administration. The narrative diagnostic report identifies problematic subjects and items. In practice, future administrations are likely to involve different subjects, so little can be done about subjects. However, consideration should be given to removing, rewording, or replacing problematic items. Moreover, overall reliability indices, such as the separation indices used, cannot capture the complete essence of accuracy, which varies over the range of the instrument. As discussed in section 2.5.6 (p. 84), this can be fine-tuned by examining the targeting of items to subjects and maximising the information provided in critical regions. In practice, this often means increasing the number of items targeted at critical points, such as a pass-fail boundary. However, it is not considered practical to automate this within the expert systems approach. Thus, it is left to the educator to inspect the information density graphs and fine-tune the instrument, if required, by adjusting the mix of items accordingly.

4.3.6. Calibration Drift

Although the focus in this section is on the expert systems nature of the software, rather than on its technical implementation, one issue that emerged during the

development process is described here. The issue relates to a time series and is termed *calibration drift*, an example of which is shown in Figure 4.13.

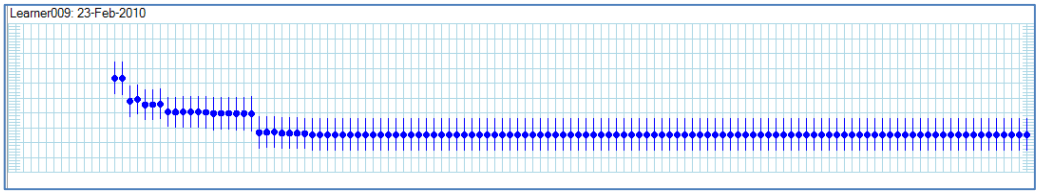


Figure 4.13: An example of calibration drift

This figure shows the change in the estimated ability of a subject, at a point in time, as additional observations are accumulated, under the constraint that the mean item threshold difficulty is set to the centre of the scale. The problem is caused by the lack of a stable common metric across time. The measurement model is only defined up to a linear transformation, and in the absence of an external criterion, is calibrated by reference to sample or item characteristics, as discussed in section 2.5.2 (p. 68). Although re-estimation of the metric is a rational approach, objective three of this thesis sets the goal of communicating the measurements in real time, and a common metric that is stable across time helps achieve this goal.

It is not unreasonable to expect learning activities that are encountered early in a course to be easier than those encountered later. Without additional information, the metric can only be established as students interact with items. There are only two options. If the early items are allowed to set the metric, the metric will be stable, but may be inaccurate because these early items may not be representative of the full course. Alternatively, if the metric is re-evaluated as the course progresses, the metric will be as accurate as possible at each stage, but the assessed difficulty of any item (and thus ability estimates) will tend to drift downwards as more difficult items are added to the mix used for calibration. The core problem is analogous to the problem of establishing a common metric in test equating and the devised solution applies equally to both test equating and time-series.

The solution involves supplying the missing external calibration by allowing some or all the item thresholds to be *anchored* by pre-specifying a difficulty estimate, and by

providing an additional calibration option. There is no need for all thresholds to have pre-specified estimates; Wright and Stone note that "... links of as few as 10 good items will always be more than enough to supervise link validity at better than .3 logits" (1999, p. 87). The calibration option re-evaluates the metric at each stage, subject to the constraint that the mean difficulty of the calibrated anchored thresholds matches the mean difficulty of the pre-specified estimates. For test equating, these pre-specified estimates are the actual results of the anchoring test. For the time-series, three approaches were used. First, estimates were produced on the assumption that difficulties increased linearly throughout the course. Second, the Salsa software, which was used for the empirical evaluation described in Chapter Seven, had information about the expected learning hours of each topic and was thus able to provide improved estimates based on expected learning time. Third, where the course had been used in a previous semester, the software was able to use the previous measurement results as estimates. Estimates for all three methods were produced automatically. Accuracy increased progressively across the three approaches, but all three effectively removed the problem of calibration drift.

4.3.7. Summary

For reasons of space, it has not been possible to describe all the features of the software. For example, the software also allows items to be grouped into topics, thus allowing an educator to work at the more abstract level of a topic. However, the decision has been taken to limit the description in this chapter to those aspects that relate directly to measurement theory and expert systems theory.

This section first described how equivalent item level measures and statistics can be derived from those at the threshold level at which the measurement model works. Issues relating to subject and case level figures were also discussed. These measures and statistics allow an educator to work at a higher level of abstraction, with more detailed figures available if there is a need to drill down into greater depth.

The information theoretic approach provides a unifying conceptual framework. Reliability was introduced as the means of identifying fitness for purpose; and

profiles of purpose, aligned to statistical strata, were introduced as a natural way for an educator to specify purpose and, thus, reliability. The traffic light metaphor was introduced as a simple and clear way of communicating measurement success to educators.

A mapping of information theoretic and statistical concepts and terminology to concepts and terms which are likely to be familiar to an educator was then presented. This mapping supports an educator through the process of planning a measurement exercise by using the purpose profile, the number of subjects, the number of items, the number of categories per item, the pass mark, and the expected pass rate as parameters, rather than the corresponding statistical terms.

The use of graphical representations, qualitative labels and narrative forms are all intended to aid understanding and interpretation by educators.

4.4. DEMONSTRATION AND EVALUATION

This section investigates the efficacy of the developed software to meet the objectives that were set out in section 4.2. The first four subsections evaluate the technical correctness of the implementation by reference to some widely available datasets. These give partial evidence of convergent and discriminant validity (Campbell & Fiske, 1959), but cannot fully validate the solution, nor can they give evidence about the various innovations made in the implementation. These points are addressed in the formal theoretical evaluation presented in Chapter Five. The next subsection relates the features of the implementation to the objectives. Finally, the key features of the implementation are summarised.

4.4.1. Correctness of Implementation

Correctness of the implementation was investigated by analysis of three widely available datasets. The first dataset (Wright & Stone, 1979, p. 31) is known as the *Knox Cube Test* and uses dichotomous items. Outputs from the model were compared to the output of the Winsteps program (Linacre, 2011) and the results are presented in subsection 4.4.2. The second dataset (Wright & Masters, 1982, p. 18) is known as the *Liking for Science* survey and uses polytomous items. Again,

outputs from the model were compared to the output of the Winsteps program (Linacre, 2011) and the results are presented in subsection 4.4.3. The third dataset was sourced from the OECD (2010) and comprised the results for New Zealand students of the Reading, Science, and Maths components of the 2009 *Program for International Student Assessment* (PISA). This dataset provided the opportunity to carry out a large scale test of the software. Outputs from the model were compared to the published OECD results and the findings are presented in subsection 4.4.4.

In general, one would expect similar results to be provided for datasets using dichotomous items (convergent validity) but would expect different results for datasets using polytomous items (divergent validity). This is because the polytomous items in Winsteps are based on the Masters/Andrich polytomous models. These models are equivalent alternative parameterisations of a conceptual model that is based on the assumption that scores are sufficient statistics for measurement. However, this assumption implies that responses to categories within an item are independent of the choice of other categories, which is clearly not true. The model presented in this thesis takes an information theoretic approach and does not make this assumption. It follows that the natural score cannot be a sufficient statistic when polytomous items are used. This alternative conceptualisation is a major difference between the model used in this thesis and the established models and this difference is addressed fully in Chapter Five. The practical consequence of the difference is that, from the perspective of the theoretical treatment proposed in this thesis, the Masters and Andrich models overestimate the available information and consequently understate the standard errors of the estimates. There will also be some differences in the estimates themselves due to the differing information weights that are consequently applied to observations in the estimation algorithms used.

4.4.2. Knox Cube Test

The *Knox Cube Imitation Test* (Knox, 1914) is a non-verbal test of intelligence in which an examinee attempts to replicate a presented sequence of taps (Richardson, 2005). In the version used (Stone & Wright, 1983), each item represented a

sequence of taps and the number of taps in the sequence ranged between two and seven. There were 18 items and 35 subjects in the dataset, and responses were coded as one if the person succeeded in replicating the sequence and zero otherwise. To preserve compatibility, the bias correction given by Wright and Stone (1999, p. 132) was applied because this correction is used in Winsteps; no other corrective options were chosen. Purpose was set at minimum separation (two strata), because Winsteps does not have any equivalent to the specification of purpose used in this thesis. The scale origin was set by constraining the mean item threshold difficulty to be zero, and a measurement unit of a logit was specified.

In the analysis, all hypotheses were supported with the exception of hypothesis 6 (local independence). Local dependence was diagnosed between items 16 and 17, and among several subjects. This was not pursued because Winsteps does not implement any test for local dependence. However, it can be noted that these items also had the lowest outfit statistics (0.11) in Winsteps. Wright and Stone give the following interpretation: “Mean square statistics less than 0.6 imply inter-item dependencies or the presence of secondary variables correlated positively with the intended variable” (1999, p. 116). Thus, the diagnosis of local dependence seems consistent and plausible. Cronbach’s alpha was 0.815, person separation reliability was 0.757 (2.7 strata), and item separation reliability was 0.959 (6.8 strata). These indicate reasonable separation for items, but fall below the level needed for triage on subjects. Fit statistics were largely in accord between the software and Winsteps. Both packages diagnosed one item (item 7) and two cases (Anne, Mike) as misfits. One case (Walter) was identified as a misfit by the software used in this thesis, but was accepted by Winsteps. However, this was marginal (infit of 2.08 compared to 1.94 in Winsteps, with a criterion of 2.0 in both cases) and is a consequence of the different treatment of extreme items. Overall, model fit was acceptable and the scale was classified as productive, excellent quality. A summary of the results is shown in Figure 4.14.

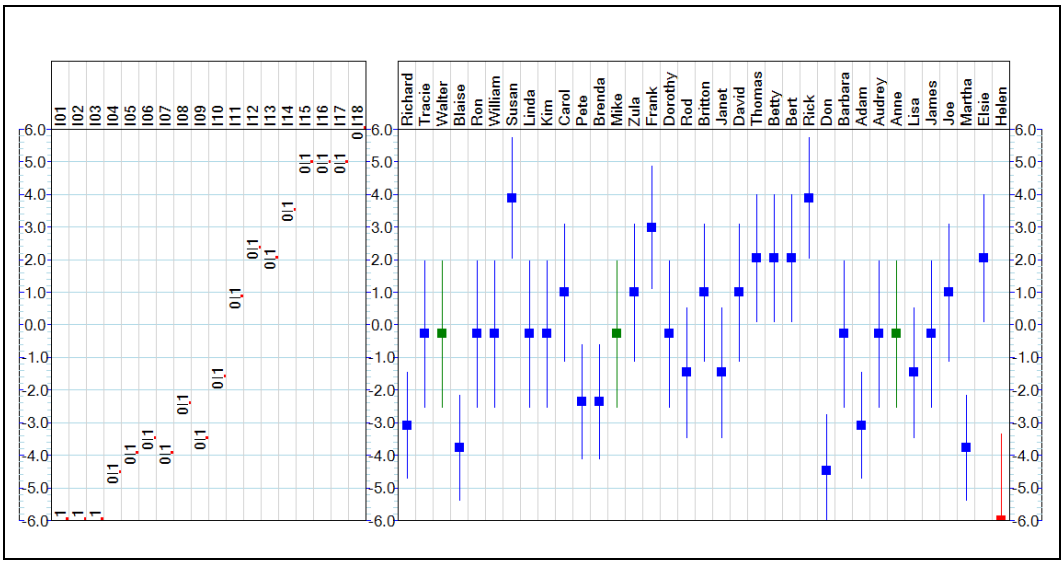
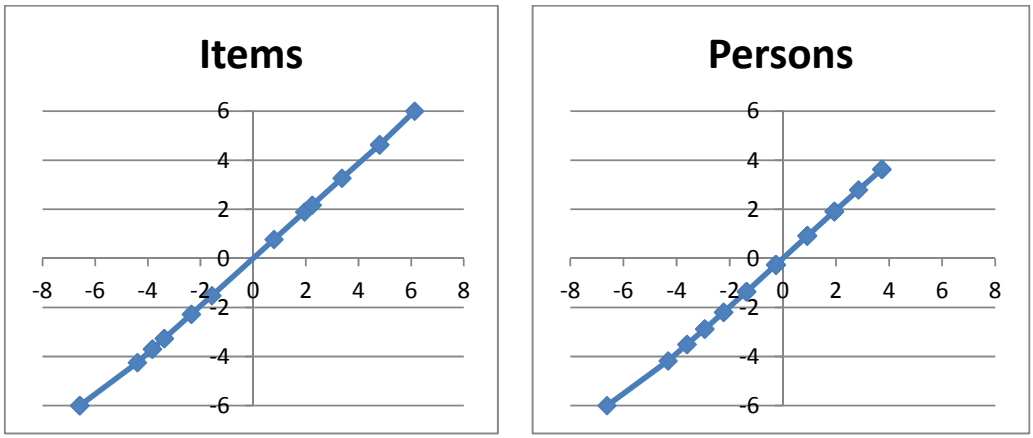


Figure 4.14: Results from the Knox Cube Test

In both software packages, three items (1, 2 and 3) were identified as not measurable because all persons succeeded in replicating the sequence, and one item (18) was identified as not measurable because no person succeeded. The treatment of these extreme items is slightly different between the packages. In Winsteps, these were located at -6.59 and 6.13 logits, respectively. In the present software, they were located at -6.0 and +6.0, respectively. Similarly, person 35 (Helen), who did not succeed on any of the tasks, was marked as immeasurable by both programs. In Winsteps, she was located at -6.62 logits, whereas she was located at -6.0 in the present software. As discussed in chapter two, the decision as to what location to allocate to such immeasurable estimates is essentially arbitrary.

The correlation between the item estimates and those produced by Winsteps was 0.9996 and the correlation between the person estimates and those produced by Winsteps was 0.9995. The relationship between the allocated estimates is shown graphically in Figure 4.15.



(a) (b)

Figure 4.15: Comparison of estimates with Winsteps software

In this figure, the X axis shows output from Winsteps and the Y axis the output from the software used for this project. All estimated standard errors were within ±0.01 logits for both items and cases, with the exception of those for the immeasurable estimates.

In summary, there was a close match between the software and Winsteps on this dataset. The few differences encountered are readily understood in terms of the differing treatment of extreme cases and items, which is a deliberate design decision. Overall, it can be concluded that the implementation of the measurement model for dichotomous items is essentially equivalent to the implementation of the dichotomous Rasch model in Winsteps. This agreement is expected because, although the software packages use different algorithms (unconditional maximum likelihood estimation in the present software and curve fitting in Winsteps), both implement an equivalent measurement model in the dichotomous case. This dataset thus provides evidence of *convergent* validity (Campbell & Fiske, 1959).

4.4.3. Liking for Science

The Cleveland Museum of Natural History created a scale to study children's liking for science by assembling 25 science-related activities. The list of activities covers a range of effort required, suggesting that they would be liked to varying degrees by children who are motivated by a liking for science. The *Liking for Science* dataset (Wright & Masters, 1982) contains the responses of 75 children to 25 rating-scale

items. There were both girls and boys in the sample, and ethnicity was mixed. Their families were of lower to middle socio-economic status and their grade levels were 1st to 6th (6 years - 11 years). Each item had three categories: dislike, neutral, and like. Output from the model was compared to the output of the Winsteps program (Linacre, 2011). To preserve compatibility, the bias correction given by Wright and Stone (1999, p. 132) was applied because this correction is used in Winsteps; no other corrective options were chosen. As with the Knox cube test, purpose was set at minimum separation (2 strata) because Winsteps does not have any equivalent to the specification of purpose. Because the Winsteps reference sample had an unusual metric (a mean person ability of 545.77, s.d. 127.33, and a mean item difficulty of 452.94, s.d. 134.82), the scale origin was set by constraining the subject mean to be 0 and applying a linear transform to the Winsteps results before comparison. A measurement unit of a logit was specified.

It should be noted that the Winsteps software implements the Partial Credit Model (Masters, 1982) for polytomous items and thus some differences are expected between the Winsteps estimates and those of the measurement model used in this thesis. In particular, item difficulties in the PCM are parameterised with an overall difficulty parameter and a number of *step* parameters in a different metric, in contrast to the measurement model used in this thesis, which places threshold difficulties on the common metric. Moreover, item difficulty is inferred from thresholds in the model used in this thesis rather than being an explicit model parameter. There is also a difference in the estimation of standard errors. The PCM assumes that category decisions are independent, whereas the measurement model used in this thesis assumes that they are not. Accordingly, the measurement model used herein should produce higher estimated standard errors than the PCM. This information theoretic approach to estimating standard errors is supported by the results of the simulations described in Chapter Five.

Three measurement hypotheses were rejected in the analysis: hypothesis 3 (unidimensionality), hypothesis 6 (local independence), and hypothesis 12 (fit statistics).

Multidimensionality was diagnosed among items, with three items loading more strongly on a second factor than on the main scale. These were: item 5 (find bottles and cans), item 20 (watch bugs), and item 23 (watch a rat). Some local dependence was also diagnosed among these items. This set of items may be seen as a liking (or dislike) of activities that are not necessarily related to a liking for science. Two other sets of items were diagnosed with local dependence. The first set comprised item 2 (read books on animals) and item 11 (find where animal lives). This set can be seen to share the notion of a liking of *animals.* The second set comprised two items: item 12 (go to museum) and item 13 (grow garden); again, these may be seen as activities that are not necessarily related to a liking for science. Thus, for each of these sets diagnosed with local dependence, inspection suggests that the items may indeed share commonality over and above that explained by a liking for science.

Some multidimensionality was also diagnosed among subjects, with three cases loading more strongly on a second factor than on the main scale. These were case 12 (Daniel Lieberman), case 71 (Dave Stoller), and case 72 (Soloman Jackson). This suggests that these three subjects interpreted the scale differently from other subjects. There was also some local dependence among subjects, with four cliques identified. Local dependence within a clique suggests that there was a shared interpretation, or shared values, within the clique that influenced responses over and above the individual's liking for science.

Three items were diagnosed as exhibiting misfit; these were the same items identified as loading on the second factor. Two of these items were also diagnosed with misfit by Winsteps, but the third (Watch bugs) was marginally accepted by Winsteps. Ten cases were diagnosed as exhibiting misfit. These included the three diagnosed as loading on a second factor and seven other cases. The three loading on a second factor and four of the others were also diagnosed by Winsteps as exhibiting misfit, but three cases (Matthew Shultz, Nancy Hwa and Andy van Damm) were marginally accepted as fitting by Winsteps. No cases or items diagnosed with misfit by Winsteps were accepted by the measurement model. This suggests that the fit criteria used by the measurement model are less tolerant of

misfit than Winsteps. Cronbach's alpha was 0.907, person separation reliability was 0.903 (4.4 strata) and item separation reliability was 0.960 (6.9 strata). These give sufficient separation for basic classification of items, and triage of subjects. Overall, model fit was acceptable and the scale was classified as *productive, excellent quality*. A summary of the results is shown in **Figure 4.16.**

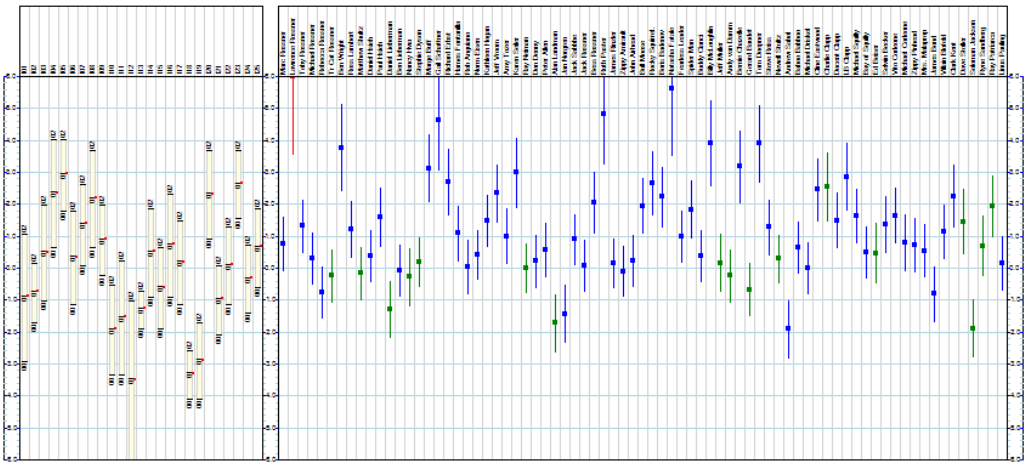


Figure 4.16: Results for the Liking for Science dataset

The correlation between the item estimates and those produced by Winsteps was 0.988 and the correlation between the person estimates and those produced by Winsteps was 0.992. The relationship between the allocated estimates is shown graphically in Figure 4.17. In this figure, the X axis shows output from Winsteps and the Y axis the output from the software used for this project.

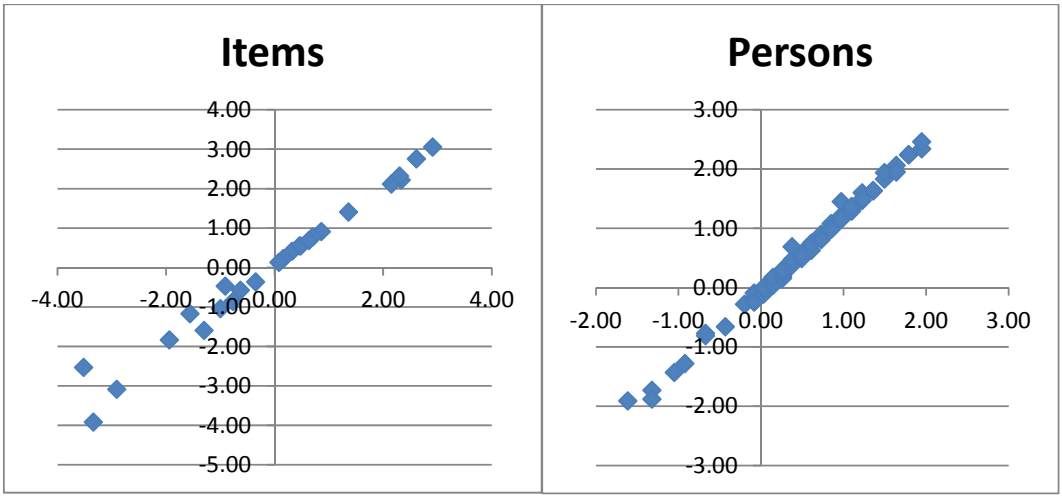


Figure 4.17: Comparison of Liking for Science estimates with Winsteps

It can be seen from this figure that several persons, and particularly some items, have noticeably different estimates. Five items had a difference of more than 0.2 logits and the greatest difference was observed with item 12 (Go to museum); this was located at -2.53 by Winsteps and -3.52 by the software used in this thesis. The response frequencies for this item are shown in Figure 4.18.

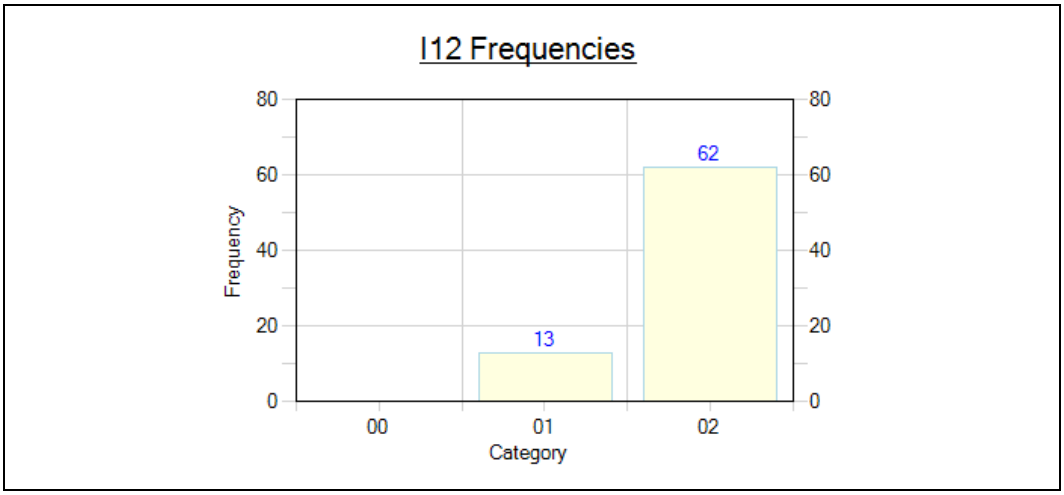


Figure 4.18: Response frequencies for item 12 in the Liking for Science dataset.

It can be seen that no subject chose option 00 and thus, the threshold separating the first two categories is immeasurable. Unobserved categories are problematic for the Partial Credit Model, because the metric used by the model is based on the relative frequencies in successive categories. However, for the software used in this thesis, the difficulty of an item is imputed as the ability at which the expected score is 50% of the maximum possible. The output for this item is shown graphically in Figure 4.19.

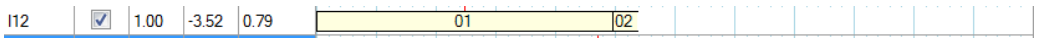


Figure 4.19: Compact representation of item 12.

It can be seen that the 00 category is effectively discarded and the imputed difficulty falls at the centre of the range of the 01 category. This is a simple and natural interpretation of item difficulty for educators, but differs from the interpretation under the Partial Credit Model. The essential point, however, is that the model used in this thesis operates at the item threshold level and imputes item ability, whereas the PCM treats item difficulty as a model parameter. Micro-analysis

of the other four items suggests that the differences are as expected under the two models.

Four persons had a difference of more than 0.3 logits. Three of these had outfit statistics classified as degrading by both packages: case 72 (.56 logits difference), case 29 (.41) and case 5 (.36). A micro-analysis is presented for the fourth, Jan Nordgren, case 30 (.38), which had acceptable fit statistics in both packages. This was estimated at -1.05 in Winsteps and -1.46 in this software, with standard errors of 0.36 and 0.46 respectively. Under the response model proposed in this thesis, only thresholds that are adjacent to a chosen category contribute information to an estimate. This subject chose the 00 category on 15 questions, the 01 category on 4 questions and the 02 category on 6 questions. Thus, the average number of thresholds per question that contribute information is 1.16 rather than the 2 that would be expected under the assumption of independence. It follows that the standard error should be roughly 31% larger ($\sqrt{2/1.16}$) than the Winsteps estimate of 0.36: i.e. 0.47. This is consistent with the estimate of 0.46.

Moreover, the case estimates are also affected by the response pattern. As discussed in section 2.5.7 (p. 86), only those thresholds that are deemed to contribute useful information are summed in determining the maximum likelihood estimates of ability. Thus, the effective weighting of the threshold difficulties used in the estimation procedure is dependent on the response pattern and, consequently, differences in estimates are expected.

On average, standard errors were 24% larger. Overall, the fit of the data to both models was acceptable: averages of the fit statistics are shown in Table 4.4.

Table 4.4: Comparison of fit statistics

	Items		Persons	
	Infit	Outfit	Infit	Outfit
Winsteps	1.03	1.10	0.99	1.08
This model	0.97	1.06	0.97	1.12

In summary, broad agreement was found between the two packages in the analysis of the Liking for Science dataset. Some differences were both expected and found. A micro-analysis suggests that the differences can be attributed to the conceptual differences between the models. This gives some evidence of both *convergent* and *discriminant* validity (Campbell & Fiske, 1959). However, evidence of convergent and discriminant validity is not sufficient to establish correctness. Although they give some assurance that the implementation is similar in the areas expected, and also different where expected, they cannot establish the correctness of the underlying model. This is treated formally in Chapter Five.

4.4.4. PISA 2009 Data

In order to provide a large scale test of the software, three datasets comprising the results for New Zealand students of the Reading, Science and Maths components of the 2009 Program for International Student Assessment (PISA) were analysed. The PISA programme is coordinated by the Organisation for Economic Cooperation and Development (OECD). It is carried out on a three yearly cycle and 65 countries and economies participated in 2009. All datasets were sourced from the OECD (2010).

The intention of the PISA assessment was not to provide individual assessment of students, but rather to provide comparisons at the country level. Accordingly, the purpose was defined for the software as minimal separation for subjects and detailed classification for items. The overall characteristics of the dataset are shown in Table 4.5.

Table 4.5: Characteristics of the PISA datasets.

Subject	N	C	Multi-dimensionality	PSR	ISR	COR
Reading	4,643	1,384	17% (17 of 101)	0.847 (3.5)	0.997 (25.2)	80%
Science	3,217	1,370	28% (15 of 53)	0.818 (3.2)	0.996 (21.9)	77%
Maths	3,215	1,421	6% (2 of 35)	0.754 (2.7)	0.997 (26.8)	79%

In this table, column N shows the number of students who participated in each subject area. Column C shows the minimum connectivity of the common metric; the study used a booklet approach in which not all students were presented with all questions, but there was sufficient overlap to achieve a common metric. The multidimensionality column shows the proportion of items loading more strongly on a factor other than the main scale. The PSR and ISR columns show the person and item separation reliabilities, with the corresponding equivalent number of statistical strata in parentheses. The COR column shows Guttman's Coefficient of Reproducibility.

The hypothesis of unidimensionality (H3) failed in all three datasets. In *Reading*, ten factors were identified, and 17 of the 101 items loaded more strongly on the second or subsequent factors than on the main scale. In *Science*, seven factors were identified, and 15 of the 53 items loaded more strongly on the second or subsequent factors than on the main scale. In *Maths*, two factors were identified, and 2 of the 35 items loaded more strongly on the second factor than the main scale. Rejection of this hypothesis is not a major concern because the PISA instruments are intentionally multidimensional and comprise a number of distinct concepts and formats, each measured by a cluster of items: the reading test used seven main clusters of questions, and the maths and science each used three main clusters (OECD, 2009, p. 29).

All three datasets failed the hypothesis of no differential item functioning (H4) with respect to booklets. This booklet effect is well known (OECD, 2009, p. 203) and a regression model was used to adjust for this, after the event, in the official PISA results. Nevertheless, stronger findings would result if this were corrected in the model or eliminated at an earlier stage.

All three datasets also failed the hypothesis of local independence (H6). This is to be expected because of the use of *testlets* in all three instruments. Each testlet presented a scenario and then asked several questions relating to the scenario. Sharing a scenario across multiple questions allows more complex tasks and

knowledge to be explored in assessment, but creates a local dependency among the set of questions.

All other measurement hypotheses were supported. The measurement software has options to manage failure of the three hypotheses that failed on the OECD datasets. The impact of these corrective options was evaluated and the findings are summarised in Table 4.6.

Table 4.6: Reliability of the PISA datasets with information correction options.

	Maths		Reading		Science	
	PSR	ISR	PSR	ISR	PSR	ISR
Baseline	0.754	0.997	0.847	0.997	0.818	0.996
a) Manage multidimensionality	0.770	0.998	0.847	0.997	0.835	0.996
b) Manage DIF	0.761	0.997	0.780	0.997	0.827	0.996
c) Manage local dependence	0.767	0.997	0.847	0.997	0.818	0.996
Manage a, b and c	0.772	0.998	0.834	0.997	0.855	0.997
Manage a and c	0.770	0.998	0.847	0.997	0.836	0.996

The approach taken by each of the management options is based on consideration of the proportion of information that can be trusted to give useful information about the quantity to be estimated, and then eliminating any untrusted information from the calculations. This can be expected to have two opposite effects. First, removing information should, in general, reduce accuracy and thus result in lower reliability indices. However, to the extent that the untrusted information degrades measurement, an increase in accuracy, and thus reliability, is to be expected.

The option to manage multidimensionality raised reliability for the Maths and Science datasets, which suggests that the presence of multidimensionality was degrading for these scales. There was no significant effect on the Reading dataset. Whether or not to apply this management option in any specific situation is a complex judgement. The option excludes items that load more heavily on a second or subsequent factor from case calibration. This may improve measurement of the first factor, but may threaten content validity because the second and subsequent

factors do not contribute to the measurement. Using separate scales for any additional factors preserves content validity, but there may not be sufficient items to measure these accurately. However, whether or not the option is used, it is a useful diagnostic tool to investigate the impact of any multidimensionality.

In both the Maths and Science datasets, the option to manage differential item functioning (DIF) raised reliability, which suggests that the *booklet effect* was degrading for measurement. However, reliability was reduced for the Reading dataset, which suggests that removing untrusted information removed too much information to compensate for the increased quality of the information retained. The management option works by excluding items diagnosed with DIF from case calibration. As with the previous option, this may improve measurement of the scale defined by the items retained, but may threaten content validity because the items diagnosed with DIF do not contribute to the measurement. In general, this option would only be recommended if relatively few items are diagnosed with DIF and content validity is not significantly threatened by the removal of untrusted items. Nevertheless, it is a useful diagnostic tool to investigate the impact of any differential item functioning.

The option to manage local dependence raised reliability for the Maths dataset, but had no significant effect on the Reading or Science datasets. Unlike the two previous options, the use of this option does not pose a major threat to content validity. This is because the information contribution of dependent items is reduced to reflect the portion of unique information contributed by items, rather than eliminating the entire contribution. In general, this would be expected to reduce reported accuracy because it should remove what would otherwise be overstated accuracy. The increased observed reliability for the Maths dataset thus suggests that the items exhibiting local dependence were also degrading to some extent.

The use of these options in combination was also investigated. The use of all three options raised reliability for the Maths and Science datasets, but reduced reliability for the Reading dataset. This is largely because of the exclusion of items exhibiting DIF, which was attributable to the booklet effect, and diagnosed for most of the

items (61 out of 101). The use of the other two options in combination raised reliability for the Maths and Science datasets, and had no effect on the reliability of the Reading dataset.

These results suggest that the option to manage local dependency should be considered whenever local dependency is diagnosed, but the other two options should be used with caution. Where relatively few items are affected and content validity is not threatened, their use is probably warranted. Otherwise, it is probably prudent to try the options in order to evaluate the impact of the diagnosed issues, but not to use the corrected measurements, or to use them with caution.

4.4.5. Evaluation

The analysis of the reference datasets presented in the foregoing sections of this chapter give some assurance of convergent and discriminant validity, but they cannot address the correctness of the underlying models and conceptual framework. This is addressed formally in Chapter Five. The analysis has also shown that the diagnostics and hypothesis framework incorporated in the software can be used to evaluate and interpret the datasets. In particular, it can be noted that multidimensionality, the booklet effect, and local dependence in the PISA dataset were all diagnosed *automatically* by the software within the conceptual hypothesis framework used.

Three objectives were set in this thesis:

- Is it possible to develop objective measures of challenge and self-efficacy from self-report data? (Objective 1)
- How can these measurements be communicated clearly to educators and learners? (Objective 2)
- Develop a practical computer software implementation that will communicate the measurements in real time (Objective 3)

The first of these objectives is evaluated in Chapter Seven, which addresses the empirical use of the model to measure self-efficacy and challenge. Accordingly, this

is not pursued further in this section. The second and third objectives are addressed in this section. The objective that measurements be communicated clearly to educators and learners will be discussed first. To achieve this, a key goal has been to supplement tables of figures with alternative representations. Figure 4.20 shows a variant of a *Wright Map* of the results of the PISA Maths dataset. This is named in recognition of Ben Wright, who pioneered the idea of presenting subjects and items on a common metric. Other examples of a Wright Map have been shown in Figure 4.14 and **Figure 4.16**. However, because of the large number of subjects in the PISA dataset, the right hand side of Figure 4.20 shows the distribution of the subjects, rather than their individual measurements; the software uses this format automatically whenever there are too many cases to show individually. The left hand side shows the categories associated with each item, separated by their thresholds at the modelled scale locations. The category label {0, 8} signifies that codes 0 and 8 are combined in this category; code 8 was used to identify unanswered questions at the end of the test and these have been treated as equivalent to 0 for this analysis. From this diagram, it is easy to see that question 20 was the most difficult for the subjects and that question 31 was the easiest.

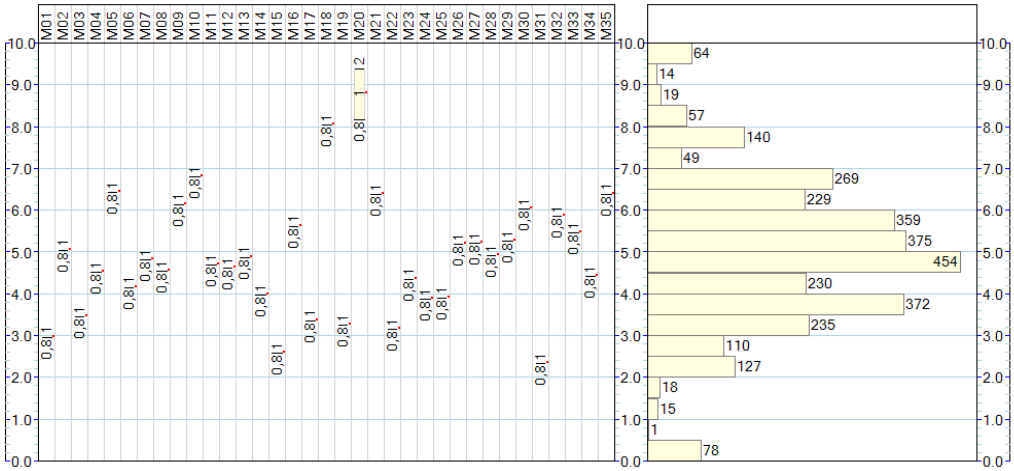


Figure 4.20: Wright Map of the PISA Maths results.

An alternate visualisation of the same results is shown in Figure 4.21.

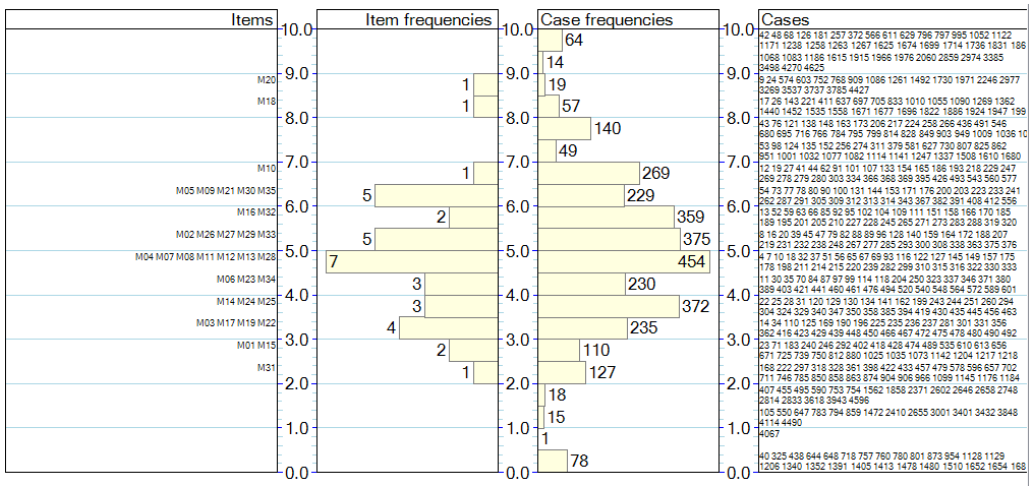


Figure 4.21: Interpretation of the PISA Maths results.

This visualisation shows the distributions of item difficulties and case abilities on the same metric, enabling easy comparison for targeting purposes. On the left, each item is also shown at its modelled difficulty location, and on the right, each case is shown at its modelled ability location. This representation makes it even clearer that question 20 is the most difficult, and question 31 is the easiest. The individual cases shown on the right are difficult to interpret for results on the scale of the PISA dataset, but are useful in a typical educational context.

These visualisations encode a substantial amount of information in a compact format. Other examples of visualisations have been given throughout this thesis. Mention has also been made of the use of verbal qualitative labels and narrative forms (see Appendix B).

The second element used to promote clear communication of findings is the traffic light metaphor, which has the simple message: do not proceed, proceed with caution, or proceed safely.

The remaining objective was to develop a practical computer software implementation that will communicate the measurements in real time. The analysis of the *Knox Cube Test* and the *Liking for Science* datasets has demonstrated that meaningful results can be achieved with sample and test sizes that are typical of an educational context. This suggests that the model is practical, and can be used in a typical educational setting. The time taken by the software for analysis is also

reasonable. For example, analysis of the PISA Maths dataset, including all hypothesis tests, took seven seconds on a modern personal computer (Intel Core I7, quad processor). Practical limits of sample size and test size are explored further in chapter five. The ability to communicate measurements in real time was addressed in section 4.3.6, which discussed the approach taken to achieve a measurement metric that is stable over time.

Three additional requirements were specified:

- The software must be *usable* by educational practitioners
- Output must be *interpretable* by educators and learners.
- The software should provide *useful* measurement in realistic educational settings.

The requirement that the software be usable by educational practitioners has led to an *expert systems* approach. The running of all hypothesis tests, measurement, and the evaluation of measurement outputs are all fully automatic and managed by the software. The educator needs only to specify purpose, which may be done by selecting from a list of predefined profiles.

The second requirement was that output be interpretable by educators and learners. As discussed throughout this chapter, interpretation has been at the heart of the software development. Multiple representations of outputs, in numeric, graphical, verbal, and narrative forms, have been used to enhance understanding. The traffic light metaphor has provided a simple and clear way of communicating measurement success. The use of a unifying conceptual framework, together with the mapping of technical concepts to educational terms and concepts, also helps to promote understanding.

The third requirement was that the software provides useful measurement in realistic educational settings. The usefulness of measurement cannot be assessed without a definition of purpose. However, consensus of the focus group suggests that *triage*, as defined in this thesis, serves as a measure of purpose that is both easily understood, and generally useful. For items, this corresponds to a

classification of easy, medium or hard. For subjects, this corresponds to three groups: lower, middle, and upper. This purpose implies a reliability index of 0.8 or more. Theoretical analysis suggests that achieving this reliability requires sample sizes of 20 to 30 students, and 20 to 30 dichotomous items or 15 to 20 five-category items. All of these are achievable in a typical educational setting.

4.4.6. Summary

This section has investigated the efficacy of the developed software to meet the objectives that were set out in section 4.2. Three widely available reference datasets were used to evaluate the correctness of the implementation. These provided basic evidence of convergent and discriminant validity. The analysis also demonstrated the efficacy of the hypothesis framework in identifying issues with the reference datasets, and the value of the diagnostics in interpreting the causes of the issues that were identified. A key focus has been on those features which ensure that the software is usable, and that the outputs are interpretable and useful, in a typical educational setting.

4.5. CHAPTER SUMMARY

This chapter has presented selected features of the proof of concept software that was implemented in connection with this thesis. The focus has been on the non-routine aspects of the development. These are concerned with the development of heuristics and encoded domain knowledge that are part of an expert system, the development of a unifying conceptual framework, the selection of appropriate metaphors, and the choice of a number of output representations. A key aspect of the development has been attention to the human factors that relate to the need for the software to be usable in a typical educational context, and to the need for a representational model that supports interpretation by educators. The feedback from the group of educators who trialled the software was invaluable in this regard.

It has been demonstrated that use of the implemented model is both practical and useful in a typical educational setting. Reference datasets were used to provide basic evidence of convergent and discriminant validity. These give some assurance

that the implementation of the model is performing as expected. However, they cannot establish the correctness of the models, which is addressed formally in the next chapter.

Chapter 5.
THEORETICAL EVALUATION

The previous chapters described the measurement model and its implementation in software. The software implementation was evaluated using some well-known reference datasets. However, a more formal evaluation is required to develop a fuller understanding of the characteristics of the model. To achieve this, simulations were undertaken to investigate a number of properties of the implemented model. The use of simulation allows the properties to be investigated systematically under controlled conditions.

Three goals were pursued in this investigation. The first goal was to evaluate how well the implemented model accords with theoretical expectations. The second was to identify characteristics that could be used to inform the expert system heuristics, guide planning, and support interpretation of the various outputs. The third goal was to investigate the robustness of the model under deviations from its core assumptions.

The key characteristics investigated were: the use of polytomous items, the number of cases and items, the effect of different item and case distributions, the effect of targeting items to cases, and tolerance of introduced errors.

5.1. APPROACH TO SIMULATIONS

Custom software was created to automate and manage the simulation and evaluation process. The simulation software was controlled by a test specification that defined the various parameters and permutations to be explored. Each test specification was run 1000 times for each specified permutation. Each of these runs involved generating case and item/threshold instances, creating a dataset with appropriate responses, running the implemented measurement software on the dataset, and then comparing the results to the generated values and logging the outcomes. This process is summarised in Figure 5.1.

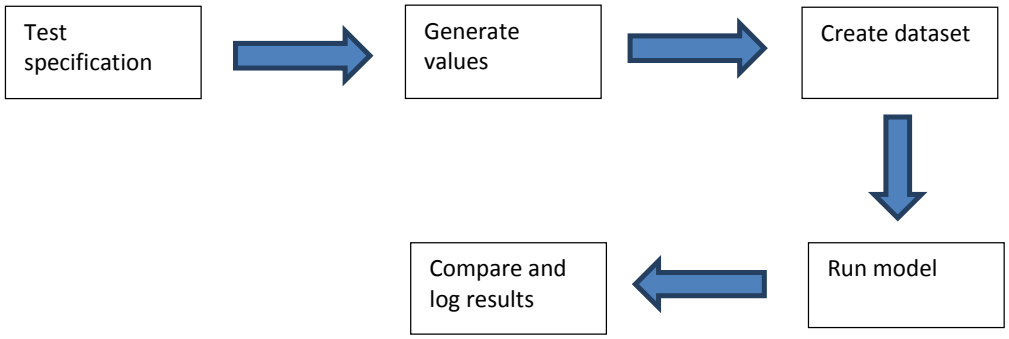


Figure 5.1: Overall schematic of simulation process

The output log of the results was then analysed and the findings of these analyses comprise the main content of this chapter.

5.1.1. Test Specifications

Case and threshold distributions were specified in terms of shape, mean, and standard deviation. Values were first allocated randomly in the unit interval {0...1}, based on the shape, and then a linear transformation was applied to each distribution to achieve the required mean and standard deviation. Three shapes were used to accommodate differing degrees of central tendency: uniform, central, and extreme. With the *uniform* shape, each value in the range had an equal probability of being chosen. With the *central* shape, ten bins were used, with target frequencies allocated in accordance with Pascal's triangle {1, 9, 36, 84, 126, 126, 84, 36, 9, 1}, and linear interpolation was used within a bin. The *extreme* shape followed a similar process, but used the inverse of this distribution {126, 84, 36, 9, 1, 1, 9, 36, 84, and 126}. These shapes are shown visually in Figure 5.2.

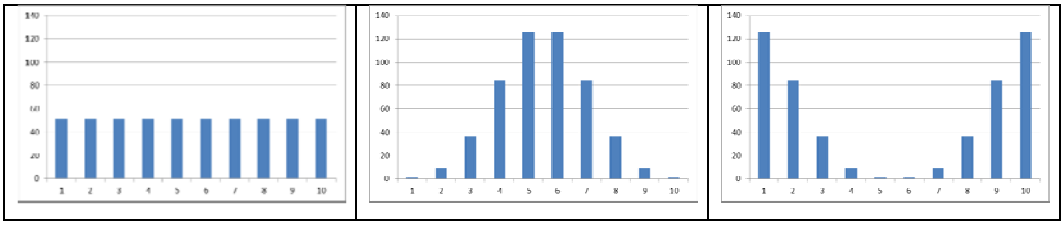


Figure 5.2: The shapes of the distributions used in the simulations

In principle, the shape of the distributions should have little effect on the estimates produced by the model, because no assumptions were made about distribution in

the derivation of the model. However, the shape of the distribution can be expected to affect the proximity of item thresholds to case abilities. This, in turn, will affect the information contributed by each observation. Consequently, it can be expected to have a moderate direct effect on estimated standard errors, and a minor indirect effect on ability and difficulty estimates due to the information weighting implicit in the estimation procedure.

The choice of means was determined as follows. It can be noted that the choice of measurement origin is arbitrary, and that the imputed measurements depend solely on the difference between case abilities and threshold difficulty estimates. Consequently, without loss of generality, the mean of the item thresholds was specified as zero, the measurement model was configured to use criterion referenced centring with a threshold mean of zero, and the effects of targeting items to cases were investigated by varying the mean of the case distribution. Moreover, because of the symmetry of the model, there is no need to investigate both positive and negative differences between case and threshold means. Accordingly, only those situations where the case mean is not less than the threshold mean were investigated. In these situations, higher differences between the case and threshold means correspond to higher pass rates. To investigate realistic scenarios, three values were chosen: 0, 1.0 and 2.0 logits. These correspond, roughly, to pass rates of 50%, 75%, and 90%, for tests with a 50% pass mark.

The targeting of item thresholds to case abilities is also affected by the spread of the distributions. The choice of standard deviations was based on the observation by Wright and Stone (1999, p. 146), that a range of ± 2 logits is appropriate for a test. Treating this effective range as roughly equivalent to ± 2 standard deviations, a standard deviation of one logit can be seen as a reasonable value. However, to allow exploration of the sensitivity of the model to different standard deviations, three values were chosen: 0.5, 1.0 and 2.0. These were labelled narrow, medium and wide.

The specification allowed items to have from 1 to 10 thresholds, corresponding to 2 to 11 categories. It also allowed for the addition of a certain amount of random noise. The noise level was specified as a proportion of the case distribution standard deviation. This provision for noise allowed investigation of the sensitivity of the model to measurement error. Finally, the number of items and the number of cases were specified. The elements of the test specification, discussed above, are summarised in Table 5.1.

Table 5.1: Summary of parameters used in simulations

Element	Characteristic	Options
Item	Number of thresholds	1 to 10
Item threshold	Shape	Uniform, Central, Extreme
	Mean	0.0
	Range (s.d.)	Narrow (0.5), Medium (1.0), Wide (2.0)
Case	Shape	Uniform, Central, Extreme
	Mean	0.0, 1.0, 2.0
	Range (s.d.)	Narrow (0.5), Medium (1.0), Wide (2.0)
Allocation	Noise	0%, 25%, 50%, 100% of case s.d.
	Number of items	10, 20, 50, 100, 200
	Number of cases	10, 20, 50, 100, 200

Unless otherwise specified, all tests were measured on a logit scale and no corrective options were chosen. A set of test specifications was associated with each characteristic to be explored, with systematic variation of selected elements of the test specification. This allowed investigation of the effect of these elements. Some of the explorations also involved a permutation of several of the elements. In most cases, the following values were used for elements that were not specifically varied: dichotomous items, central shape for thresholds and cases, medium range for thresholds and cases, a case mean of 1, no allocation noise, 50 items and 50 cases. These were considered realistic values for a typical educational setting.

5.1.2. Dataset Generation

Case abilities and threshold difficulties were allocated by drawing samples from the specified distributions. Where an item had multiple thresholds, each threshold was drawn randomly from the specified distribution and the thresholds for an item were then sorted into difficulty order to preserve the ordinal response pattern.

The dataset was then generated by assigning the responses of cases to items in accordance with the model. Each generated response, by a subject to an item, was based on the allocated ability of the subject, the amount of noise specified, and the allocated locations of the thresholds associated with the item, as follows. First, a noise sample was drawn from a normal distribution with a mean of zero and the specified standard deviation. Second, the allocated subject ability was adjusted by this noise sample to produce a perturbed ability estimate. Third, the probabilities of a response in each category were calculated for the perturbed case ability and the allocated threshold difficulties, as defined in equation 2.53 (p. 49). A random sample was then drawn from this empirical distribution to determine the actual value allocated.

5.1.3. Comparisons

After running the measurement model, the automation software determined and logged the outcomes. The automation software had access to the actual values used to derive the generated responses that formed the inputs to the model. This allowed an external assessment of the performance of the model. Accuracy of the estimates was derived from the Pearson correlation (r) between the allocated values and the estimates. From this, an estimate of the standard error is given by:

$$\hat{\sigma} = \sqrt{1 - r^2} \tag{5.1}$$

This was termed the *actual* standard error. This was compared to the *model* standard error, which was calculated as the root mean square of the standard error estimates produced by the model, and then standardised by dividing by the standard deviation of the appropriate distribution. This standardisation allows both actual and model standard errors to be expressed in the same dimensionless units.

However, a simple mean is not appropriate for correlational measures (Fisher, 1915). Accordingly, averages of correlational measures were achieved by using Fisher's z transformation to transform them into a suitable metric, averaging in the transformed metric and then translating back for interpretation. Fisher's z transformation (p. 521) is $tanh^{-1}(r)$:

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \tag{5.2}$$

The inverse transformation is given by:

$$r = \frac{e^{2z}-1}{e^{2z}+1} \tag{5.3}$$

Similarly, averages for counted fractions were achieved by transforming them into the logistic metric, averaging in the transformed metric and then translating back for interpretation. The logistic transform is:

$$z = \log\left(\frac{f}{1-f}\right) \tag{5.4}$$

The inverse transformation is given by:

$$f = \frac{e^{z}}{e^{z}+1} \tag{5.5}$$

The treatment of the various statistics, discussed above, is summarised in Table 5.2.

Table 5.2: Averaging of the measures and statistics used in comparisons

Statistic	Method
Actual error	Calculated from the proportion of variance not explained by the model.
Modelled error	Root mean square of standard errors reported by the software.
Cronbach's alpha	Averaged using Fisher's z transformation
Person separation	Averaged using Fisher's z transformation
Item separation	Averaged using Fisher's z transformation
Reproducibility	Averaged using logistic transformation
Infit	Simple arithmetic mean
Outfit	Simple arithmetic mean

5.2. THEORETICAL EXPECTATIONS

To provide a context for the evaluation of the simulations, the main results expected from a theoretical perspective are set out in this section. Firstly, it can be noted that one would expect item (threshold) estimates to be affected mainly by case parameters, and case estimates to be affected mainly by item parameters. On the other hand, one would expect quality control and fit statistics to depend jointly on the characteristics of both cases and items.

For the number of cases, items and thresholds, it is clear that each item provides an observed response that contributes to the ability estimate of a case. Similarly, each case provides an observed response that contributes to the estimates of item and threshold values. Accordingly, one would expect that a greater number of cases would result in improved accuracy of item (threshold) estimates and that an increased number of items and thresholds would result in improved case ability estimates. More specifically, from an information theoretic perspective, increasing the number of cases from n_1 to n_2 should improve the accuracy of item threshold estimates by a factor of $\sqrt{n_1/n_2}$, but should have negligible effect on the accuracy of case estimates. Likewise, increasing the number of items from n_1 to n_2 should improve the accuracy of case ability estimates by a factor of $\sqrt{n_1/n_2}$ but should have negligible effect on the accuracy of item estimates.

If the natural score were a sufficient statistic, as assumed in the derivation of the Masters and Andrich models, polytomous items should have a similar improvement in accuracy, proportional to $\sqrt{t_1/t_2}$, where t_1 and t_2 are the respective threshold counts. However, the information theoretic approach used in this thesis suggests that at most two thresholds will contribute information to a case estimate and, thus, a smaller improvement is predicted. This is a major difference between the model proposed in this thesis and the established Masters and Andrich models, and this difference is discussed fully in section 5.7.

One would expect modelled standard errors to reflect the accuracy of actual estimates, and thus expect reliability measures, such as Cronbach's alpha to

improve with greater accuracy of estimates. Conversely, one would expect reproducibility, as measured by Guttman's coefficient of reproducibility (COR), to fall a little with improved accuracy. This is to be expected because increases of either the number of items or the number of thresholds per item are likely to cause an increase in the density of thresholds in the region of each case estimate. Likewise, an increase in the number of cases is likely to increase the number of cases in the region of each threshold. These increase the information density (and thus accuracy) since information is at a maximum when the maximum uncertainty is discharged by the observation. However, since the calculation of the Guttman COR is deterministic, this is precisely the situation in which the coefficient is most likely to have discordant observations. Finally, one would expect the effect of the number of cases, items and thresholds on the infit and outfit statistics to be neutral. The infit and outfit statistics are based on an analysis of residuals which, in turn, depend jointly on both case abilities and item threshold difficulties. One would therefore expect these statistics to depend jointly on case and item characteristics. Furthermore, both item and case versions of these statistics represent different summations, across cases and items respectively, of the same underlying measure. Accordingly, one would expect the infit statistics of both case and item thresholds to be comparable and similarly, the outfit statistics of both should be comparable.

Since the information provided by an observation (and thus accuracy) is at a maximum when case ability and item threshold difficulties are closely matched, one would expect the range, distribution and targeting of items to have an effect on the accuracy of estimates. Broadly speaking, increasing the range and extremity of the distributions is likely to reduce the information density in the region of an estimate, and thus reduce accuracy. Similarly, close targeting of item and case estimates is likely to increase information density, and thus accuracy. In general, one would expect a greater difference between the means of the item threshold and case ability to increase the average distance between individual case ability estimates and item threshold difficulty estimates. This should therefore lead to a reduction in accuracy and Cronbach's alpha, an improvement in reproducibility, and a relatively

neutral effect on fit statistics. Finally, one would expect a deterioration of accuracy of estimates, reproducibility and fit statistics as noise is added or observations deviate from the model in any other way.

A standardised approach has been taken for the interpretation and reporting of findings. All findings are first tested for statistical significance. This is tested mainly by ANOVA, but where the assumptions of the ANOVA method do not hold, a Kruskal-Wallis non-parametric method is used as a fall-back. Findings are reported as *not significant* (ns) if they are not statistically significant at the $p < .05$ level. However, with the relatively large number of replications used, even relatively small effect sizes are likely to be detected as statistically significant. To aid consistent interpretation, a number of qualitative terms are used to characterise the effect size of statistically significant results: the term *negligible* is used for effect sizes below 1%, *minor* for effect sizes between 1% and 5%, and *major* for effect sizes of 5% or more. These terms and values are, of course, arbitrary, but, nevertheless, reasonable in the context of educational measurement. The theoretical expectations are summarised in Table 5.3.

Table 5.3: Predicted findings

Element	Accuracy		Reliability		Infit		Outfit		Other	
	Item	Case	Item	Case	Item	Case	Items	Case	Alpha	COR
Polytomy		⬆⬆		⬆⬆					⬆⬆	⬇
No items		⬆⬆		⬆⬆					⬆⬆	⬇
No cases	⬆⬆		⬆⬆							⬇
Item shape	↧↥	↧↥	↧↥	↧↥						
Item range	↧↥	⬆	↧↥	⬆						
Case shape	↧↥	↧↥	↧↥	↧↥						
Case range	⬆	↧↥	⬆	↧↥						
Case mean	⬇⬇	⬇⬇	⬇⬇	⬇⬇						⬆
Noise	⬇	⬇	⬇	⬇	⬆	⬆	⬆	⬆	⬇	⬇
Notes	The symbol ⬆⬆ represents a major positive effect, ⬇⬇ a major negative effect, ⬆ a minor positive effect, ⬇ a minor negative effect, and ↧↥ a complex interaction. All other effects are predicted to be not significant or negligible.									

5.3. POLYTOMOUS ITEMS

The first area investigated was the effect of the number of item categories. This is a key point of difference between the proposed model and existing measurement models. The following questions were investigated:

- What is the effect on estimates and estimated errors?
- What is the effect on reliability and fit statistics?

Two series of tests were used in the experiment. In the first series, the dataset comprised 50 cases and 50 items. In the second series, there were 50 cases and 100 items. The same specification was used for both series: item thresholds were allocated from the central medium range distribution; case abilities were allocated from the central medium range distribution with a mean of 0; no noise was added; threshold counts ranged from 1 to 10, corresponding to 2 to 11 categories. The experiment used a fully crossed factorial design: 2 [series] x 10 [number of thresholds]. Each test was repeated 1,000 times, giving a total of 20,000 tests.

To evaluate the effect on estimates, a non-parametric Kruskal-Wallis test, comparing ranks across categories defined by the number of thresholds, was used. This test was chosen because the variance was not homogeneous across groups, which is one of the assumptions of ANOVA. The results are shown in Table 5.4.

Table 5.4: Significance of the effects of the number of thresholds on estimates

		Items	H(9)	ϕ_C	ϕ_C^2	Significance	Comment
Threshold	Actual	50	61.31	0.0261	0.1%	<0.0001	Negligible
		100	54.05	0.0245	0.1%	<0.0001	Negligible
	Model	50	185.82	0.0454	0.2%	<0.0001	Negligible
		100	146.28	0.0403	0.2%	<0.0001	Negligible
Case	Actual	50	2855.03	0.1781	3.2%	<0.0001	Minor
		100	2769.73	0.1754	3.1%	<0.0001	Minor
	Model	50	6160.88	0.2616	6.8%	<0.0001	Major
		100	7211.65	0.2831	8.0%	<0.0001	Major
	Naïve	50	9802.54	0.3300	10.9%	<0.0001	Major
		100	9865.21	0.3311	11.0%	<0.0001	Major

The average actual and model standard errors found are shown in Table 5.5.

Table 5.5: Effect of the number of thresholds on accuracy of estimates

		Threshold estimates		Case estimates		
Items	Thresholds	Actual	Model	Actual	Model	Naïve
50	1	0.314	0.313	0.314	0.312	0.312
50	2	0.318	0.315	0.279	0.277	0.225
50	3	0.319	0.316	0.266	0.264	0.184
50	4	0.319	0.316	0.260	0.257	0.160
50	5	0.320	0.317	0.254	0.252	0.143
50	6	0.319	0.317	0.252	0.249	0.131
50	7	0.321	0.317	0.249	0.247	0.122
50	8	0.320	0.317	0.248	0.245	0.114
50	9	0.322	0.317	0.247	0.245	0.108
50	10	0.320	0.317	0.245	0.243	0.102
100	1	0.319	0.316	0.225	0.227	0.227
100	2	0.321	0.317	0.201	0.201	0.162
100	3	0.322	0.317	0.191	0.191	0.133
100	4	0.322	0.318	0.186	0.185	0.115
100	5	0.322	0.318	0.183	0.182	0.103
100	6	0.323	0.318	0.180	0.180	0.094
100	7	0.323	0.318	0.180	0.178	0.087
100	8	0.323	0.318	0.178	0.177	0.082
100	9	0.323	0.318	0.177	0.176	0.077
100	10	0.324	0.318	0.176	0.175	0.073

As discussed earlier, one would expect the number of thresholds to have a negligible effect on item and threshold estimates. Although the number of item categories shows statistically significant effects on threshold estimates, the effect size is small (much less than 1%). It can be seen that, in each test series, accuracy tends to decrease slightly with an increasing number of categories. This can be understood as having insufficient cases (50), and therefore observed responses, to discriminate among the large number of thresholds (up to 1000 with 100 items and 10 thresholds). However, overall, as expected from theory, it can be concluded that there was a negligible effect on the accuracy of item estimates or modelled standard errors.

For case estimates, however, a significant and meaningful effect was found, as expected, in all series. The estimates are shown graphically in Figure 5.3.

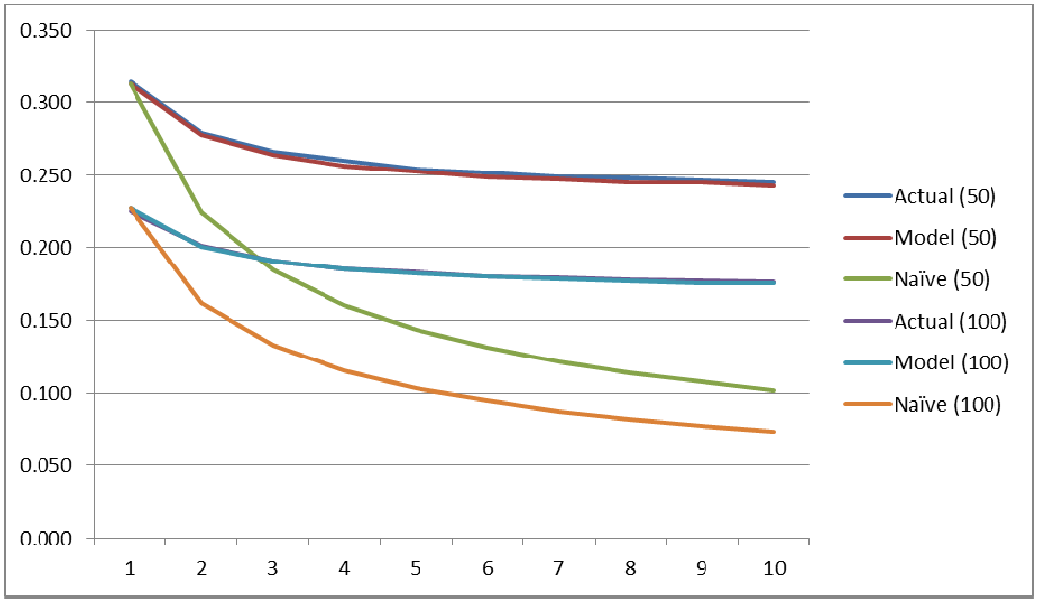


Figure 5.3: Effect of the number of item thresholds on case estimates

From this figure, it can be seen that there is a monotonic decreasing relationship between the number of thresholds and the standard errors. It can also be seen that there is a close agreement between the actual and modelled standard errors. Conversely, the naïve standard error calculation, which makes the assumption of threshold independence, clearly underestimates the true standard error. It can also be seen that the increased accuracy associated with additional thresholds is less than that associated with additional items. For example, both model and actual errors are around 0.32 for 50 items with 10 thresholds, whereas, they are around 0.23 for 100 items with 1 threshold. It can also be seen that there is relatively little improvement after the first few thresholds. This is consistent with the proposition made in this thesis that only thresholds adjacent to the chosen response category contribute information to the case estimate.

To summarise, it was found that the use of additional response categories has a negligible effect on the accuracy of item threshold estimates. There is a significant effect, however, on the accuracy of case ability estimates.

The effect of the number of thresholds on reliability and fit statistics was also investigated. The relevant data are set out in Table 5.6.

Table 5.6: Effect of number of item categories on fit statistics and reliability

No of Items	No of Thresholds	Alpha	COR	Case		Item		Separation	
				Infit	Outfit	Infit	Outfit	Item	Person
50	1	0.904	0.742	1.000	1.002	1.000	1.003	0.902	0.903
50	2	0.921	0.740	0.999	1.001	0.999	1.001	0.901	0.923
50	3	0.927	0.739	0.999	0.999	0.999	0.999	0.900	0.931
50	4	0.930	0.739	0.999	1.000	0.998	1.000	0.900	0.934
50	5	0.932	0.738	0.998	0.998	0.998	0.998	0.900	0.936
50	6	0.933	0.739	0.998	1.000	0.998	1.000	0.900	0.938
50	7	0.934	0.738	0.998	0.999	0.998	0.999	0.900	0.939
50	8	0.935	0.739	0.998	0.999	0.997	0.999	0.900	0.940
50	9	0.935	0.738	0.997	0.997	0.997	0.997	0.900	0.940
50	10	0.936	0.739	0.998	0.998	0.997	0.998	0.899	0.941
100	1	0.949	0.738	1.000	1.001	1.000	1.001	0.900	0.948
100	2	0.959	0.737	1.000	1.000	0.999	1.000	0.900	0.960
100	3	0.962	0.737	0.999	1.000	0.999	1.000	0.900	0.964
100	4	0.964	0.736	0.998	0.999	0.998	0.999	0.899	0.966
100	5	0.965	0.736	0.998	0.999	0.998	0.999	0.899	0.967
100	6	0.965	0.736	0.998	0.998	0.998	0.998	0.899	0.968
100	7	0.966	0.736	0.998	0.998	0.997	0.998	0.899	0.968
100	8	0.966	0.736	0.997	0.997	0.997	0.997	0.899	0.969
100	9	0.967	0.736	0.998	0.998	0.997	0.998	0.899	0.969
100	10	0.967	0.736	0.997	0.997	0.997	0.997	0.899	0.969

A comparison of Cronbach's alpha detected a significant difference in ranks across the thresholds. The statistics were:

- $H_{(9)}$=4863.15; ϕ_C=0.2325; ϕ_C^2=5.4%; p=0.0001 (50 item dataset).
- $H_{(9)}$=5786.11; ϕ_C=0.2536; ϕ_C^2=6.4%; p=0.0001 (100 item dataset).

A plot of Cronbach's alpha for the two datasets is given in Figure 5.4.

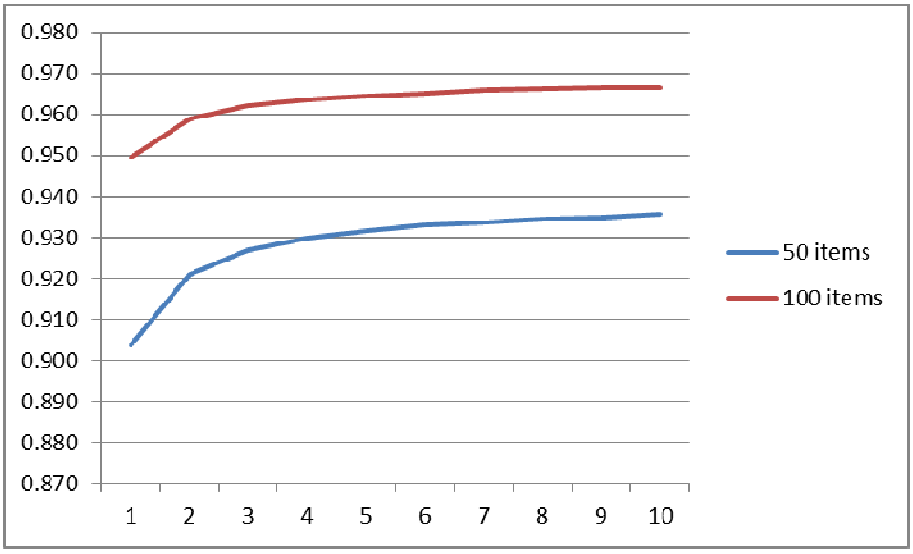


Figure 5.4: Effect of the number of categories on Cronbach's alpha

For both datasets, a progressive increase in Cronbach's alpha is visible with higher number of thresholds. However, the increase levels off after the first few thresholds. This suggests that the use of polytomous items with up to about five categories can be expected to produce a noticeable increase in Cronbach's alpha, but there is little benefit in using more than around five categories. It is also noticeable that, even with 10 thresholds, alpha for the 50 item series is lower than that for 100 items with 1 threshold.

As expected, there was also a small, but statistically significant effect, on Guttman's coefficient of reproducibility (COR). The statistics were:

- $H_{(9)}$=157.18; ϕ_C=0.0418; ϕ_C^2=0.2 %; p<.0001 (50 item dataset).
- $H_{(9)}$=124.13; ϕ_C=0.0371; ϕ_C^2=0.1 %; p<.0001 (100 item dataset).

A plot of the data is shown in Figure 5.5. Both the 50 and 100 item datasets show a reduction in reproducibility as the number of thresholds increases. This reduction is larger in the 50 item dataset. In both cases, the reduction is more noticeable for the first few thresholds. Taken as a whole, the 100 item dataset also shows lower reproducibility than the 50 item dataset. This is in accordance with theoretical expectations. However, the overall effect size is relatively small: there is less than 1% difference between the lowest reproducibility (73.6%) and the highest (74.2%).

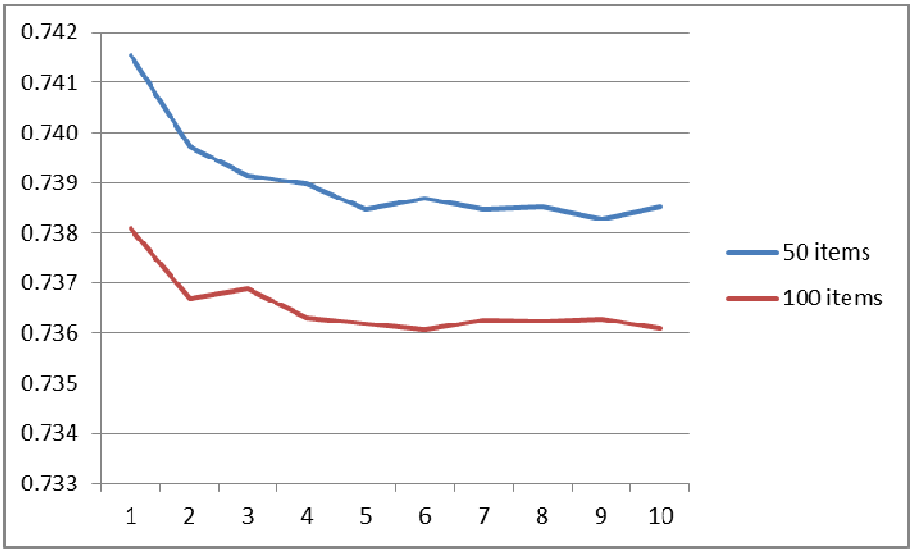


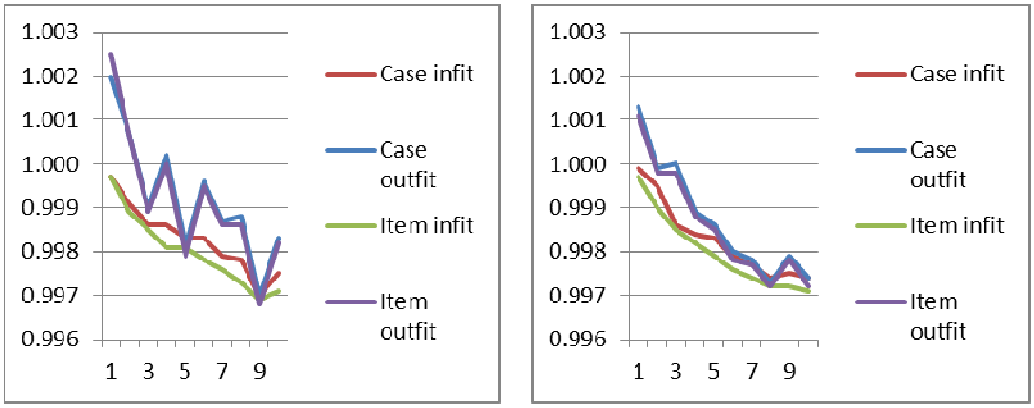
Figure 5.5: Effect of the number of categories on reproducibility

Small but statistically significant differences were found among the rank association of case and item fit statistics with the number of thresholds (Table 5.7).

Table 5.7: Significant effects of the number of item categories on fit statistics.

Dataset	Statistic	$H_{(9)}$	φ_C	φ_C^2	Significance	Comment
50 items	Case infit	146.73	0.0404	0.2%	<.01	Negligible
	Case outfit	33.11	0.0192	0.0%	<.01	Negligible
	Item infit	381.50	0.0651	0.4%	<.01	Negligible
	Item outfit	34.34	0.0195	0.0%	<.01	Negligible
100 items	Case infit	368.68	0.0640	0.4%	<.01	Negligible
	Case outfit	82.99	0.0304	0.1%	<.01	Negligible
	Item infit	761.23	0.0920	0.8%	<.01	Negligible
	Item outfit	86.01	0.0309	0.1%	<.01	Negligible

Plots of the fit statistics are shown graphically in Figure 5.6. A decreasing trend is noticeable in each of the statistics for an increasing number of thresholds. It is also noticeable that the infit statistics, which are less sensitive to outliers, show a smoother pattern than the outfit statistics. It can also be seen that this smoothness increases with a greater number of items. However, the overall magnitude of the effect is negligible when compared to the expected range of 0.5 to 2.0 of the fit statistics. This is consistent with theoretical expectations.



50 item dataset 100 item dataset

Figure 5.6: Effect of the number of item categories on infit and outfit statistics

Significant effects were also found for person and item separation. The results of a Kruskal-Wallis test are shown in Table 5.8.

Table 5.8: Significant effects of the number of thresholds on reliability.

Dataset	Statistic	$H_{(9)}$	φ_C	φ_C^2	Significance	Comment
50 items	Person separation	6160.77	0.2616	6.8%	<.01	Major
	Item separation	176.09	0.0442	0.2%	<.01	Negligible
100 items	Person separation	7211.61	0.2831	8.0%	<.01	Major
	Item separation	126.38	0.0375	0.1 %	<.01	Negligible

As expected, the number of thresholds has a major effect on person separation reliability (PSR), but only a minor effect on item separation reliability (ISR). The values of the indices are shown graphically in Figure 5.7. In this figure, it can be seen that the person separation reliability (PSR) increases monotonically with the number of thresholds in both datasets. This is to be expected because items with more categories give more information about case estimates than those with fewer. It is also noticeable that item separation reliability (ISR) is relatively unaffected by the number of thresholds.

Some tentative conclusions can also be drawn from these figures about the possible recommendations that could be given to educators about the number of cases and items, and the number of categories per item, to give useful measurement in a typical context.

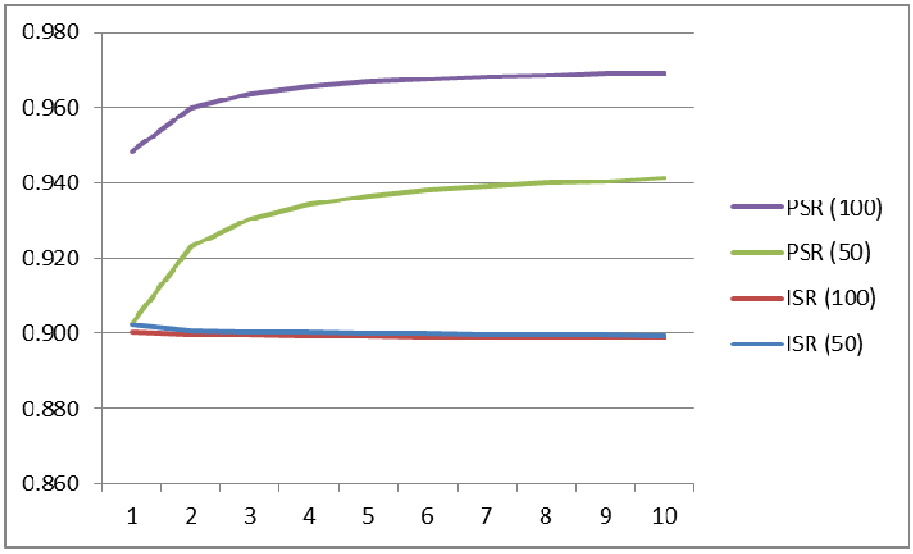


Figure 5.7: Effect of the number of thresholds on reliability.

The theoretical relationship between item separation reliability and the number of ideal subjects or cases was defined in the last chapter as $4/(1-ISR)$. Likewise, the number of ideal dichotomous items is $4/(1-PSR)$. Ideal in both cases refers to perfect targeting and perfect items or subjects. It therefore represents a minimum base line, rather than a realistic estimate.

With this approach, it is possible to evaluate the use of polytomous items in terms of the number of equivalent dichotomous items. It is then also possible to express the contribution of a polytomous item, relative to a dichotomous item, as a percentage. The appropriate calculations are set out in Table 5.9. In this table, it can be noticed that the number of *ideal* cases is approximately 40 in all circumstances. The lack of variation reflects the fact that item separation is only minimally affected by the number of items or thresholds. The level of 40 ideal cases, compared to the actual number of 50 cases, reflects the fact that perfect targeting is an unachievable ideal. Each real case is effectively contributing, on average, 80% of the maximum possible information for an ideal case. A similar conclusion can be drawn from the number of ideal items in the dichotomous case. The values of 41 and 78 can be seen as approximately 80%, again representing the lack of perfect targeting.

Table 5.9: Relative contribution of polytomous items

No of Items	No of Thresholds	Separation		Ideal Cases	Ideal Items	Relative Contribution
		Item	Person			
50	1	0.902	0.903	41	41	100%
50	2	0.901	0.923	40	52	**127%**
50	3	0.900	0.931	40	58	140%
50	4	0.900	0.934	40	61	**148%**
50	5	0.900	0.936	40	63	153%
50	6	0.900	0.938	40	65	157%
50	7	0.900	0.939	40	66	160%
50	8	0.900	0.940	40	67	162%
50	9	0.900	0.940	40	67	163%
50	10	0.899	0.941	40	68	165%
100	1	0.900	0.948	40	78	100%
100	2	0.900	0.960	40	99	**128%**
100	3	0.900	0.964	40	110	142%
100	4	0.899	0.966	40	116	**150%**
100	5	0.899	0.967	40	120	155%
100	6	0.899	0.968	40	123	159%
100	7	0.899	0.968	40	126	162%
100	8	0.899	0.969	40	128	164%
100	9	0.899	0.969	40	129	167%
100	10	0.899	0.969	40	130	168%

In the relative contribution column, it can be seen that in the 100 item dataset, a three category item contributes 28% (27% in the 50 item dataset) more information than a dichotomous item. Likewise, a five category item contributes 50% more information (48% in the 50 item dataset). These values, 28% and 50%, have been taken as realistic values for the expert system heuristics which are used to guide educators in the planning stage of measurement.

The findings in this section can be briefly summarised as follows. Using polytomous items has a negligible effect on item and threshold estimates but has a significant effect on case ability estimates. This is in accordance with theoretical expectations. Naïvely treating item thresholds as independent dichotomous items would grossly overstate the information available in the dataset about case ability estimates and consequently understate the standard error of the estimate. The theoretical implications of this are discussed in section 5.7. Using polytomous items can both improve the accuracy of estimates and improve reliability as measured by indices

like Cronbach's alpha or the Rasch separation indices. However, there is a diminishing return with increasing number of categories. A three category item gives approximately 28% more information than a dichotomous item, and a five-category item approximately 50% more. Thus, 10 five-category items contribute information roughly equivalent to 15 dichotomous items. The corresponding reductions in standard errors, compared to the dichotomous case, are approximately 11.5% for three-category items, and 18% for five-category items. The increase of accuracy from the use of polytomous items results in a slight loss of reproducibility although the effect size is small and negligible. There is also a negligible effect on infit and outfit statistics.

5.4. NUMBER OF CASES AND ITEMS

The following questions were investigated:

- How does the number of items affect estimates?
- How does the number of cases affect estimates?
- What are realistic lower limits for the number of cases and items?

Four series of tests were used in the experiment. In the first series, the dataset comprised 50 dichotomous items and the number of cases was varied between 10, 20, 25, 50 and 100. In the second series there were 50 cases and the number of dichotomous items was varied between 10, 15, 20, 25, 50, and 100. In the third series, the effect of using 3-category and 5-category items was investigated for 50 cases and 10, 15, 20, and 25 items. In the fourth series, selected combinations of dichotomous and 5-category items were used to explore feasible lower limits.

The same specification was used for all series: both item thresholds and case abilities were allocated from the central, medium range, distribution with a mean of 0; no noise was added. There were a total of 27 tests and each test was repeated 1,000 times giving a total of 27,000 tests. The results are shown in Table 5.10.

Table 5.10: Test results for effect of number of cases and items

No of	No of	Cats/	Case Std. Err.		Item Std. Err.		Separation indices		No of ideal	
Cases	Items	item	Actual	Model	Actual	Model	Item	Person	cases	Items
10	50	2	0.277	0.290	0.623	0.674	0.606	0.916	10	48
20	50	2	0.303	0.305	0.473	0.470	0.784	0.907	18	43
25	50	2	0.305	0.308	0.430	0.426	0.820	0.905	22	42
50	50	2	0.316	0.313	0.316	0.313	0.902	0.902	41	41
100	50	2	0.320	0.316	0.227	0.228	0.948	0.900	77	40
50	10	2	0.624	0.675	0.281	0.291	0.916	0.605	48	10
50	15	2	0.535	0.538	0.293	0.298	0.912	0.726	45	15
50	20	2	0.474	0.470	0.299	0.303	0.908	0.785	44	19
50	25	2	0.428	0.425	0.308	0.307	0.906	0.821	42	22
50	50	2	0.315	0.312	0.312	0.312	0.903	0.903	41	41
50	100	2	0.227	0.228	0.319	0.316	0.900	0.948	40	77
50	10	3	0.563	0.556	0.296	0.300	0.911	0.704	45	14
50	15	3	0.476	0.464	0.303	0.305	0.907	0.787	43	19
50	20	3	0.424	0.413	0.309	0.310	0.904	0.830	42	24
50	25	3	0.384	0.377	0.314	0.312	0.903	0.858	41	28
50	10	5	0.519	0.500	0.302	0.305	0.911	0.752	45	16
50	15	5	0.440	0.429	0.309	0.310	0.904	0.816	42	22
50	20	5	0.391	0.382	0.313	0.313	0.902	0.854	41	27
50	25	5	0.355	0.348	0.314	0.314	0.902	0.879	41	33
20	10	2	0.597	0.630	0.426	0.434	0.815	0.624	22	11
20	10	5	0.502	0.486	0.454	0.456	0.796	0.762	20	17
10	20	2	0.420	0.435	0.597	0.629	0.629	0.808	11	21
10	10	2	0.550	0.600	0.553	0.595	0.670	0.639	12	11
10	10	5	0.473	0.471	0.599	0.647	0.622	0.776	11	18
20	20	2	0.452	0.456	0.448	0.454	0.795	0.794	20	19
20	15	5	0.432	0.422	0.467	0.468	0.786	0.823	19	23

The separation indices were used as indicators of accuracy. These were also expressed as the equivalent number of ideal cases or dichotomous items, to give an indicator that is approximately linear. In the first series, the person separation index was relatively unaffected by the variation of the number of cases. This is expected because it depends on the information provided by the items and the number of items was constant. The item separation index showed a strong dependence on the number of cases, again, as expected. The reverse pattern was found in the second series in which the number of cases was held constant. The number of ideal cases and dichotomous items corresponding to the reliability indices is shown graphically in Figure 5.8.

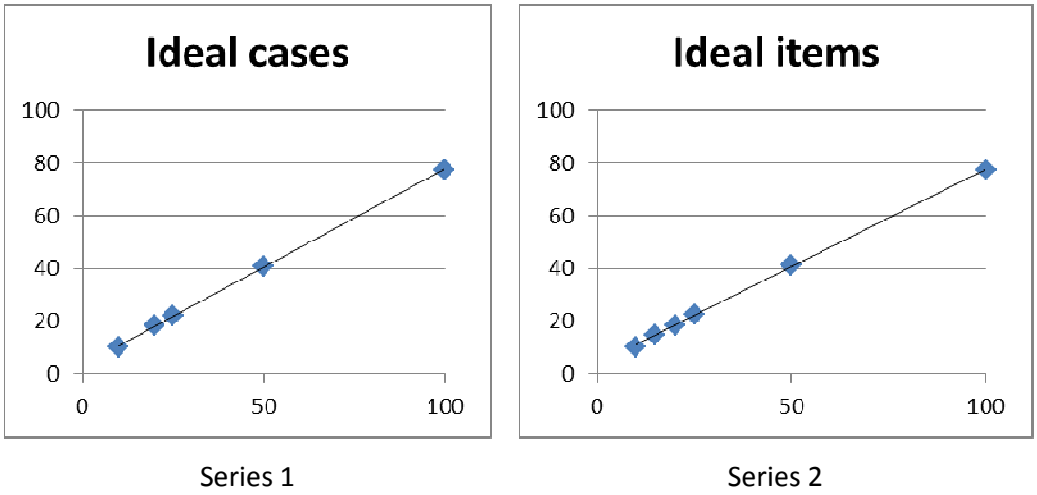


Figure 5.8: Relationship between ideal and real cases and items

The relationship shown is almost perfectly linear: the best fit line has an R^2 value of 0.9997 for series 1 and 0.9996 for series 2.

As discussed in section 4.3.3, the minimum reliability for measurement purposes is 0.610, corresponding to two distinct statistical strata. By inspection of the reliability indices, this suggests a practical minimum of 10 cases for the minimal purpose of classifying 50 dichotomous items as easier or harder. Similarly, a minimum of 10 dichotomous items are required to classify 50 subjects into an upper or lower category.

The third test series explored the effect of using polytomous items on the number of items required. Figure 5.9 shows the person separation reliability, expressed as the equivalent number of dichotomous items, for dichotomous, 3-category and 5-category items, with 50 cases and 10, 15, 20 and 25 items.

The linear relationship between accuracy and the number of items can be seen clearly in this figure; the R^2 values for dichotomous, three-category, and five-category items were 0.9986, 0.9999, and 1.0000 respectively. Extrapolation from these linear relationships suggests that the practical minimum, for classification of 50 subjects into two strata, is ten dichotomous items, eight three-category items or seven five-category items. Item separation remained above 0.9 throughout these tests: a level of accuracy corresponding to 4.3 strata.

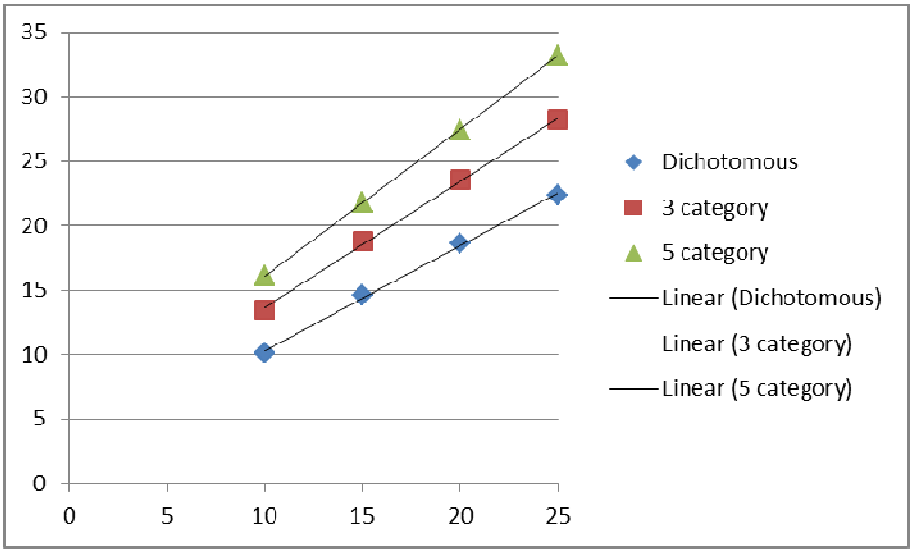


Figure 5.9: Effect of polytomous items on the number of items required.

However, the use of 50 cases provided a stable basis for the tests and it is not clear whether the same results would result with a smaller number of cases. To investigate this, the fourth series explored smaller numbers of items and cases. As it eventuated, the model remained stable in this context, and it can be seen that all of the tests produced reliabilities above 0.610. This suggests that the minimal measurement purpose of two distinct statistical strata can be achieved with as few as 10 cases and 10 dichotomous items. A more realistic purpose, however, is triage, as defined in this project, which requires separation indices of 0.8 or more.

The combination of 10 cases and 20 dichotomous items achieves the purpose of triage for subjects and minimal separation of items. A similar result could be achieved with 10 cases and about 12 five category items. The purpose of triage on both items and cases would require around 20 cases and 20 dichotomous items or 15 five category items.

In summary, these findings suggest that it is possible to achieve useful measurement with as few as 20 subjects or items. These numbers are realistic and achievable in many educational settings.

5.5. CASE AND ITEM THRESHOLD DISTRIBUTIONS

The following questions were investigated:

- How does the distribution of abilities and difficulties affect estimates?
- How do these distributions affect reliability and fit statistics?

For the investigation, datasets with 50 cases, and 50 items, were used. These values were chosen as typical of an educational setting. A mean case ability of 1.0, which corresponds approximately to a 75% pass rate, and the central shape, and medium range, were also chosen as typical. To allow systematic exploration of its effects, these typical values were used as a baseline and each of the elements was then permuted, in turn, in accordance with the permutation scheme summarised in Table 5.11.

Table 5.11: Permutation scheme for testing distribution effects

Series	Item shape	Item range	Case shape	Case range	Case mean
1	Central	Medium	Central	Medium	Permute
2	Permute	Medium	Central	Medium	1.0
3	Central	Permute	Central	Medium	1.0
4	Central	Medium	Permute	Medium	1.0
5	Central	Medium	Central	Permute	1.0

Three options were permuted in each series, giving a total of 3,000 tests in each series and 15,000 tests overall. A summary of the results is shown in Table 5.12.

From these results, it can be seen that the shapes of the distribution had little effect. All of the indicators remained noticeably stable across the shape permutations for both items and cases. This accords with theoretical expectations, because no assumptions about the shape of the distribution were made in the derivation of the model. However, there were noticeable effects, as expected, for variations in range and case mean.

The case mean represents the alignment of case abilities and item difficulties, or equivalently, the targeting of items to cases. As expected, this targeting had a major effect on the accuracy of the estimates. Accuracy, as measured by standard errors, reduced by about .04 between 0 and 1 logits and by about .09 between 1 and 2. The corresponding information loss was approximately 17% and 60%, respectively. These values correspond, roughly, to pass rates of 50%, 75%, and 90% for a pass mark of 50%. This illustrates that targeting, or equivalently the pass rate, has a major effect on accuracy.

Table 5.12: Results of the tests of distributional effects

	Case errors		Item errors				Case fit		Item fit		Reliability	
Series	**Actual**	**Model**	**Actual**	**Model**	**Alpha**	**COR**	**in**	**out**	**in**	**out**	**item**	**case**
Case Mean												
0	0.31	0.31	0.31	0.31	0.90	74%	1.00	1.00	1.00	1.00	0.90	0.90
1	0.35	0.35	0.35	0.36	0.90	77%	1.00	1.00	1.00	1.00	0.88	0.88
2	0.44	0.47	0.50	0.65	0.92	84%	0.98	0.98	0.98	0.98	0.80	0.80
Item shape												
Central	0.35	0.35	0.35	0.36	0.90	77%	1.00	1.00	1.00	1.00	0.88	0.88
Extreme	0.35	0.35	0.35	0.34	0.90	77%	1.00	1.00	1.00	1.00	0.88	0.88
Uniform	0.35	0.35	0.35	0.34	0.90	77%	1.00	1.00	1.00	1.00	0.88	0.88
Item range												
Narrow	0.36	0.38	failed	0.99	0.91	75%	0.99	0.99	1.00	1.00	0.09	0.87
Medium	0.35	0.35	0.35	0.35	0.90	77%	1.00	1.00	1.00	1.00	0.88	0.88
Wide	0.36	0.36	0.26	0.35	0.87	81%	1.00	1.00	0.95	0.96	0.94	0.87
Case shape												
Central	0.35	0.35	0.35	0.36	0.90	77%	1.00	1.00	1.00	1.00	0.88	0.88
Extreme	0.34	0.34	0.35	0.36	0.90	77%	1.00	1.00	1.00	1.00	0.88	0.89
Uniform	0.35	0.34	0.35	0.36	0.90	77%	1.00	1.00	1.00	1.00	0.88	0.88
Case range												
Narrow	failed	0.99	0.38	0.41	0.07	75%	1.00	1.00	0.99	0.99	0.87	0.09
Medium	0.35	0.35	0.35	0.35	0.90	77%	1.00	1.00	1.00	1.00	0.88	0.88
Wide	0.24	0.29	0.36	0.36	0.97	81%	0.95	0.95	1.00	1.01	0.87	0.94

Model fit was very close to the expectation of 1.0 in the first two tests, but may be seen to fall slightly in the third test. Case and item fit statistics, and similarly infit

and outfit statistics were also in accord. However, the effect on the fit statistics was small when compared to the effective range (0.5 to 2.0) of these statistics. This is in accordance with theoretical expectations. Guttman's coefficient of reproducibility increased noticeably as the distance between case and item means was increased. Again, this is as expected from theory. Finally, Cronbach's alpha also increased slightly as the distance increased. This is also a known effect (Linacre, 1997).

The range of item difficulties and case abilities also had a major effect. First, measurement failed for many items when the item range was narrow and for many cases when the case range was narrow. This is because, with the default targeting of a one logit difference between case and item means, the narrow range of 0.5 logits placed many of the estimates in an extreme part of the distribution where subjects were likely to get items "all right" or "all wrong", thus leading to immeasurable estimates. The corresponding reliability indices can also be seen to reflect this loss of accuracy.

For the remaining medium and wide ranges, the following effects can be seen. First, case accuracy increased with a wider case range. Likewise, item accuracy increased with a wider item range. The corresponding reliability indices also increased accordingly. Person fit was also more deterministic with a wider case range, corresponding to infit and outfit statistics less than 1. Likewise, item fit was more deterministic with a wider item range. Guttman's coefficient of reproducibility also increased with range, reflecting the reduced uncertainty associated with more deterministic responses. Finally, Cronbach's alpha increased with a wider case range, but reduced with a wider item range. These are also known effects (Linacre, 1997).

In summary, the shape of the distribution has little effect on estimates, reliability indices or fit statistics. Conversely, both targeting and the ranges have a major effect. In general, accuracy is reduced as the difference between case and item means increases, and this situation typically corresponds to higher pass rates. Also, the use of a wider range tends to increase accuracy. Little can be done, in practice, about the distribution of subject abilities, but this observation leads to the general

advice to use a wide range of item difficulties when constructing measurement instruments.

5.6. INTRODUCED ERRORS

The following questions were investigated:

- How does the introduction of error affect estimates?
- How does the introduction of error affect reliability and fit statistics?

Four series of tests were used in the investigation. All used the central shape, medium range, and zero case mean. A dataset of 50 items and 50 cases was used. The introduced error (noise) was drawn from a normal distribution with a zero mean, and a specified percentage of the case standard deviation. This was added to the allocated case ability to create a perturbed ability that was used for the allocation of responses. The noise levels used were: 0%, 25%, 50%, and 100%. Each test was repeated 1,000 times, for a total of 4,000 tests. The results are shown in Table 5.13.

Table 5.13: Results of the tests of introduced noise.

Added	Case errors		Item errors				Case fit		Item fit		Reliability	
Noise	Actual	Model	Actual	Model	Alpha	COR	in	out	In	out	item	case
0%	0.315	0.313	0.315	0.314	0.903	0.742	1.00	1.00	1.00	1.00	0.902	0.902
25%	0.318	0.315	0.318	0.315	0.902	0.740	1.00	1.00	1.00	1.00	0.901	0.901
50%	0.322	0.320	0.324	0.321	0.899	0.736	1.00	1.00	1.00	1.00	0.897	0.898
100%	0.346	0.343	0.345	0.343	0.883	0.722	1.00	1.00	1.00	1.00	0.883	0.883

From these results, it can be seen that the model is relatively robust under the introduction of random noise. As expected, increased levels of added noise resulted in increased standard errors, and a reduction of accuracy. The separation reliability indices of both case and item estimates also reduced accordingly. However, no effect on fit statistics was observed. This was unexpected because the literature (Wright & Linacre, 1994) suggests that they should increase. On reflection, this makes perfect sense. The fit statistics are based on the standardised residual where the standardisation is achieved by dividing by the *modelled* variance; for example,

see equation 3.36 (p. 140). Any model can be conceptualised as: observed = modelled + unexplained error. The standardisation by modelled variance, rather than observed variance, means that the fit statistics measure the departure of the model *component* from its expectations, and are thus not affected by the magnitude of the error component. Cronbach's alpha also reduced progressively with increased noise, as did Guttman's coefficient of reproducibility. With the exception of the fit statistics, all of these effects are as expected by theory. The effect on reliability is shown graphically in Figure 5.10.

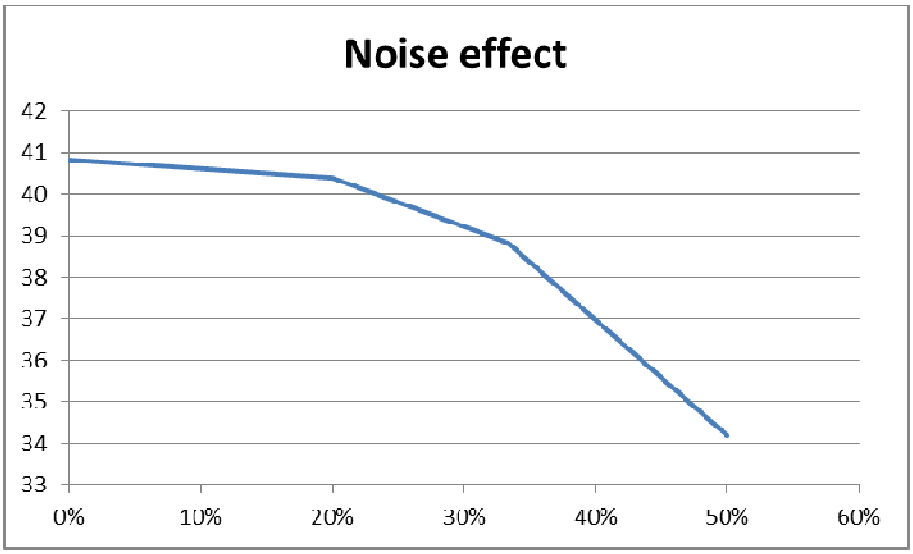


Figure 5.10: Effect of added noise on reliability

In this figure, the x-axis shows the percentage of noise, expressed as a proportion of the total: signal + noise. The y-axis shows the case separation reliability, expressed as the number of equivalent dichotomous items. The figures for item reliability are essentially the same and not shown in this figure. It can be seen that a 25% noise level (20% of the total) results in minimal loss of information and that, although the information loss increases monotonically, the loss of information is proportionately less than the proportion of noise. Even at 100% noise (50% of the total), there is only approximately 20% loss of information. This robustness is characteristic of many stochastic measurement models: as long as the error is uncorrelated with case or item characteristics, there is relatively little degradation in measurement.

This observation leads to the final series of tests carried out: the systematic distortion of measurement. As discussed in section 2.4.1, the central assumption of the model is that of conditional (or local) independence. Accordingly, the effect of breaching this assumption was investigated. Four series of tests were used in the investigation. All tests used the central shape, medium range, zero case mean, and zero noise. A dataset of 50 items and 50 cases was used. To create the dependency, each case had a specified percentage chance of being averaged with the previous case, thus creating a perturbed ability with local dependence. The first series used a specified chance of 0% to create a baseline. The remaining series specified the chance as: 20%, 40%, and 60%. Each test was repeated 1,000 times, for a total of 4,000 tests. The results from these investigations are set out in Table 5.14.

Table 5.14: Effects of introduced dependency.

Introduced	Case errors		Item errors				Case fit		Item fit		Reliability	
Dependency	Actual	Model	Actual	Model	Alpha	COR	in	out	in	out	item	case
0%	0.315	0.313	0.315	0.314	0.903	0.742	1.00	1.00	1.00	1.00	0.902	0.902
20%	0.360	0.342	0.315	0.315	0.883	0.732	1.00	1.00	1.00	1.00	0.901	0.883
40%	0.439	0.371	0.314	0.314	0.860	0.725	1.00	1.00	1.00	1.00	0.901	0.863
60%	0.543	0.396	0.315	0.314	0.840	0.720	1.00	1.00	1.00	1.00	0.902	0.844

From these results, it can be noticed that there is a negligible effect on the accuracy of item estimates, item reliability or on the fit statistics. There is, however, a significant effect on case estimates, person separation reliability, Cronbach's alpha, and Guttman's coefficient of reproducibility. As expected, both actual and modelled standard errors increase as the level of dependency increases, but the modelled standard error increases more slowly and significantly underestimates the actual error at the upper end. This disagreement between model and actual is an artefact of the specific mechanism used to create the dependency. The modelled standard error is based on the variance of observations around the modelled location, whereas the actual standard error is based on the variance between the modelled and the actual allocated ability which was known only to the test management software. The actual error thus comprises two components: the actual measurement error, and a component representing the difference between the

perturbed ability used for a percentage of the time, and the real allocated ability, used for the rest of the time. Nevertheless, this difference highlights the need for caution in using any outputs when the assumptions of the model do not hold. The person separation reliability index also indicates the reduced accuracy as the level of local dependence is increased. As expected from theory, both Cronbach's alpha and Guttman's coefficient of reproducibility progressively reduce as the level of dependency increases.

The effect is shown graphically in Figure 5.11.

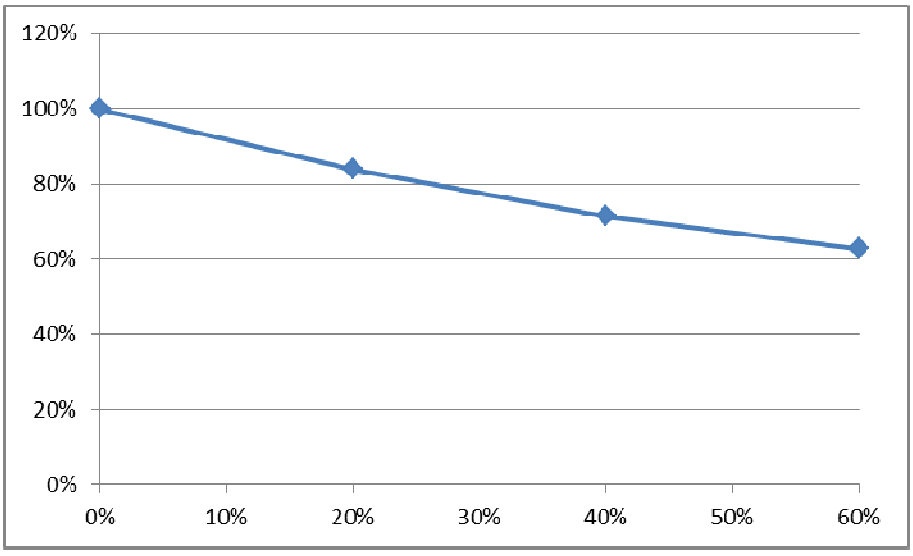


Figure 5.11: Effect of introduced local dependency.

In this figure, the x-axis shows the percentage of cases for which the ability estimate was perturbed and the y-axis shows the person separation reliability, expressed as the number of equivalent ideal dichotomous items, and shown as a percentage of the number of ideal items for zero distortion. It can be seen that the relationship is approximately linear. As expected, local dependence has a major impact on the accuracy of the estimates.

5.7. THEORETICAL IMPLICATIONS

The conceptualisation and model used in this thesis for polytomous items differs from that of the widely-used Andrich and Masters models. Moreover, the simulations described in this chapter provide support for the proposed model.

Thus, this thesis makes a theoretical contribution to the understanding of generic measurement, and its practical implementation. This contribution is presented in the first subsection. Also, the information theoretic approach used allows the software to move beyond diagnosis of local dependence to the implementation of a correction for diagnosed dependence. This is described in the second subsection.

5.7.1. Polytomous Items

The contribution to theory is summarised in the following theorems:

- Theorem 1: *The natural score cannot be a sufficient statistic for case estimates when polytomous items are used*.
- Theorem 2: *The information about a case estimate comes from thresholds, not from categories.*
- Theorem 3: *Only those thresholds that immediately bound a category contribute information to case estimates.*

Formal proof of these theorems is outside the scope of this thesis, but the logical argument for them is presented here. In the dichotomous case, the model used in this thesis is isomorphic to the dichotomous Rasch model, as are the models of Andrich and Masters. One characteristic of the Rasch model is that the natural score is a sufficient statistic, and this is taken as the starting point for the derivation of the Andrich and Masters models. However, another, equally important, characteristic of the Rasch model is that items are exchangeable, as are cases. That is, unlike the various IRT models, the characteristic curves are all parallel and this equivalence of slope leads to exchangeability. Moreover, it is this exchangeability that leads to the natural score as a sufficient statistic in the dichotomous model. Essentially, the logic is that, if all items (or cases) are alike, then counting successes captures all the available information. However, the model used in this thesis takes as its starting point the assumption of conditional independence, rather than that of a score as a sufficient statistic, and the model is then developed using standard information theory. Theorem 1 asserts that, although the natural score is a sufficient statistic for the dichotomous model, it cannot be for a polytomous model. This assertion is justified as follows.

In any test, a candidate may succeed on some items and not succeed on others. The principle of exchangeability leads to the conclusion that it does not matter which specific items constituted successes. Thus, under the standard Rasch model, it is sufficient to count successes to determine ability. Likewise, candidates are considered equivalent, and item difficulty can be determined by counting the number of candidates who succeeded on the item. Moreover, a candidate's responses to the items are independent. For any pair of items, it is possible for a candidate to succeed on both, on neither, or on either one but not the other. A similar observation can be made from an item perspective. For any pair of candidates, both could succeed, or neither, or just one. This independence is at the heart of the Rasch model and it allows simple summation of the information provided by each response to determine estimates and the associated standard errors. Each response contributes one unit of information, and the model is symmetrical with respect to items and subjects.

The use of polytomous items removes this symmetry. Each item has a number of mutually exclusive categories, and it is not possible for a candidate to choose each, none, or an arbitrary combination of categories: these choices are not independent. When the categories are ordinal, they can be conceptualised as defined by cut points on an underlying continuum. A response can then be treated as a judgement that the underlying value is above one cut point and below another. Each judgement therefore represents a comparison of the underlying value to two cut points, or in the case of extreme categories, to a single cut point. Each comparison, the judgement that the value is above or below the cut point, contributes one unit of information. The essential point is that the *information* comes from the comparative judgement, rather than from any property associated with the category (theorem 2). However, the comparisons are not independent. It is not possible to record a judgement that the value is simultaneously below a lower threshold and above an upper. The consequence of this lack of independence is that the information cannot simply be summed across thresholds to determine case estimates and standard errors. The converse does not apply, however. When estimating any specific item threshold, subjects remain independent: for any pair of

subjects, both, neither, or just one could be judged as above the threshold. Thus, it is clear that the estimation process must be different for cases and thresholds.

For case estimates, the approach of simply summing information across thresholds is termed *naïve* in this thesis, by analogy to the use of that term for the independence assumption of the Bayes classifier. This naïve approach clearly overstates the available information, thus underestimating standard errors, as demonstrated by the simulations in section 5.3. From the description above, it can also be seen that, at most, two thresholds can contribute information (theorem 3). Consequently, the model used in this thesis sums information across the thresholds that immediate bound a chosen category to determine case estimates. The results shown in Figure 5.3 (p. 231) show the close agreement between actual standard errors and the model errors determined by this process. It is important to note that the simulation process used responses generated directly from category probabilities, and did not make the assumption that only boundary thresholds contribute information. Thus, the simulations provide strong empirical evidence for the correctness of this logic.

The natural score is simply a count of the number of thresholds that have been passed. It has been argued above that only those thresholds that immediately bound a category can contribute useful information. Further, the response category chosen is not independent of the choice of other categories. Moreover, unlike dichotomous items, thresholds in polytomous items are not exchangeable either within or between items. It follows that the natural score cannot be a sufficient statistic.

Further evidence for the proposition can be given as follows. When creating responses, it is not necessary for a candidate or scorer to consider all the available categories; it is sufficient to check the boundary conditions around the chosen category. From an information theoretic perspective, the information provided comes from the comparative judgements that are made. However, it is clear that no new information can be provided by judgements that are not made - only those thresholds that are actually considered can contribute information. Although

categories below the chosen category can be deemed to have been reached, and those above not to have been reached, this is implicit in the ordering of the categories and does not represent any new information.

5.7.2. Information Correction

As discussed under the relationship to the Naïve Bayes Classifier (section 2.4.1), the central assumption of the model is that of conditional independence. Where this does not hold, for example when the response of a person depends in part on their response to another question, the unique information provided by the response is overstated and, consequently, the standard errors based on this response are understated. This is well understood (Wang & Wilson, 2005) and can be diagnosed by statistics such as Yen's Q_3 (1984). Local dependence is usually treated as an error to be remedied by removing or replacing the items exhibiting dependency from the scale. However, this precludes the use of complex scenarios associated with a set of questions, and such items can be a useful part of a test designer's toolkit.

To enable the use of such items, an option was added to the software used in this thesis to manage local dependence by evaluating the degree of local dependence and then making an appropriate correction to the assessed information contribution. Details of the method were given in section 3.6. A similar correction can be applied to remedy diagnosed dependency among cases. The analysis in section 5.6 shows a near-linear relationship between local dependence and accuracy, as measured in terms of information contributed. This suggests that the linear regression approach used in the management option will produce a very close approximation to the actual standard error.

Because of the central importance of the assumption of local independence to the model, diagnosed dependence is still considered an issue to be considered carefully before measurements are used. However, the availability of the corrective option makes it possible for test designers to consider using sets of related items, judiciously, as discussed above.

5.8. PRACTICAL IMPLICATIONS AND GUIDANCE

Some practical guidance for test construction can be inferred from the findings presented in this chapter. The investigation of polytomous items suggests that items with more categories provide more information about case abilities, and thus, greater accuracy, but that the benefit falls off rapidly, and little benefit is seen beyond four or five categories. However, accuracy of measurement depends on the proximity of the difficulty of item thresholds to case abilities. In addition to the number of categories, some comments can be made about the ideal width of each category. This is determined by the relative difficulty of the various cut points. If these are too close together, the categories might not achieve sufficient separation of subjects to provide useful information. If they are too far apart, they might not discriminate well enough between the subjects in the range and additional cut points might improve accuracy. The relationship between the width of a category (in logits), and the extent to which a category is discriminated from its neighbours, is shown in Figure 5.12.

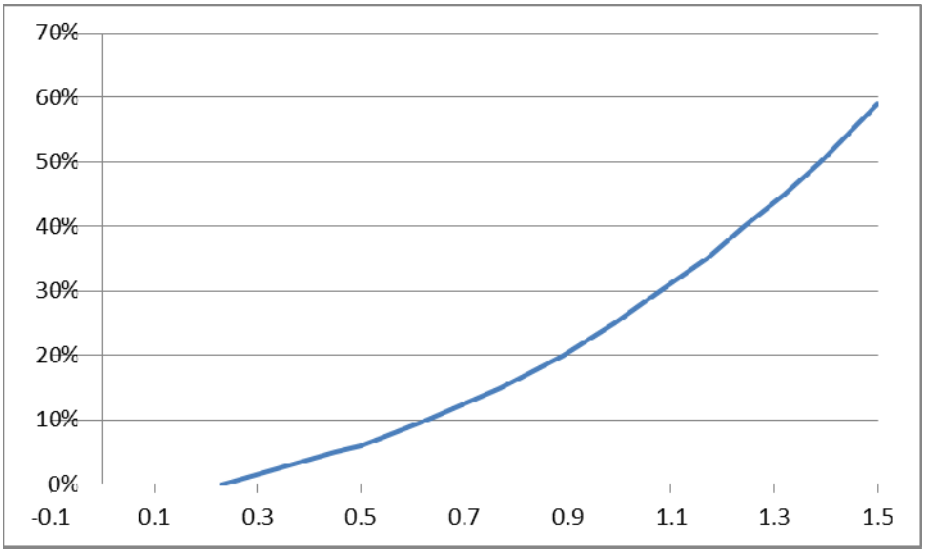


Figure 5.12: Dominance of a category by logit category width

In this figure, the x-axis represents the category width in logits, and the y-axis the relative likelihood of a choice in the category compared to a choice in the adjacent categories. For example, a value of 20% denotes that a subject with ability at the centre of the category is 20% more likely to choose the category than an adjacent

category. With a category width of 0.23 logits or less, a category does not become modal: it is more likely that a response in the categories to either side will be chosen in preference to the category. This is a clear indication that the category width is too small. A width of 1.0 seems a practical guideline. At this width, a response in the category is 25% more likely than in its neighbours and a five-category item will have an effective width of three logits for its inner categories. At a width of 1.5, the response in the category is 60% more likely than in its neighbours; this might be a good choice for a three-category item. These widths can also be expressed in terms of success rates. If the lower end of a one logit interval represents a 50% chance of success for a student, the upper end represents a 27% success rate. At 1.5 logits, the upper end would represent a success rate of about 18%.

It was found that the shape of the case and item distributions has little effect on accuracy, but both range and targeting have a major effect. The use of a wider range tends to increase accuracy. Since little can be done about the distribution of subject abilities, this suggests using a wide range of item difficulties when constructing measurement instruments. Targeting is also important. Accuracy is reduced as the difference between case and item means increases. This situation corresponds to higher pass rates or endorsement rates. For general research purposes, ideally, the overall expectation should be 50% endorsement. However, for general educational measurement, the expectation is generally for success rates well above 50% and this is potentially challenging for measurement.

Accuracy of measurement depends on the proximity of item thresholds to case abilities. This suggests the use of polytomous items in preference to dichotomous items for educational measurement. Polytomous items can provide information over a wider range and thus help ensure the proximity of, at least some, thresholds to cases. The use of up to five categories is a reasonable choice for most educational purposes.

From this discussion and the simulation results, the following general guidelines can be set out.

- Fifteen to twenty five-category items should be sufficient to achieve the purpose of triage for subjects.
- Twenty to twenty five subjects should be sufficient to classify items as easy, medium or hard.
- A measurement instrument should have a mix of easy, medium, and hard, items to promote accuracy over a reasonable range of subject abilities.
- Sufficient items should also be targeted at the region of any critical decision points.

As an example of this last guideline in an educational context, targeting sufficient items at the pass/fail boundary in high stakes assessment will increase accuracy for subjects in that critical region.

5.9. ROBUSTNESS AND LIMITATIONS

The simulations have demonstrated that the model is stable down to as few as ten cases and items. They have also shown that useful information can be produced with relatively few subjects and items. This suggests that the implementation is practical in realistic educational settings.

As shown in section 5.6, the model is tolerant of added noise, and is reasonably well behaved under departures from its core assumption. However, accuracy is affected and, in the latter case, imputed standard errors may be inaccurate unless the software option to manage this has been chosen.

This last point highlights a key feature of the model. Ultimately, no model can produce accurate measurement from poor data. Thus, an important feature is the measurement hypothesis tests that are carried out, automatically, in conjunction with the measurement process: the inherent strength of the model is that it is readily refutable. This leads, logically, to its implementation in a context of continuous improvement. Each administration is then both a test of the model and instrument, and a source of data that can inform future improvements.

There are two key limitations to the findings. First, although the use of simulation allows systematic exploration of the model, and the use of 1,000 replications of

each simulation gives reasonable confidence in the findings, any simulation may lack ecological validity. In particular, the test instruments were constructed randomly, in accordance with the specified criteria. Real test instruments are constructed purposefully, and with a great deal of care and attention. Thus, some of the constructed test instruments may have contained some items with poor characteristics that would have been culled on inspection in a real context. Broadly, this is likely to lead to an understatement of accuracy in the simulations.

The second limitation relates to the heuristics and guidelines derived from these simulations. The aim of these is to simplify a complex reality so as to give guidance that is useful in practice. Inevitably, there will be an element of oversimplification in doing this. Some of this guidance may be only approximate, and there may even be situations in which the guidance is misleading.

5.10. CHAPTER SUMMARY

Three goals were pursued in this chapter. The first was to evaluate how well the implemented model accords with theoretical expectations. The second was to identify characteristics that could be used to inform the expert system heuristics, guide planning, and support interpretation of the outputs. The third was to evaluate robustness by exposing the implementation to data that do not fit the model.

There was close agreement between the results of the simulations and theoretical expectations. Item (threshold) estimates were affected mainly by case parameters, and case estimates by item parameters. The relationship between the number of cases and items and their information contribution was very close to a perfect linear relationship (R^2 above .999) for both dichotomous and polytomous items. This strongly supports the essential correctness of the information theoretic approach taken in this thesis. The only indicator that was not materially as originally predicted was that added noise did not noticeably increase the fit statistics. However, with hindsight, this makes perfect sense, as discussed in section 5.6.

The simulations have also allowed the second goal to be achieved. All of the simulations suggest that the separation reliability indices provide an appropriate indication of accuracy, and thus fitness for purpose. This supports the use of reliability as the central organising element of the expert system. Practical guidance has been derived that can inform useful test design in an educational context. Furthermore, the simulations suggest that useful measurement can be achieved with relatively few subjects and items, thus supporting the proposition that useful measurement can be achieved in realistic educational settings.

The model was robust under the introduction of random noise. As expected, this resulted in reduced accuracy, but the degradation was not disproportionate. Moreover, the imputed standard errors correctly reflected the actual accuracy achieved. The systematic distortion of measurement, by violation of the core assumption of local independence, also reduced accuracy but, again, the degradation was not disproportionate. However, in this case, the imputed standard errors were possibly inaccurate. Nevertheless, the local dependence was diagnosed by the software under the test of hypothesis 6, and this should be taken as a warning that the reported standard errors may be inaccurate. Accordingly, it is recommended that the software option to manage local dependence is chosen if the decision is taken to use the output measurements when local dependence is diagnosed. It is important to note that no model can give valid measurement from poor data and that, consequently, the tests of the measurement hypotheses are a necessary part of the measurement process.

In addition to these goals, the simulations support the contributions to measurement theory made in this thesis, as discussed in section 5.7.

To this point, the measurement model has been described, from its theoretical foundations, through to its implementation in software and theoretical evaluation. The presented model is suitable for generic measurement in an educational setting. The next two chapters provide an empirical evaluation of the model in the specific context of the measurement of self-efficacy and challenge that is the focus of this thesis.

Chapter 6.

SELF-EFFICACY AND CHALLENGE

Man often becomes what he believes himself to be. If I keep on saying to myself that I cannot do a certain thing, it is possible that I may end by really becoming incapable of doing it. On the contrary, if I shall have the belief that I can do it. I shall surely acquire the capacity to do it, even if I may not have it at the beginning. ~ Mahatma Gandhi

The quotation by Mahatma Gandhi above emphasises the importance of self-belief as an influence of one's actions and behaviour in life. In general society, the effect of self-belief on success has been long understood. In act 1, scene IV, of Shakespeare's *Measure for Measure* (Alchin, 2005), Lucio remarked "Our doubts are traitors, and make us lose the good we oft might win, by fearing to attempt". However, within the field of psychology, its importance has been recognised only much more recently.

There are many aspects to self-belief. The focus in this chapter is on one in particular: self-efficacy. Bandura (1994) defined self-efficacy as "people's beliefs about their capabilities to produce designated levels of performance that exercise influence over events that affect their lives" (p. 71). Bandura's background was in behaviourism and, to set the context for his theory, the first section gives a brief overview of the behaviourist agenda and history, which were antecedents of the development of Bandura's *Social Cognitive Theory*. This theory is then outlined in the second section. The third section describes Bandura's concept of self-efficacy and introduces the concept of challenge as its mirror image. The fourth section reviews the implications and the importance of self-efficacy for educators. The fifth section discusses the need for objective evidence and regular monitoring of self-efficacy. The sixth section discusses the approach to measurement of self-efficacy. Finally, the chapter is summarised in the last section.

6.1. BEHAVIOURISM

From a metaphysical perspective, the origins of behaviourism can be traced to the atomist views of the pre-Socratic Greek philosophers Democritus and Lucretius. In more recent times, the origins can be traced to Julien Offray de La Mettrie. In his 1748 work, *L'homme machine* [man a machine] (La Mettrie & Busey, 1912), he put forward a materialistic view that conduct is governed entirely by positive and negative reinforcements. Condillac combined La Mettrie's physiological materialism with empiricism and in his 1754 *Treatise on Sensations* (Falkenstein, 2010), foreshadowed what would become central tenets of behaviourism. He argued that sensation was the foundation of thought and perception, and that experience not only provides us with the raw materials for knowledge, but also teaches us how to focus attention, remember, imagine, abstract, judge, and reason. For example, a sensation might contain information that goes unnoticed by a subject simply because it is not attended to. People come to perceive what they have already sensed by selectively attending to each of its parts in turn, and then noting how these parts are related to one another. This act of attention is not innate so we need to learn how to attend to what we sense, with experience as our teacher. We attend first to what promises to satisfy our needs and interests. These in turn are a product of past experience that has made us aware of what objects are connected with the frustration or satisfaction of those needs and interests, developed as a consequence of a past experience of pleasure and pain. He also argued that when multiple stimuli regularly occurred together, they would be perceived as a single event with multiple characteristics. Thus, substance is a collection of sensations, qualities, or properties, which are commonly observed to occur together. Condillac set the ground for modern behaviourism by articulating an exclusively mechanistic theory of behaviour grounded in a stimulus-response model.

However, there remained the question of precisely how stimuli and their effects determine human conduct, and this led to a methodological search for laws of behaviour analogous to the laws of physics. Thorndike (1898) carried out a series of experiments in which cats were placed in *puzzle boxes*, and needed to discover how to operate a release mechanism to get access to food. He concluded that the cats

used a trial and error strategy, and that behaviours that led to a successful outcome were *stamped in*, and unsuccessful behaviours were *stamped out*. Once behaviour was learned in this way, the cat would carry out the action at once when it was shut in the box. From this observation he formulated the basic law of operant learning. This *law of eff*ect held that:

> Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; … The greater the satisfaction … , the greater the strengthening [of the bond]. (Thorndike, 1911, p. 244)

Although the law of effect was to contribute to the later theory of connectionism, his main contribution to psychology was, perhaps, methodological. Prior to Thorndike, psychology was concerned largely with introspective accounts of the furniture of the mind, and with attempts to organise these into coherent theory. Thorndike used a systematic experimental approach and it is this combination of Condillac's theory with scientific method that perhaps best characterises modern behaviourism.

Another early influence on behaviourism was the work of Ivan Pavlov, who is widely known for his demonstration of a *conditioned reflex*. He began his work by studying the digestive system of dogs. He published his results in 1897 (Pavlov, 1910) and received the 1904 Nobel Prize for medicine in recognition of this work. He believed that the digestive system was controlled jointly by physiological processes and the psyche. He persisted for several years in the belief that appetite produced a *psychic secretion* (Todes, 1997). According to Todes (1997), it was Snarskii, one of his co-workers, who in his 1901 doctoral thesis challenged the idea that the psyche had an active role, preferring the explanation that the process was simply one of recognition by means of newly-established associations. Tolochinov, another of Pavlov's co-workers, discovered in 1902 the phenomenon that would later be called *extinction*. He proposed that the phenomenon of salivation during irritation of the dogs at a distance by foodstuffs be considered a reflex at a distance, which Pavlov

then termed a *conditional reflex*. However, it was not until 1903 that Pavlov began to endorse publicly the concept of a conditional reflex rather than an interpretation of psychic secretion. Nevertheless, the experimental methodology of his animal studies largely defined the scientific approach used by early behaviourists.

This scientific approach was strongly advocated by J. B. Watson, who is often regarded as the founder of behaviourism. He argued that psychology should be a purely objective experimental branch of natural science with the goal of the prediction and control of behaviour. No dividing line should be placed between man and animal since the aim is to develop a unitary scheme of animal response. Moreover, introspection should form no essential part of its methods, nor should the scientific value of its data be dependent upon the readiness with which they lend themselves to interpretation in terms of consciousness (Watson, 1913). From these premises, Watson argued that behaviour, rather than consciousness, should be the focus of study in psychology. In large part, though, his argument was pragmatic, based on the difficulty of obtaining objective data on consciousness in an experimental setting:

> Our minds have been so warped by the fifty-odd years which have been devoted to the study of states of consciousness that we can envisage these problems only in one way. We should meet the situation squarely and say that we are not able to carry forward investigations along all of these lines by the behavior methods which are in use at the present time. (p. 174)

Furthermore, he did not completely close the door on possible future inclusion of consciousness in behaviourism:

> As our methods become better developed it will be possible to undertake investigations of more and more complex forms of behavior. Problems which are now laid aside will again become imperative, but they can be viewed as they arise from a new angle and in more concrete settings. (p. 175)

In essence then, his argument was that psychology could be advanced more productively, and scientifically, by making behaviour rather than consciousness the

focus of study. Most early advocates of behaviourism thus renounced cognition and advanced the doctrine that learning can occur only by performing responses and experiencing their effects. However, not all rejected the notion of including consciousness. For example, Tolman (1922) proposed a broad framework in which stimulating agencies and behaviour-cues, objects, and acts could be studied. This framework could be seen to include ideas, images, feelings and emotions. Several of his experiments produced evidence of a cognitive component in behaviour. In one experiment, (Tolman & Honzik, 1930), rats who had initially explored a maze without reward were able to learn more rapidly when a reward was introduced than those who had a reward from the beginning, suggesting that learning was initially happening in the absence of reinforcement. Another experiment (Tolman, Ritchie, & Kalish, 1946) suggested that rats built a *cognitive map* of the location of the food, independent of the conditioning received in accordance with the maze topology.

Nevertheless, most behaviourists rejected such a purposive component or indeed any cognitive dimension, believing instead in a purely physiological basis for behaviour. Hull (1935) argued for a formal mathematical approach on the premise that, by its nature, any truly scientific theory would necessarily converge on truth. His *mathematico-deductive* theory (Hull, et al., 1940) proposed a detailed deductive theoretical system incorporating definitions, postulates, corollaries and theorems to explain rote learning, which he then extended and redefined in his major theoretical work (Hull, 1943) on the principles of behaviour. His theory was ambitious in scope and included a prominent role for motivation, unlike other behaviouristic theories (Brogden, 1944). Despite general acclaim, several theorists pointed out theoretical weaknesses and inconsistencies in the theory. Koch (1944) identified inconsistent treatment of variables as observable antecedents, process or intervening variables, and pointed out that Hull had derived equations based on no more than three or four empirical data points. Meele (1945) noted that applying some of the equations to data given for other cases in the book produced conflicting answers, and that the equations require that conditioned responses to stimuli approach the same final state of strength and have the same growth rate,

despite much contradictory empirical evidence. However, Hull is nowadays perhaps best known for his *Drive Reduction Theory*. He believed that all motivation came originally from biological imbalances or needs, and that behaviour could be regarded as an outward expression of the organism's attempts to remedy the imbalance. He used the word drive to describe the degree of imbalance; hence drive was something the animal tried to reduce. In his terminology, an animal searched for food in order to reduce the hunger drive. He believed an animal would repeat any behaviour that reduced a drive, if the same need occurred again. He also believed that secondary non-biological needs were learnt by conditioning through association with biological needs and were thus weaker than biological needs. This was widely accepted until an experiment by Harlow (1958), which demonstrated that a baby monkey's need for comfort and affection was much stronger than the need to feed.

The most influential behaviourist of the 20th century was B.F. Skinner who introduced the term *radical behaviourism* to emphasise the explicit exclusion of any cognitive mechanism in behaviour: "What is felt or introspectively observed is not some nonphysical world of consciousness, mind, or mental life but the observer's own body" (1974, p. 17). Although he made many contributions to the field, he remained committed to a narrow interpretation of Darwin's evolutionary theory as the basis of behaviour:

> The process of operant conditioning presumably evolved when those organisms which were more sensitively affected by the consequences of their behavior were better able to adjust to the environment and survive. (Skinner, 1971, p. 120)

A broader interpretation of Darwin's theory, incorporating a cognitive perspective, did not emerge until the work of Richard Dawkins who introduced the concept of a meme (1976/1989, pp. 189-201).

The hallmarks of behaviourism, as described above, were carefully controlled and described laboratory experiments. These experiments contributed a great deal to understanding of animal behaviour and learning. However, a dark side of

behaviourism also emerged. Pavlov also carried out experimental surgical procedures for research, used electric shocks on dogs, and experimented on children. In a controversial study, (Watson & Raynor, 1920), widely known as the "Little Albert" study, a baby who was nine months old at the start of the trial was conditioned to fear a white rabbit and a rat, which he had previously accepted without fear. Although this may have added to knowledge on how fear might be learned, the impact of the experiment on the baby was disregarded. The baby was not "deconditioned" and nothing is known about whether the baby later suffered adverse effects. In a sad footnote, it appears that the baby, who it is believed was named Douglas, died at the age of six from hydrocephalus (Beck, Levinson, & Irons, 2009).

Many behaviourist experiments involved subjecting animals to what, from a 21st century perspective, is tantamount to torture. For example, in Harry Harlow's *well of despair* experiment, monkeys were separated from their mothers a few hours after birth and kept in total social isolation in a stainless-steel chamber for periods of up to 12 months (Harlow, Dodsworth, & Harlow, 1965). They reported "The effects of 6 months of total social isolation were so devastating and debilitating that we had assumed initially that 12 months of isolation would not produce any additional decrement. This assumption proved to be false; 12 months of isolation almost obliterated the animals socially" (p. 94). Moreover, despite mounting evidence of the ineffectiveness of electric shocks and other punishments from many experiments (Skinner, 1953, p. 182), many behaviourists continued to use electric shocks in their experiments. In one of the worst cases, a study on *learned helplessness* (Overmier & Seligman, 1967), dogs were immobilised with curare before being exposed to repeated electric shocks.

Inevitably, such experiments were undertaken with an *end justifies the means* rationale. From a 21st century perspective, it is difficult to look past these negative aspects of the behaviourist agenda to recognise its achievements. There were other sources of unease. Skinner argued that a theory of learning was unnecessary and that behaviourist research would progress better without theory: "Theories are fun. But it is possible that the most rapid progress toward an understanding of learning

may be made by research that is not designed to test theories" (1950, p. 215). This rejection of the scientific method may have motivated Noam Chomsky to write a scathing, but influential, critique of Skinner's *Verbal Behaviour*. Discussing Skinner's concept of the scientific method, Chomsky comments:

> To cite just one example, Skinner defines the process of confirming an assertion in science as one of "generating additional variables to increase its probability" (425), and more generally, its strength (425-29). If we take this suggestion quite literally, the degree of confirmation of a scientific assertion can be measured as a simple function of the loudness, pitch, and frequency with which it is proclaimed, and a general procedure for increasing its degree of confirmation would be, for instance, to train machine guns on large crowds of people who have been instructed to shout it. (1959, p. 54)

Moreover, Chomsky questioned the ecological validity of Skinner's extrapolation from rat and pigeon behaviour to that of humans: "This creates the illusion of a rigorous scientific theory with a very broad scope, although in fact the terms used in the description of real-life and of laboratory behaviour may be mere homonyms, with at most a vague similarity of meaning" (1959, p. 51). He continued his attacks on behaviourism: "Whatever function 'behaviorism' may have served in the past, it has become nothing more than a set of arbitrary restrictions on 'legitimate' theory construction" (1972, p. 17), and on Skinner's disregard for the scientific method: "It would be hard to conceive of a more striking failure to comprehend even the rudiments of scientific thinking" (1972, p. 46).

One clue to understanding the vehemence of the growing opposition to behaviourism, and to Skinner in particular, was concern about the underlying motives of behaviourism. Many humanists endorse Goldstein's (1934/1995) concept of self-actualisation, to realise one's potential, as an ideal goal for the betterment of humanity. *To be all that one can be* is the highest level in Maslow's (1943) hierarchy of needs, and self-actualisation occurs when a person's *ideal self* (who they would like to be) is congruent with their actual behaviour and *self-image* (Rogers, 1951). In contrast, there was a suspicion that behaviourism had an

underlying motive of the control of human behaviour. Indeed, Skinner openly advocated a benign authoritarian state in which the citizens' behaviour was controlled for their own good (1948). However, he also understood why many would find this unacceptable:

> This possibility is offensive to many people. It is opposed to a tradition of long standing which regards man as a free agent, whose behavior is the product, not of specifiable antecedent conditions, but of spontaneous inner changes of course. The alternative point of view insists upon recognizing coercive forces in human conduct which we may prefer to disregard. It challenges our aspirations, either worldly or otherworldly. Regardless of how much we stand to gain from supposing that human behavior is the proper subject matter of a science, no one who is a product of Western civilization can do so without a struggle. We simply do not want such a science. (1953, pp. 6-7).

The use of controlled laboratory animal experiments led to doubts about ecological validity, and generalisation to broader contexts. Furthermore, the scientific approach had become associated in many minds with reductionist experiments and a perceived lack of richness in results. This misunderstanding of the scientific approach can, perhaps, best be understood in terms of the much wider quantitative/qualitative divide (Caporaso, 1995); there is no reason why a scientific approach cannot be used to study consciousness, or any other abstract concept.

A more fundamental concern is that the behaviourist goal of a unitary theory of animal behaviour leads to a focus on the commonalities between all organisms. From this perspective, individual differences are seen as "nuisance" parameters which must be controlled or eliminated as far as possible. However, it is precisely these individual differences, comprising what might be broadly termed personality, that are of central interest to many psychologists.

However, perhaps the strongest objections arose from behaviourism's rejection of a cognitive component in behaviour. Having renounced cognition, advocates of behaviourism advanced the doctrine that learning can occur only by performing responses and experiencing their effects. This restricted view on how learning

happens confined the scope of research to learning from the consequences of one's acts. This view was based on extrapolation from animal experimentation involving simple responses in artificial situations, using mainly rats and pigeons: animals with quite limited symbolic capacities compared to the richness of the human use of language.

In applying findings to human learning, behavioural practitioners soon discovered that relying solely on the effects of actions is a tedious and ineffective way of teaching new behaviours (Bandura, 1978a). The acquisition process can be substantially accelerated by using instructional and observational modes that capitalize on human cognitive capabilities. The most important of these is observational learning or *modelling*. In observational learning, people develop conceptions of new behaviours by observing the performances of others. They transform what they observe into generalised conceptions that capture the essential features of the modelled patterns. These conceptions are not transformed into faultless performances on first attempt, but they enable learners to produce at least a rough approximation of the activity from the outset (Bandura, 1978a). Skills are then perfected by corrective adjustments. The performance feedback that is most informative, and achieves the greatest gains, relies on corrective modelling (Vasta, 1976). In this approach, troublesome segments of a performance are identified, and adept ways of performing them are modelled by those who are proficient at it. In some cultures, complex occupational skills are developed almost entirely through prolonged intensive observation; Nash (1967) found that apprentices stand beside proficient machine operators, but do not make any attempt to operate the machines until they believe they have understood all the necessary details. They then typically operate the machines adeptly on first trial.

Some activities are mastered more rapidly through combined cognitive and physical practice than by physical practice alone. For physical activities, such as sports or gymnastics, there is substantial evidence that *cognitive rehearsal*, in which individuals visualize themselves executing the correct action sequences, improves subsequent motor performance (Corbin, 1972). Moreover, when learning is defined in terms of acquisition of novel patterns of thought and action, socially mediated

experiences enable people to go beyond what they already know and do. Indeed, a considerable body of research confirms that virtually all learning phenomena resulting from direct experience can occur through observational experience as well (Bandura, 1978a). Modelling can also teach observers effective ways of dealing with challenging or threatening situations.

It is mainly through observational learning that children extract syntactic rules from the speech they hear around them. Once they acquire syntactic rules they can generate new sentences they have never heard (Bandura, 1989). This is clear evidence of a cognitive component to learning. Most parents will recognise the stage of language development in which a child will use a word like *knowed* rather than *known*. It is hard to see how this could be construed in any way other than that the child has developed a concept of past tense and its usual representation in English. This is strong evidence against the notion that language is learnt simply by reinforcement.

The state of research in psychology, at the time of Bandura's work, is now briefly summarised. The behaviourist school which had dominated research for most of the 20th century was losing favour with many scholars. There was growing evidence of the importance of a cognitive dimension, which was explicitly rejected by the behaviourist agenda. There was also mounting evidence of the importance of social learning mechanisms and observational learning. Finally, the fundamental goal of the behaviourist agenda, which was the search for mechanisms that were universal and applied to all animals, precluded research on any aspects that were distinctly human.

6.2. SOCIAL COGNITIVE THEORY

Bandura (1986) brought the behavioural, cognitive and social perspectives together into a coherent unified theory that he named *social cognitive theory*. The choice of the term cognitive served both to distance it from the prevalent social learning theories of the day, and to emphasize the critical role that cognition plays in people's capability to construct reality, self-regulate, encode information, and

perform behaviours. Bandura believed that “what people think, believe, and feel affects how they behave” (p. 25).

There are two central insights in this theory. The first is that people are producers, as well as products, of social systems. The behaviourist view is that “a person does not act upon the world, the world acts upon him” (Skinner, 1971, p. 211). In contrast, Bandura (1978b) uses the term *reciprocal determinism* to emphasise that people create and activate environments, as well as respond to them. Human functioning is seen as the product of a dynamic interplay of personal, behavioural, and environmental influences. This triadic relationship is depicted graphically in Figure 6.1.

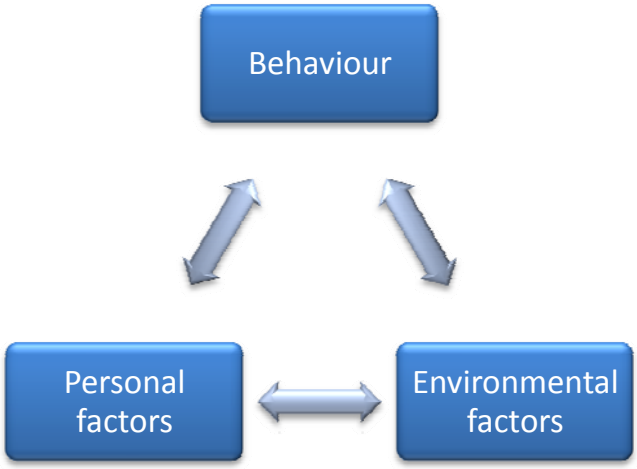


Figure 6.1: Triadic relationship between behavioural, personal and environmental factors

Bandura (1978b) used the everyday example of television viewing to illustrate this triadic reciprocal determinism:

> Personal preferences influence when and which programs, from among the available alternatives, individuals choose to watch on television. Although the potential televised environment is identical for all viewers, the actual televised environment that impinges on given individuals depends on what they select to watch. Through their viewing behavior, they partly shape the nature of the future televised environment. Because production costs and commercial requirements also determine what people are shown, the options provided in the televised environment partly shape the viewers' preferences. Here, all three

> factors—viewer preferences, viewing behavior, and televised offerings—reciprocally affect each other (p. 346).

The second insight is that people are not just reactive organisms who are shaped by environmental forces, or driven by concealed inner impulses, but are active agents. They are actively engaged in their own development and can make things happen by their actions. Bandura (2001) identifies four key elements of this perspective of agency: intentionality, forethought, self-reactiveness and self-reflectiveness. An *intention* is not simply an expectation or prediction of future actions but a proactive commitment to bring them about (p. 6). Through the exercise of *forethought*, people motivate themselves and guide their actions in anticipation of future events. They set goals for themselves, anticipate the likely consequences of prospective actions, and select and create courses of action likely to produce desired outcomes and avoid detrimental ones (p. 7). However, having adopted an intention and an action plan, one cannot simply sit back and wait for the appropriate performances to appear. Although strong interest and engrossment in activities is sparked by challenging goals, the self-regulative effectiveness of goals depends greatly on how far into the future they are projected. Distal goals set the general course of pursuits but are too far removed in time to provide effective incentives and guidance for present action. Proximal sub-goals mobilise self-influences and direct what one does in the here and now. A *self-reactive* agent monitors behaviour, and the cognitive and environmental conditions under which it occurs, and compares it with distal and proximal goals, and personal standards, thus giving shape to appropriate courses of action and motivating and regulating their execution (p. 8). Furthermore, people are not only agents of action, but self-examiners of their own functioning. The metacognitive capability to reflect upon oneself and the adequacy of one's thoughts and actions is another, distinctly human, core feature of agency. Through reflective self-consciousness, people evaluate their motivation, values, and the meaning of their life pursuits. It is at this higher level of *self-reflectiveness* that individuals address conflicts in motivational inducements and choose to act in favour of one over another (p. 10). Thus, through intentionality, forethought, self-reactiveness and self-reflectiveness, people are viewed as self-organising,

proactive, self-reflecting and self-regulating rather than as reactive organisms that are shaped and shepherded by environmental forces, or driven by concealed inner impulses.

A key concept is that, from a *phenomenological* perspective, personal meaning is derived from perceptions of reality, rather than from objective reality. Thus, it is perceived prospects that are of central psychological impact, rather than objective prospects. Likewise, perceived empowerment is important, rather than objective empowerment.

Social cognitive theory is not elaborated in full herein but selected key features have been introduced to give context for the concept of self-efficacy.

6.3. SELF-EFFICACY

Within the framework of social cognitive theory, no mechanism of personal agency is more central or pervasive than people's beliefs in their capability to exercise some measure of control over their own functioning and over environmental events (Bandura, 2006). These efficacy beliefs, which Bandura (1977) termed *perceived self-efficacy*, are the foundation of human agency. Unless people believe they can produce desired results, and forestall detrimental ones, by their actions, they have little incentive to act, or to persevere in the face of difficulties. Whatever other factors may operate as guides and motivators, they are rooted in the core belief that one has the power to produce effects by one's actions. Such beliefs influence whether people think pessimistically or optimistically and in ways that are self-enhancing or self-hindering.

Efficacy beliefs also play a central role in the self-regulation of motivation through goal challenges and outcome expectations. Efficacy expectations are distinguished from outcome expectancies: outcome expectancy is defined as a person's belief that a given behaviour will lead to certain outcomes, whereas an efficacy expectation is the conviction that one can successfully execute the behaviour required to produce the outcomes. These expectations are differentiated because individuals can believe that a particular course of action will produce certain

outcomes, but such belief does not influence their behaviour if they entertain serious doubts about whether they can perform the necessary activities (Bandura, 1977, p. 193). Moreover, perceived self-efficacy occupies a pivotal role in the causal structure of social cognitive theory because efficacy beliefs affect adaptation and change, not only in their own right, but through their impact on other determinants (Schwarzer, 1992).

It is partly on the basis of efficacy beliefs that people choose what challenges to undertake, how much effort to expend in the endeavour, how long to persevere in the face of obstacles and failures, and whether failures are motivating or demoralising. To undertake an activity, people need to want to do it, and must have the necessary skills, or the willingness to acquire them. Given appropriate skills and adequate incentives, however, efficacy expectations are a major determinant of people's choice of activities, how much effort they will expend, and how long they will sustain effort in dealing with stressful situations (Bandura, 1977, p. 194). A strong sense of coping efficacy reduces vulnerability to stress and depression in taxing situations, and strengthens resilience to adversity.

According to Bandura (1977, p. 195), efficacy expectations vary on several dimensions that have important performance implications. They differ in *magnitude.* Thus, when tasks are ordered in level of difficulty, the efficacy expectations of different individuals may be limited to the simpler tasks, extend to moderately difficult ones, or include even the most taxing performances. Efficacy expectations also differ in *generality.* Some experiences will create mastery expectations that relate only to a specific situation. Others will instil a more general sense of efficacy that extends well beyond the specific situation. In addition, expectancies vary in *strength.* Weak expectations are easily extinguishable by disconfirming experiences, whereas individuals who possess strong expectations of mastery will persevere in their coping efforts despite disconfirming experiences. Those who are socially persuaded that they possess the capabilities to master difficult situations, and are provided with provisional aids for effective action, are likely to mobilise greater effort than those who receive only the performance aids. However, to raise expectations of personal competence by persuasion, without

arranging conditions to facilitate effective performance, will most likely lead to failures that discredit the persuaders and further undermine the recipients' perceived efficacy (p. 198).

When tasks are performed in ambiguous or complex situations in which there is a variety of evocative stimuli, the informational value of the resultant arousal will depend on the meaning imposed upon it: failure in a task could be ascribed to personal, or to situational factors. People who perceive their arousal as stemming from personal inadequacies are more likely to lower their efficacy expectations than those who attribute their arousal to situational factors. Moreover, a proneness to ascribe arousal to personal deficiencies can lead to heightened attention to internal events, which can lead, in turn, to a cycle of reciprocally escalating arousal (Bandura, 1977, p. 202). A common example of this, in an educational context, is when students have to give a presentation of their work to the whole class for the first time. Even in a supportive classroom environment, the fear of doing this can be very real and debilitating for a student. It is therefore important for an educator to identify this risk ahead of time and to devise, with the student, strategies for managing the performance.

6.3.1. Effects of Self-efficacy

From the discussion above, it can be seen that self-efficacy has an effect on many facets of behaviour. There are four main aspects to this: cognitive, motivational, affective, and selection. The *cognitive* aspect is that self-efficacy influences the choices people make as they take on particular tasks. For example, it may affect the goals that people set themselves, whether they take a positive or negative view of future scenarios, and how well they function under pressure. The *motivational* aspect arises because most human motivation is generated cognitively. Self-efficacy affects how much effort someone puts into a task, and how well they sustain that effort in the face of difficulties. It influences what people think is possible, whether they attribute success or failure to their own efforts or to natural ability, and how resilient people are to failure. There is also an *affective* component. For example, anxiety, stress, arousal, and depression are subjectively determined by the

individual's perceived self-efficacy in dealing with threat, rather than directly reflective of the level of threat itself. Finally, there is a *selection* effect. Self-efficacy influences the choices people make, which, in interaction with their environment, can determine their life course beyond the moment of decision. For example, it could have an impact on a person's choice of career.

Given the wide-ranging effects of self-efficacy, the question arises as to whether a high level of self-efficacy is always good, or whether it can lead to overconfidence. Self-efficacy theory adopts a conditional view regarding negative effects of an elevated sense of personal efficacy. The functional value of high perceived self-efficacy differs between the preparatory and performance stages of a task. A distinction can also be made between two different aspects of perceived self-efficacy: belief that one has the capability to undertake a task, and belief that one has the capability to *learn* how to carry out the task. In preparing for challenging endeavours, some self-doubt about one's performance efficacy provides incentives to acquire the knowledge and skills needed to master the challenges. In the skills development phase, a high sense of learning self-efficacy serves a positive function, promoting engagement and persistence. Thus, observation of peers can enhance children's beliefs in their own efficacy for learning, which, in turn, predicts both their rate of progress during instructional sessions and their eventual level of competency (Schunk & Hanson, 1989).

Whereas, a high degree of learning self-efficacy seems always positive, the question remains as to whether an elevated level of performance self-efficacy could have a negative effect at the preparatory stage. From *perceptual control theory* (Powers, 1973), motivation is the drive to reduce discrepancies between goals and perceived achievements. From this perspective a strong elevated sense of self-efficacy should undermine the willingness to invest time in learning. Bandura and Locke, however, disagree:

> Thus, even in the preparatory phase of functioning, one need not undermine a sense of efficacy to motivate self-investment in activities, as the control theory

> under discussion would prescribe. On the contrary, instilling a strong sense of learning efficacy enhances the development of competencies. (2003, p. 96)

Nevertheless, although an elevated sense of learning self-efficacy may be positive, an elevated sense of performance self-efficacy has little value, and may indeed be negative. Consequently, establishing a realistic level of performance self-efficacy would be prudent, and seems more likely to be optimal.

6.3.2. Sources of self-efficacy

Bandura identifies four sources of self-efficacy: mastery experiences, vicarious experiences, social persuasions, and physical factors. *Mastery experiences* are defined as "the experience of overcoming obstacles through perseverant effort" (Bandura, 1997, p. 80). These are believed to be the most effective sources of increased self-efficacy (Bandura, 1977, p. 197). *Vicarious experiences* are based on modelling or observational learning. These experiences are provided by social models and entail seeing people similar to oneself succeed by sustained effort (Bandura, 1994). The effect of such modelling is strongly influenced by perceived similarity to the models. Models provide a social standard, transmit knowledge, and teach skills. *Social persuasions* can be characterized as verbal persuasion to overcome self-doubt. Negative persuasions which decrease self-efficacy are more influential than positive ones. However, the impact of social persuasion is weaker and more fragile than mastery or modelling. As Bandura notes, "In the face of distressing threats and a long history of failure in coping with them, whatever mastery expectations are induced by suggestion can be readily extinguished by disconfirming experiences" (1977, p. 198). Social persuasion is thus best used to motivate engagement in activities that can foster self-efficacy. Self-efficacy can also be affected by *physical factors*, such as somatic and emotional states. A person's perception of their physical responses (such as stress, arousal, depression, and mood) to threatening environments and situations influences their self-efficacy beliefs. From the social learning perspective, it is mainly perceived inefficacy in coping with potentially aversive events that makes them fearsome and generates arousal (Bandura, 1978a, pp. 255-256).

In summary, self-efficacy is considered to be malleable and can be affected negatively or positively by engagement in activities, and by interventions. The degree to which people raise their perceived efficacy through performance successes will depend upon the difficulty of the task, the amount of effort they had to expend, the amount of external aid they received, and the temporal pattern of their successes and failures (Bandura, 1978a, p. 252).

For the study of self-efficacy, two of these ideas merit further comment: the difficulty of the task, and the temporal variability of self-efficacy. The difficulty of the task is defined, for the purpose of this thesis, as *challenge*. It is clearly not possible to investigate self-efficacy without a conjoint analysis of challenge. The temporal variability of challenge also suggests the need to consider the use of a *time-series* in analysis.

6.4. IMPLICATIONS FOR EDUCATORS

Research has shown that self-efficacy has an impact on many areas of concern to an educator. First, it has been shown to be a major predictor of achievement. For example, in a path analysis study, Pajares (1996, p. 333) found that self-efficacy was a stronger predictor (β=.387) of success in mathematical problem solving than GPA (β=.293), or cognitive ability (β=.242); together, these three explained 51% of the variability. Self-efficacy has also been linked to the use of strategies such as self-regulated learning (Pintrich & De Groot, 1990), to motivational constructs such as persistence (Multon, Brown, & Lent, 1991), and to goals and goal setting (Schunk & Ertner, 1999).

It has also been found to be an important factor in the adjustment of students to a college environment (Brady-Amoon & Fuertes, 2011). Finney and Schraw (2003) demonstrated its connection with test anxiety. Solberg and Villarreal (1997) established its relationship to affective constructs such as stress, distress and anxiety. All of these have an impact on the retention of students and their persistence in their programme of study.

In general, individuals with higher levels of self-efficacy tend to be more motivated, use more productive learning strategies, have higher achievement, experience less stress and anxiety, be more resilient in the face of adversity, and persist in their study. It is natural, then, for educators to look for ways in which self-efficacy can be fostered and enhanced in students. At the planning stage, there are two broad areas in which educators can do this. First, the design of a course should provide opportunities for students to engage in challenging tasks that require perseverant effort for success. Second, the design should also provide opportunities for students to observe others engaged in challenging tasks. Group work provides an opportunity to do this, but the make-up of a group is important. If a group contains members of different perceived abilities, then the impact on self-efficacy for those who see themselves as having less ability could be negative whether the group succeeds in the task, or not. If the group succeeds, the success could be attributed to the contribution of more capable others. If the group is perceived as failing, the failure could be attributed to their own inability. This suggests that groups should be relatively homogeneous.

How an educator interacts with students during the course is also important. Since the impact of challenging activities on self-efficacy could be positive or negative, it is important that the success of students in challenging tasks is closely monitored, and appropriate interventions are applied if an activity becomes too challenging. This suggests that success in challenging activities should be closely monitored and appropriate scaffolding (Wood, Bruner, & Ross, 1976) put in place if success seems to be at risk. Verbal persuasion by an educator to overcome self-doubt seems a natural intervention, but the nature of this intervention is critical.

According to *attribution theory* (Weiner, 1985), ascriptions of reasons for success or failure play a key role in determining the emotional impact of the feedback and the effect on on-going motivation. These ascriptions affect a variety of common emotional experiences, including anger, gratitude, guilt, hopelessness, pity, pride, and shame. Typical perceived ascriptions share three common properties: locus, stability, and controllability. The locus identifies whether people ascribe success or failure to personal factors, situational factors, or external factors such as luck. The

perceived stability of causes also influences changes in expectancy of success. Stable factors suggest that little can be done about the causes, whereas changeable factors suggest that these can be modified. Finally, controllability refers to the extent to which a person feels they can control the factors to which they attribute the success or failure. The effect on self-efficacy will depend on the specific combination of these properties. Thus, for example, attribution of failure to insufficient ability, which is presumed to be internal, stable and uncontrollable, is likely to reduce self-efficacy beliefs, whereas attribution to luck, which is presumed to be external, unstable, and uncontrollable, would have little effect. On the other hand, a belief that a person is intrinsically lucky or unlucky (stable, uncontrollable), would lead to a reduction in self-efficacy.

Thus, the effect of success or failure on self-efficacy is complex and depends on the details of the ascription. Consequently, an educator should be sensitive to the nuances of the attribution in any intervention. This suggests that an educator should use interpersonal discussion, when intervening, in preference to any formal transmitted information. To establish the necessary climate and rapport, the guidelines of empathy, genuineness, and unconditional positive regard, from *Person Centred Therapy* (Rogers, 1957) seem as relevant today as they were over half a century ago. Viewed from the perspective of *motivational interviewing* (Miller & Rollnick, 2002), one could add the need to allow a student to develop their own strategy for change, rather than to suggest one, and to affirm rather than confront in any interactions.

In summary, self-efficacy has an impact on many areas of concern to an educator and an educator should look for ways in which it can be fostered and enhanced. The design of a course should provide opportunities for students to engage in challenging tasks that require perseverant effort for success, and should provide opportunities for students to observe others engaged in challenging tasks. Relatively homogeneous groups should be used for group work. Student engagement in challenging tasks should be closely monitored and appropriate scaffolding should be put in place, as needed. The educator should also use judicious intervention when needed. The nature of verbal persuasion and

counselling is critical and should generally allow a student to develop their own solutions. The intervention should affirm rather than confront. These points are summarised in Figure 6.2.

6.5. NEED FOR OBJECTIVE AND TIMELY EVIDENCE

With small classes, an educator may be able to spend enough time talking to each individual student to get to know the student well and form a clear impression of their perceived self-efficacy as it varies from week to week. However, many educators have larger classes, and limited time in which they can develop such a relationship with students. In such a situation, better use of the time available can be made if objective evidence is available that identifies which students are in most need of intervention. As discussed earlier, it is important that student success in challenging activities is monitored closely as an indication of whether, and when, intervention is needed.

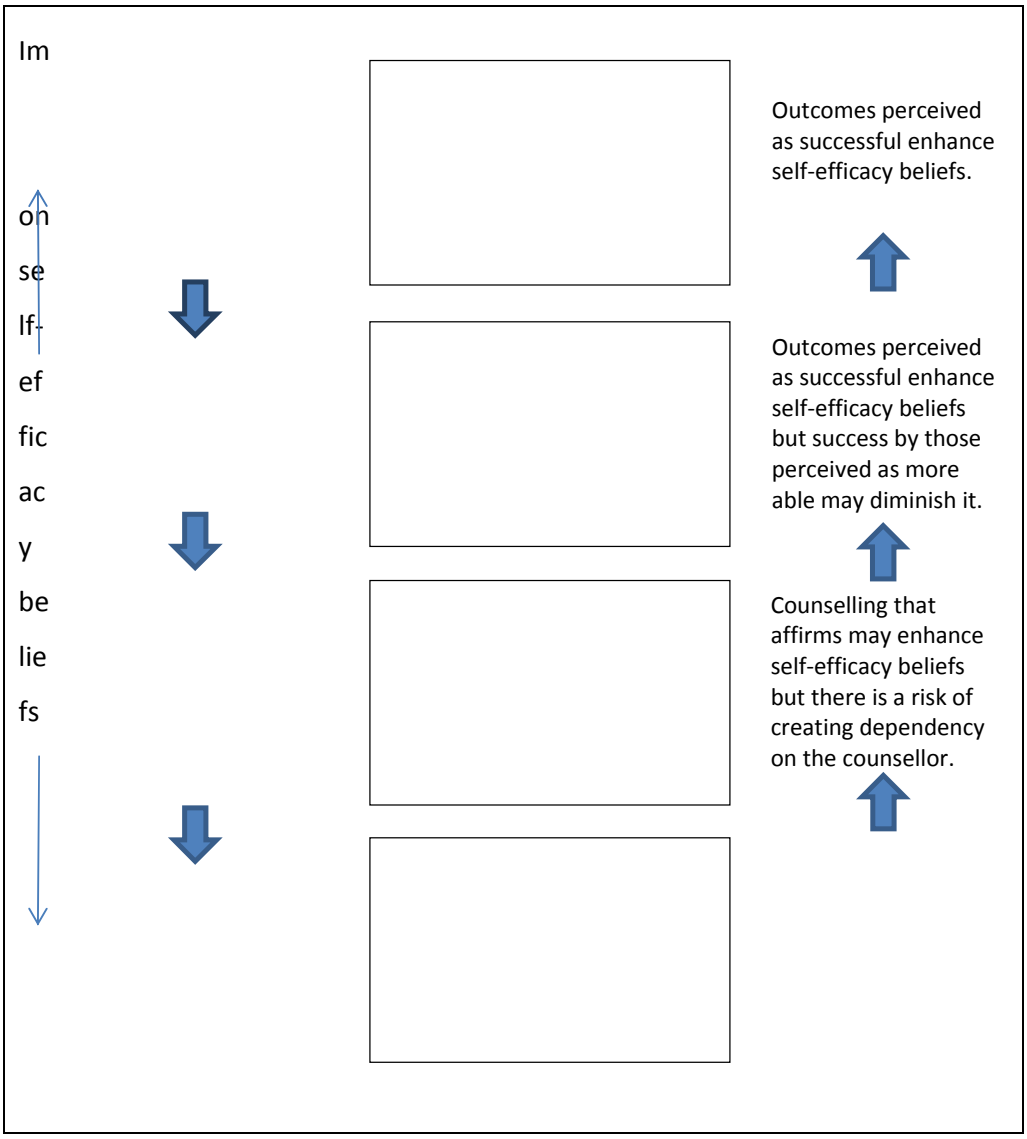


Figure 6.2: Sources of self-efficacy beliefs.

In the literature, most research on self-efficacy has involved administering and analysing a questionnaire that is designed to capture students' perceived efficacy. However, this places an additional burden on students. It is difficult to carry out such data capture frequently enough to capture the temporal variability of self-efficacy without this burden becoming excessive. This poses a quandary for an educator: choosing between an acceptable workload for students, or having sufficient information to guide them appropriately.

One solution is to look to naturally occurring data for the information. An obvious source of such data is the summative assessment data that is routinely associated

with a course. However, closer examination suggests a problem with this approach. It is important that the information is *timely*. Summative assessment normally relates to a topic, or set of topics, and occurs after completion of the associated learning activities, whereas intervention is typically required while the learning activities are progressing. This suggests that the source data need to be gathered as the activities are undertaken, rather than after the event.

In the courses studied in this thesis, such data were readily available from a software tool: *Salsa* (Lopez, 2005). This tool was used by the students to help them plan and monitor their learning progress throughout the course. In general, though, such a tool would not be available. However, as discussed later in this chapter, it might be possible to infer self-efficacy from the pattern of engagement in activities. If this were to prove possible, then reporting and analysis of self-efficacy could be produced from engagement in activities, and this, in turn, could possibly be made available automatically from an on-line learning environment such as Moodle (Dougiamas & Taylor, 2003).

6.6. MEASURING SELF-EFFICACY

Since it is a person's *belief* in their capabilities, rather than an objective measure of ability, that is central to self-efficacy, it is reasonable to base measurement on self-report data. Bandura (2006) gives several recommendations for the construction of self-efficacy scales. Many of the points made are relevant to the construction of any scale, but two are specific to self-efficacy. First, he recommends that items are specific to the context, rather than general: "Scales of perceived self-efficacy must be tailored to the particular domains of functioning that are the object of interest" (pp. 307-308). This is because behaviour "… is better predicted by people's beliefs in their capabilities to do whatever is needed to succeed than by their beliefs in only one aspect of self-efficacy relevant to the domain" (p. 310), and because general terms may leave "much ambiguity about exactly what is being measured or the level of task and situational demands that must be managed" (p. 307). The second point is that items should clearly reflect capabilities, rather than intentions: "The items should be phrased in terms of can do rather than will do" (p. 308).

Although perceived self-efficacy is a major determinant of intention, the two constructs are conceptually and empirically separable (p. 307). Thus, self-efficacy is best measured by a series of specific "can do" items denoting performance tasks that relate directly to the subject matter.

However, he also recommends that respondents use a finely graduated response format: they should: "… record the strength of their efficacy beliefs on a 100-point scale, ranging in 10-unit intervals …" (p. 312). He argues that this helps avoid extremity bias and achieves a better distribution of responses. The use of such a finely graduated scale is at odds with the approach taken in this thesis. Respondents are unlikely to be able to judge their confidence with such precision, and, thus, there is a risk of creating the illusion of greater accuracy than is really there. There is a difference here between the research perspective and a measurement perspective. With the research perspective, the usual goal is to assess the magnitude of an effect at a *collective* level. Thus large numbers of participants are often involved and statistical techniques are used, speaking colloquially, to separate the signal from the noise. With a measurement perspective, the goal is to assess an effect at the *individual* level with known accuracy. This requires careful attention to the trusted information content of each response. However, if used with due care, it also allows more efficient use of the information in a dataset and, as has been shown in Chapter Five, useful measurement can often be achieved with relatively few participants.

Thus, the approach taken in (and recommended by) this thesis is to use relatively few response categories in each item, but to use sufficient items to achieve the desired level of accuracy. With this exception, the approach used for a direct self-report of self-efficacy follows Bandura's advice.

However, a second approach, based on inference from engagement in activities, is also introduced. This approach is speculative and is evaluated in Chapter Seven. From the theory set out, the degree of perceived challenge should have a major effect on how students engage with activities. According to Bandura's theory, students with high self-efficacy are likely to start challenging activities with

confidence, and persevere with them in the face of difficulties. Conversely, those with low self-efficacy may avoid starting such activities, and may abandon them as difficulties are encountered. This suggests that it may be possible to infer the level of self-efficacy by observing the pattern of engagement in activities. However, achieving this also requires an assessment of the degree of challenge posed by the activities. Challenging activities are likely to be associated with a pattern in which a relatively smaller proportion of students readily start the activity, and those who start take longer to complete the activity. On the other hand, less challenging activities are likely to exhibit a pattern in which a larger proportion of students readily start the activity and the activity is completed more rapidly. This suggests that a measure of challenge can also be derived from observation of the same data: starting and completing activities.

With this motivation, a model can be defined for the conjoint measurement of both self-efficacy and challenge. This model is summarised in Figure 6.3.

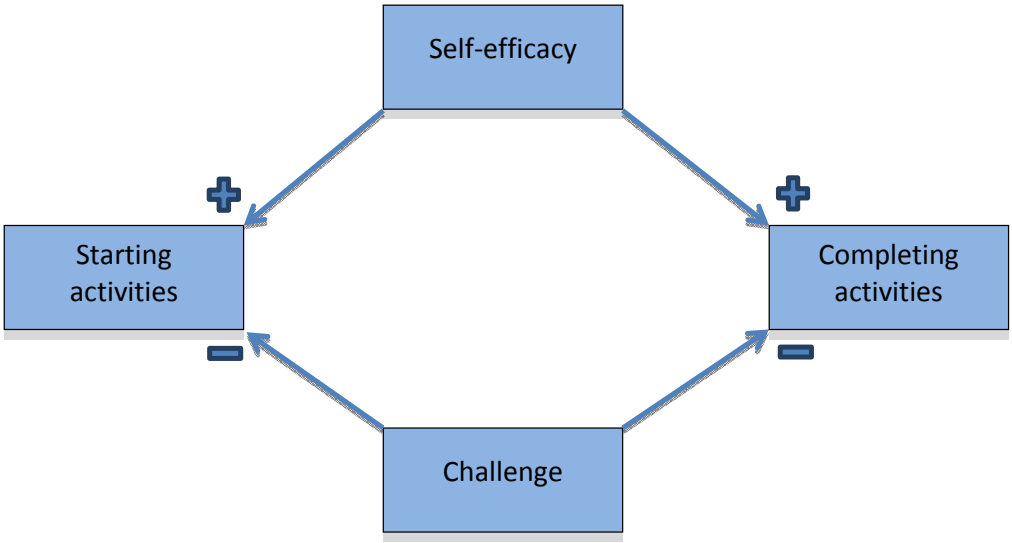


Figure 6.3: Conceptual measurement model for self-efficacy and challenge

It can be seen that the pattern for starting activities is the same as that for completing activities. Consequently, there are two approaches that could be taken. First, separate measurements could be derived from each, thus producing independent estimates of challenge and self-efficacy. Alternatively, a single scale could be used, with the start and completion of activities corresponding to separate

steps on a polytomous model. If the model holds, this latter approach should provide more accurate estimates of self-efficacy and a more informative estimate of challenge. However, it is not clear that the model should hold: the difficulty of an activity, as perceived before it is started, may differ from the perception when the activity is underway. Combining these on a single scale would mean that the first threshold is defined by the perceptions before starting and the completion threshold by the perceptions after starting. On the other hand, this confounding poses no real issue for interpretation of the output measurements.

6.7. CHAPTER SUMMARY

This chapter has presented self-efficacy, the key component of Bandura's *Social Cognitive Theory*, together with its place in that theory, and an overview of the behaviourist tradition that provides a context for understanding the impact and relevance of the theory.

The relevance to education has been briefly outlined and their implications for educators have been discussed.

This chapter has also introduced two potential approaches to the measurement of self-efficacy and challenge. The first is based on direct self-reporting of perceived capabilities. The second is based on inference from an observed pattern of engagement in activities. The next chapter describes an empirical evaluation of these two methods.

Chapter 7.
EMPIRICAL EVALUATION

The previous chapters have presented the measurement model from its theoretical foundations, through to its implementation in software, and theoretical evaluation. A review of self-efficacy, and its importance for educators, has also been presented. This chapter presents an empirical evaluation of the measurement model in the context of the measurement of self-efficacy and challenge in a classroom setting. An empirical evaluation is needed because how an instrument is used in practice, and how people interpret and use a rating scale, may differ from theoretical expectations. In particular, with self-report data, interpretation and use is likely to vary from individual to individual. This is in contrast to common assessment situations, where a single scorer is used, or where several scorers are used but there is a process of agreeing on common interpretation and use.

This chapter is organised as follows. Data for the evaluation were sourced from the Salsa software (Lopez, 2005). The first section gives a brief summary of Salsa, the relevant available data, and the mapping of these data to the self-efficacy constructs. The second section presents the key research questions addressed, and describes the methodology that was used to investigate them. The third section describes the datasets that were used for the evaluation, and the initial screening of these datasets. The fourth section presents and discusses the results of the measurement hypothesis tests on these datasets; these hypothesis tests are considered an integral part of the measurement process. The fifth section investigates the construct validity of the measures. The sixth section reviews the measurement issues that arose during the investigations and discusses practical considerations concerning use of the measurement model. The seventh section explores the possibility of using a measure of self-efficacy that is derived from patterns of engagement in activities. Finally, conclusions are drawn in the eighth section.

7.1. SALSA DATA

The basic organisational unit in the Salsa software is a course. An educator uses the software to prepare an extended electronic course descriptor, and to identify the students who may use the course in Salsa. This course descriptor includes general information about the course and syllabus, together with detailed information about the topics covered. Each topic is mapped to syllabus items, and may have other topics as prerequisites. Each topic is also associated with a set of intended learning outcomes, and a set of expected learning activities.

Students then use the web-based Salsa software to file status reports, request assistance, and monitor their learning progress. After logging in, students are presented with a list of the Salsa courses available to them. Selection of a course presents the students with an overview of the course, including progress by topic. From this overview, students can view graphical representations of their learning progress, and status reports. Students record and report their learning progress by selecting a topic and updating the status of the various learning activities and outcomes. A sample of the screen that students see for a topic is shown in Figure 7.1.

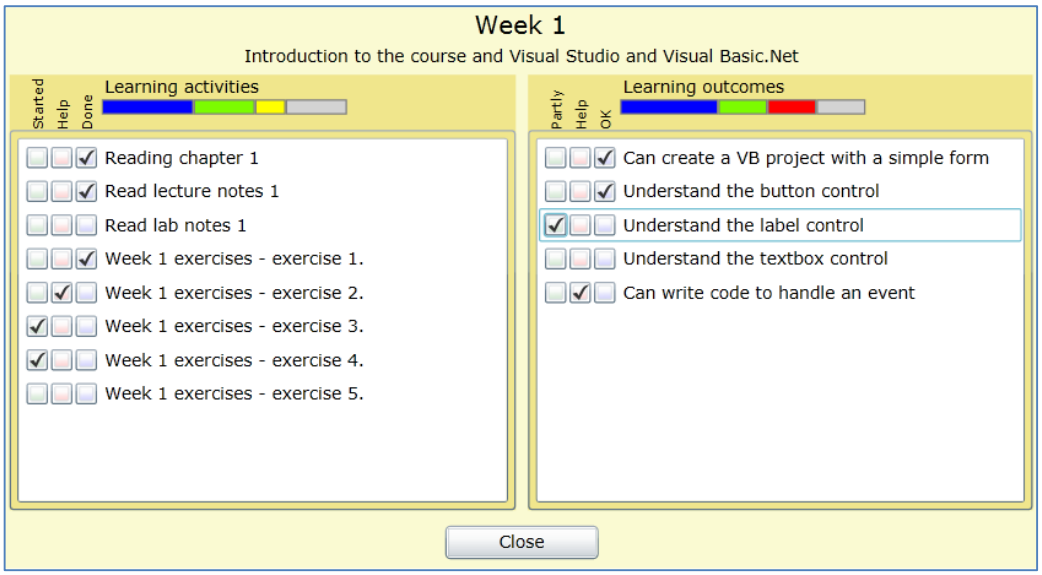


Figure 7.1: Student reporting of achievement

There are two sections in this figure. The left section shows the expected learning activities, and the right section shows the expected learning outcomes. Students record their learning status by clicking on the appropriate status box in each section. Three status boxes are used in each section.

For the expected learning activities, a student clicks the *started* box when they have started the activity, the *help* box if they have started the activity but need help to be able to complete the activity, and the *done* box if they have completed the activity. By default, the status is *not started,* and clicking on a ticked start box resets the status to this value to accommodate any error in data entry. The coloured bar at the top the section represents counts of the number of activities in each status: completed (blue), started (green), need help (red) and not started (grey).

For the intended learning outcomes, a student clicks the *partly* box if they believe they can sometimes carry out the task but are not confident they could always do it, or for knowledge based learning outcomes, if they have some understanding of the concept, but do not believe they have grasped it fully. They choose the *help* box if they believe they need help before they will be able to master the task or concept. They choose the *OK* box if they are confident they understand the concept, or can do the task. The default status is *not understood* (or cannot do). Clicking on a ticked *partly* box resets the status to this default value to accommodate any error in data entry. Because this is self-report data, the use of terms like *understand* is appropriate. This is in contrast to the normal language used for assessing learning outcomes, for which the use of vocabulary relating to unobservable factors is generally unacceptable. The coloured bar at the top the section represents counts of the number of outcomes in each status: confident (blue), partly (green), need help (red) and not understood (grey).

Tutors use the Salsa software to monitor and manage requests for help and to monitor the learning progress of students. This use is not described herein. The essential point, for the purpose of this thesis, is that the complete history of the student status reports is recorded in a database, and an extract of these data, for those students who gave research consent, was used as the source data for the

analysis. Custom software was written to manage this extraction process and to depersonalise the data by allocating pseudonyms.

The mapping of the Salsa data to the self-efficacy constructs is now described. For intended learning outcomes, the Salsa data express the confidence students have that they can carry out the specified tasks or understand the specified topics. The *partly* status and the *need help* status were combined into a single category, resulting in three levels of confidence. In terms of performance of a task, these levels can be expressed as:

- cannot do the task
- can do sometimes or partly, but is not confident one can always do the task
- confident that one can do the task

For understanding of concepts, the corresponding levels are:

- does not understand the concept
- partly understands, but is not confident one has complete understanding
- confident that one understands the concept completely

These data thus express the confidence students have in their understanding, or in their ability to carry out the specified tasks. Consequently, they map directly to Bandura's conception of self-efficacy: a three-category polytomous item can capture the level of confidence. The self-efficacy scales which are derived from these data are termed the *direct* measures of self-efficacy in this chapter.

As discussed in section 6.6, it may also be possible to develop a measure of self-efficacy that is based on inference from engagement in activities. For engagement in activities, the *started* and *need help* statuses are combined into a single category, again leading to three levels of classification:

- Not started
- Started but not completed
- Completed

As with learning outcomes, a three-category polytomous item can capture this. The scales derived from these data are termed the *imputed* measures of self-efficacy in this thesis. As discussed in section 6.6, whether, or not, such a measure is valid and useful is speculative; the investigation of this approach is described in section 7.6.

Two additional independent measures were also created by further combining the levels to form scales based on dichotomous items. The first of these combined the *started but not yet completed* and *completed* levels to create a scale based on the dichotomy not started or started; this is termed the *started* scale. The second combined the *not started* and *started* levels to create a scale based on the dichotomy not completed or completed; this is termed the *completed* scale.

7.2. RESEARCH QUESTIONS AND METHOD

The key research questions addressed in this evaluation are:

- Is the measurement model used for self-efficacy valid?
- Is the *imputed* approach to measurement a suitable proxy for *direct* measurement?
- Can useful measurement be achieved in a real classroom setting?
- What measurement issues arise in such a setting?

Validity of the measurement model is addressed by exploring the extent to which the characteristics expected from theory are evidenced in the empirical datasets. At the heart of Bandura's theory is the concept of reciprocal determinism. The practical consequence of this is that all the relevant variables are expected to be highly inter-correlated in what has been called a complex *web of causation* (MacMahon, Pugh, & Ipsen, 1960). Unpacking the causal elements of this web is challenging (Rutter, 2007); the approach taken herein is that of *path analysis*. This approach allows the investigation of the effect of a variable, after controlling for the influence of other correlated variables. However, in many cases, the proportion of variance among these covariates is expected to be larger than the residual effect, thus leading to small effect sizes. Nevertheless, the following propositions can be made from Bandura's theory.

First, mastery experiences are believed to be the strongest source of self-efficacy. It follows that task completion should be a significant predictor of reported self-efficacy, after controlling for activity starts. Using this as a covariate identifies the unique contribution of activity completion. The conceptual path analysis model for this proposition is shown in Figure 7.2.

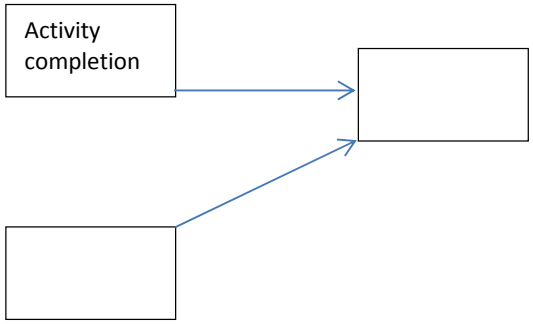


Figure 7.2: Conceptual model for source of self-efficacy

Second, persons with higher self-efficacy are more likely to persevere with tasks in the face of difficulties. Mapping this concept to activity starts and completion suggests that reported self-efficacy should explain a significant portion of activity completions, after controlling for activity starts. Again, the use of this covariate isolates the unique component of activity completion. The conceptual path model for this analysis is shown in Figure 7.3.

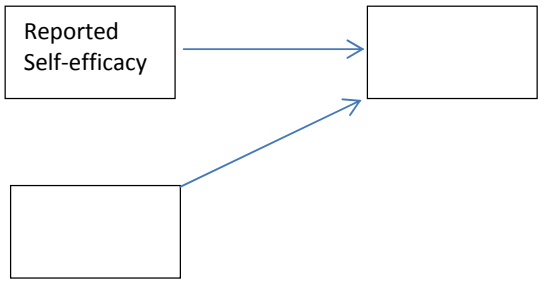


Figure 7.3: Conceptual model for perseverance in activities

Third, those with higher self-efficacy are more likely to engage in activities, whereas those with lower self-efficacy may shy away from the tasks. This suggests that reported self-efficacy should explain a significant portion of activity starts, after

controlling for activity completions. The conceptual path model for this analysis is shown in Figure 7.4.

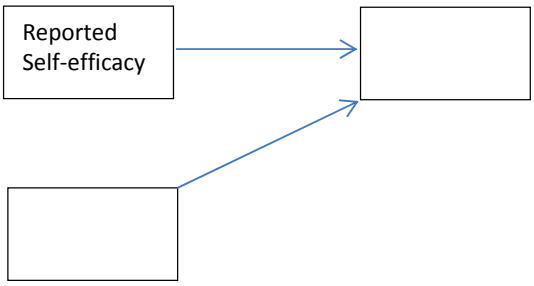


Figure 7.4: Conceptual model for engagement in activities

Taken together, these three propositions can lend support for convergent validity of the model. Further evidence of convergent validity, and evidence of discriminant validity, can be supplied by examining the temporal variation of the effects. The conceptual framework for this examination is shown in Figure 7.5.

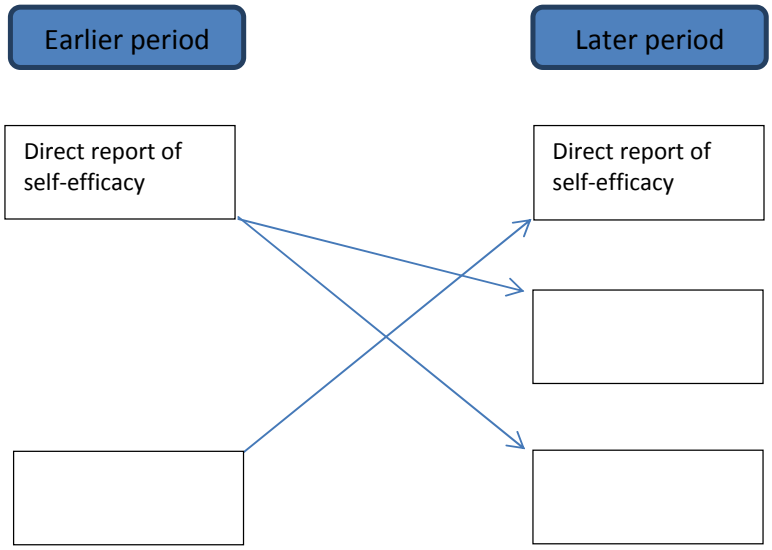


Figure 7.5: Conceptual framework for testing temporal effects.

For each of the effects mentioned above, one would expect the strength of the effect to diminish as more time passes. Thus, one would expect success in activities, as captured by the completed scale, to have a stronger effect on self-efficacy, as reported at the same time as completion, than on that reported in future reports. Similarly, one would expect a direct report of self-efficacy to have a stronger effect

on activity starts that are closer to the time of report than on later starts. Finally, one would expect a direct report of self-efficacy to have a stronger effect on perseverance and thus on activity completions that are closer to the time of report than on later completions. The above has outlined the logic of the approach to the investigation of validity; the investigation itself is presented in section 7.5.

The second research question asks whether the *imputed* approach to measurement is a suitable proxy for *direct* measurement. The conceptual approach used to investigate this is to examine the reliability of *parallel forms* of the two measures. The direct and imputed forms of the measures are compared by investigating the correlation between them. However, it is well known that correlation is attenuated in the presence of measurement error (Spearman, 1904). An estimate of the true correlation in the population (ρ) is given by $\rho = r/\sqrt{R_1 R_2}$, where R_1 and R_2 are the reliabilities of the respective measures. Consequently, the sample correlation is first calculated from the respective self-efficacy measures. The population correlation between the measures is then estimated from this by using the Person Separation Reliability estimates as the reliabilities R_1 and R_2 in the foregoing formula. However, this is likely to overestimate the population effect to some degree because these separation reliabilities are slightly conservative; the conservatism arises because, as discussed in section 3.13, reliability is not constant for an instrument, but varies across the range of an instrument. Nevertheless, it is convenient to summarise reliability in a single measure, and a conservative measure is appropriate in most situations. The consequence of this caveat is that the estimated population coefficients should be considered as upper bounds to the true population coefficients. The started and completed scales are also considered as alternative approaches to imputed measurement; a similar correlational approach is used to investigate these alternatives. The investigation of both aspects of this research question is described in section 7.6.

The third research question asks whether useful measurement can be achieved in a real classroom setting. This is addressed by examining the reliabilities achieved in practice by the various datasets used in this chapter. As discussed in section 4.3.3, useful measurement is defined by purpose, and a minimum reliability of 0.8 is

required for the purpose of triage. Consequently, this reliability level is used as a benchmark for the evaluation.

The fourth research question aims to identify practical measurement issues that may arise in a real classroom setting. The identification of these is based on a review of the issues arising from the hypothesis tests described in section 7.4.

An examination of these last two research questions is presented in section 7.7.

7.3. DATASETS USED

The self-efficacy measurement model was used in several courses over a period of two years. The three datasets chosen were taken from the last semester of the study. These reflected the outcome of approximately two years of development and refinement of the model. Overall, the model worked well. However, some issues still remained and the main criterion for selection of the datasets presented here was that they illustrated these issues well. Within a semester, the broad pattern observed was a good fit to the model at the early stages, followed by a deterioration of fit, and then an improvement of fit, leading finally to a good fit in the later stages of a course. These datasets were taken as a snapshot at week nine, a point at which the level of misfit was at its greatest.

The courses in which the model was used were first year courses, at level five in the New Zealand qualifications framework, in a three year computing degree. The courses chosen cover the three main specialisations in the degree: Computer Science (S dataset), Information Technology (H dataset), and Information Systems (D dataset). In each of the courses, the Salsa software was made available to the students as an aid to their management of their learning, but use of the software was on voluntary basis, and not all students chose to use the software. Moreover, the datasets used contain data only for those students who gave permission for their data to be used for research purposes. Accordingly, the datasets cannot be treated as representative of the level of self-efficacy in students, even within these courses.

Nevertheless, the purpose of the evaluation was to explore issues concerning the *measurement* of self-efficacy, rather than to report any inferences made about the actual level of self-efficacy found. Whereas inference to a population requires a sample representative of that population, exploration of the practical issues arising from the use of smaller convenience samples associated with a typical classroom situation requires samples representative of that situation. Thus, these datasets provide authentic data for this exploration. The summary characteristics of the datasets are shown in Table 7.1.

Table 7.1: Summary characteristics of datasets

		Outcomes	Activities	
Dataset	No of students	No of items	No of items	Context
S	43	66	38	Computer Science
H	28	38	104	Information Technology
D	23	58	162	Information Systems

Preliminary screening of the datasets was carried out using summated scales based on natural scores. The dimensionality of the *samples* is shown in Figure 7.6. This figure shows scree plots (Cattell, 1966) of a Principal Component Analysis of the *inter-subject* correlations. In each scree plot, the x-axis represents the components in decreasing order of contribution from left to right. The y-axis represents the contribution made by that component to the overall variability. The grey line shows the contribution expected under random distribution of components using a broken stick distribution.

For measurement purposes, the ideal pattern is a single dominant component. All three datasets show a single dominant component for both activities and outcomes, with residual components at noise levels. This suggests that, broadly speaking, students interpreted the scales in a similar manner. Although, analysis of scores, rather than measurements, is not optimal, these analyses are nevertheless a useful part of the screening process.

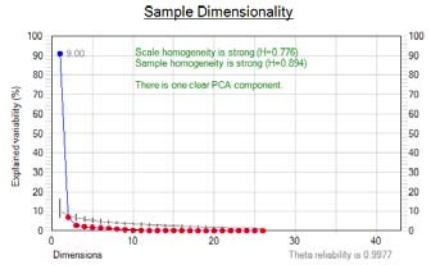
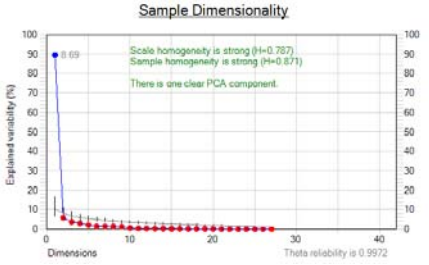
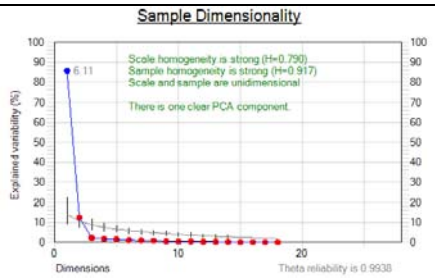
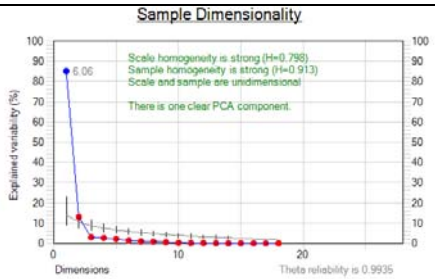
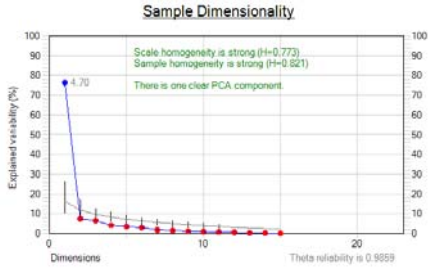
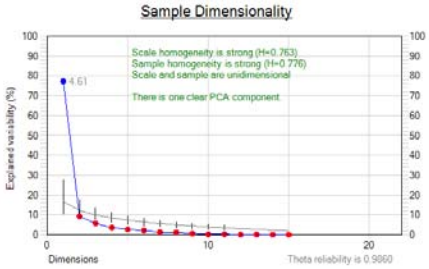
Set	Activities (imputed measure)	Outcomes (direct measure)
S		
H		
D		

Figure 7.6: Sample dimensionality of the three datasets

The dimensionality of the *instruments* is shown in Figure 7.7. This figure shows scree plots of a Principal Component Analysis of the *inter-item* correlations in the datasets. There is some evidence of multidimensionality in each of the three datasets, but the magnitude is relatively small, and there is a clear single dominant component in each case. This suggests that each of the instruments is relatively homogeneous. Moreover, all samples and instruments show *strong* scales under the criteria of Mokken and Lewis (1982, p. 422).

The presence of multidimensionality only poses a threat to validity if the scale is intended to be unidimensional. These datasets were from first-year courses which were relatively broad in content. Accordingly, some multidimensionality is to be expected. Nevertheless, for the purposes of this evaluation, it was assumed that the datasets were intended to be unidimensional.

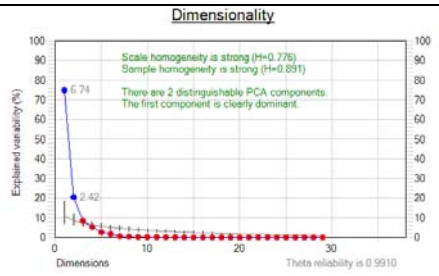
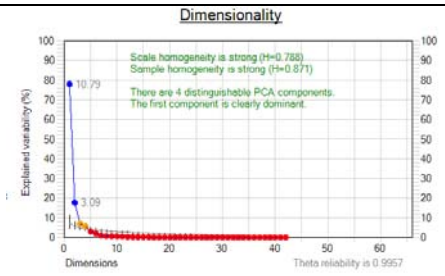
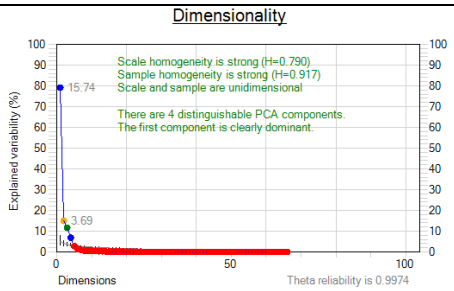
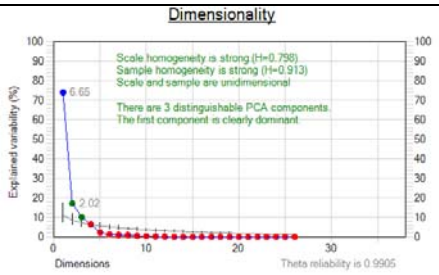
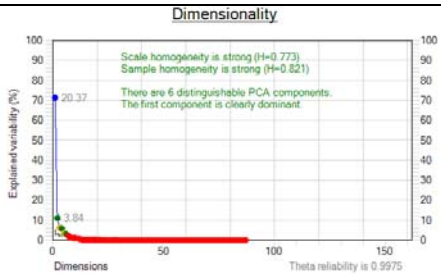
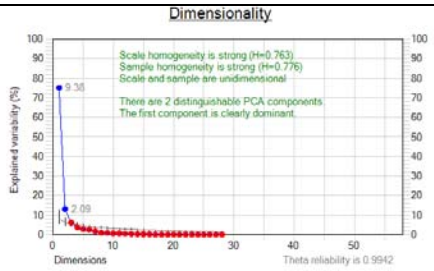
Set	Activities (imputed measure)	Outcomes (direct measure)
S		
H		
D		

Figure 7.7: Instrument dimensionality of the empirical datasets

Two conventional statistics were also calculated for screening purposes: Cronbach's alpha, and Guttman's coefficient of reliability. The connectivity of the scale was also assessed; this determines whether a common metric for subjects and items can be established. These summary statistics for the three datasets are shown in Table 7.2.

Table 7.2: Summary statistics for the three empirical datasets

	Cronbach's alpha		Reproducibility		Connectivity (cases)	
Dataset	Activities	Outcomes	Activities	Outcomes	Activities	Outcomes
S	0.941	0.974	91%	92%	81	78
H	0.982	0.957	91%	91%	101	101
D	0.994	0.986	88%	87%	83	80

The values for Cronbach's alpha suggest that the instruments are relatively homogeneous: all values are above 0.95, which corresponds to an upper limit of 5%

on possible distortion from multidimensionality, as measured by the proportion of variance explained. Moreover, the basic measurement purpose for this thesis has been defined as triage, corresponding to a conceptual reliability of 0.8. All of the values of Cronbach's alpha are well above this level, thus suggesting than the purpose of triage is achievable in these datasets. In all cases, Guttman's coefficient of reproducibility was within the confidence limits expected from theory, thus lending support to the proposition that the data are likely to fit the model. Finally, connectivity was adequate. According to Wright and Stone: "... links of as few as 10 good items will always be more than enough to supervise link validity at better than .3 logits" (1999, p. 87). The minimum connectivity of 78 cases is well above this level, and thus gives confidence that there is adequate connectivity to achieve a common metric.

Overall, all of the values found in this initial screening suggest that the datasets are suitable for measurement purposes.

7.4. HYPOTHESIS TESTS

Although the preliminary screening described above suggest the possibility of measurement, formal testing of the measurement hypotheses is also required. These tests are considered an integral part of the measurement process and provide a comprehensive framework for quality assurance, from verification of the model assumptions through to fitness for purpose. A summary of the results of these tests is given in Table 7.3.

From this table, it can be seen that all of the hypotheses were supported except for hypotheses 3, 5, 6 and 12. None of these is necessarily critical because the software has corrective options for each of these situations. However, it is important that the cause of rejection is investigated; consequently, an investigation of each of these hypothesis test failures is presented below.

Table 7.3: Results of the hypothesis tests

		Hypotheses												
Set	Method	1	2	3	4	5	6	7	8	9	10	11	12	13
S	Direct	✓	✓	×	✓	×	×	✓	✓	✓	✓	✓	×	✓
H	Direct	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
D	Direct	✓	✓	✓	✓	×	×	✓	✓	✓	✓	✓	✓	✓
S	Imputed	✓	✓	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
H	Imputed	✓	✓	✓	✓	×	×	✓	✓	✓	✓	✓	×	✓
D	Imputed	✓	✓	✓	✓	×	×	✓	✓	✓	✓	✓	×	✓

Hypothesis 3 states that the dataset is unidimensional from both sample and instrument perspectives. This was supported in the H and D datasets, but was rejected in the S datasets. In each S dataset, the sample was unidimensional, but two feasible components were identified among items. As discussed earlier, some multi-dimensionality is to be expected. Nevertheless, for the purpose of this analysis, it was assumed that unidimensionality was required. On inspection, three of the 38 items in the activity scale, and five of the 66 items in the outcomes scale, were associated with a second factor. Moreover, each of these items was associated with a single topic that had not been covered in lectures or tutorials at the time of the dataset. In this course, all resources, including lecture notes and assessments, are available to students from the outset. Thus, students were free to cover the material in any sequence they chose, subject to assessment deadlines. The identification of this topic as comprising a second factor suggests that confidence in the learning outcomes, and the pattern of engagement with these activities, was different from the other topics. Since this topic was the only topic that has not been covered in the lectures and tutorials at the time of the dataset, such a difference seems reasonable.

The software option to manage multidimensionality was chosen to investigate and quantify the effect on estimates. With this option, subjects are measured using only responses to items associated with the first item factor; item measurements were not affected because the sample was diagnosed as unidimensional. On investigation, the change in self-efficacy estimates resulting from the correction

was, in each case, less than 0.001%. Accordingly, it was concluded that the degree of multidimensionality observed had no material effect on the estimates produced.

Hypothesis 5 states that there is no response set. Significant response set was diagnosed in four of the datasets. In each of these datasets, the dominant response set found was an extreme response style (ERS), representing underuse of the central category. The central category represents activities that are started, but not yet complete, or topics that are partly understood, but not understood completely and confidently. This response set can be understood as a consequence of the way students actually used the Salsa software in practice: students tended to defer reporting their learning status until activities were completed, and concepts were understood. If they were part way through an activity, and believed they would be able to complete it within a reasonable time, they were likely to choose to spend their time on the activity, rather than reporting status. They would then update the Salsa status on completion of the activity. The effect of this was underuse of the central category, when compared to what would be expected based on their ability.

In principle, this should have no effect on the threshold estimates or associated standard errors. Likewise, the self-efficacy estimate for subjects should not be affected, but a larger standard error will be assessed. This is because both bounding thresholds contribute information for responses in the central category, whereas only a single threshold contributes information for an extreme category. With the diagnosed response set, the students' reports are successive snapshots at times when the *in-progress* aspect of learning is at a minimum. Consequently, the estimates of self-efficacy are correct. However, under-reporting of the in-progress aspect leads to less information in the dataset, and thus less accuracy in self-efficacy estimates, than would otherwise eventuate. This analysis suggests that the response set poses no threat to the validity of the measurements, but is less than ideal for measurement accuracy.

When the software option to correct for response set is chosen, subjects diagnosed with response set are eliminated from the item threshold estimation procedure. Thus, use of this option may threaten validity. In general, the option is appropriate

when relatively few cases are diagnosed, and the response set is believed to degrade measurement, but is not appropriate when there is a systematic response set across many cases. With a systematic response set, the correction option is likely to introduce a systematic bias to the estimates. Accordingly, due to the systematic nature of the response set in the analysis above, the use of the software correction was not deemed appropriate for these datasets.

However, to explore the consequences, the software option to manage response set was investigated. The data from this exploration are summarised in Table 7.4.

Table 7.4: Corrections for Response Set

		Uncorrected		Corrected		Effect	%	Correlation	
Set	Measure	PSR	ISR	PSR	ISR	Size	Cases	Case	Item
S	Direct	0.965	0.922	0.959	0.786	0.6%	23%	1.000	0.982
D	Direct	0.970	0.939	0.971	0.648	4.4%	53%	0.989	0.949
H	Imputed	0.980	0.944	0.978	0.853	2.2%	17%	0.961	0.990
D	Imputed	0.989	0.903	0.989	0.574	5.0%	70%	0.979	0.915

This table shows the person (PSR) and item (ISR) separation reliabilities with and without the correction, the response set effect size, the proportion of cases diagnosed with response set, and the correlation between the case and item estimates, with and without the software correction. As expected, there was a small impact on the accuracy of subject estimates, but a major impact on item estimates, with the item separation reliability falling below the minimum level for triage on three of the four datasets. This reliability can be seen to fall rapidly with the proportion of cases diagnosed with response set. The item correlations show a similar deteriorating pattern with increasing proportion of diagnosed cases. This is expected because these cases are ignored in the estimation process, and therefore contribute no information to the item estimates. Overall, the effect size of response set was relatively small, but response set affected a large proportion of the cases.

In summary, this investigation suggested that the use of the response set correction was not appropriate for these datasets. The analysis also supports the idea that the

response set correction option should be used judiciously, and should be avoided whenever there is suspicion of a systematic response set.

Hypothesis 6 is that there is no local dependence among items or subjects. This hypothesis was rejected for each of the datasets. In each dataset, some dependence was diagnosed both among subjects, and among items. For items, the pattern found was that the items associated with a topic formed a cluster, which was diagnosed with mutual local dependence. This dependence can be understood in terms of a *topic effect*, analogous to a booklet effect, or to the use of a scenario for a set of items. In the Salsa software, both expected learning activities, and intended learning outcomes, were grouped by topic. When students filed reports, they had a natural inclination to defer filing a report until they had completed all the activities for a topic. The effect of this was that, for many students, the items tended to move as a group from a status of *not started*, to *completed*, thus creating an artificial local dependency. This analysis suggests that using the software option to manage dependency might not be required in this case.

Local dependence among subjects occurs when the responses made by a subject are associated in some way with the responses made by another subject. This commonly occurs when students study together, and thus share misconceptions or the same pattern of domain knowledge. The main effect of this is to cause the software to understate the estimated standard error of the item threshold difficulty estimates.

The software option can correct this understatement by estimating the proportion of common information, and then making a corresponding adjustment to the information weight used in item threshold difficulty estimation. From theory, local dependence among items should affect the accuracy of the person estimates, and local dependence among subjects should affect the accuracy of the item estimates. Neither should have a major effect on the mean value of the actual estimates. The software management option was used, for both items and subjects, to quantify the consequences of the diagnosed local dependence. The results are shown in Table 7.5.

Table 7.5: Effects of software management of local dependency

		Before		After		Correlation	
Set	Method	PSR	ISR	PSR	ISR	Person	Item
S	Direct	0.965	0.922	0.962	0.860	0.984	0.992
H	Direct	0.950	0.946	0.938	0.859	0.946	0.970
D	Direct	0.970	0.939	0.971	0.879	0.991	0.983
S	Imputed	0.924	0.919	0.920	0.884	0.958	0.988
H	Imputed	0.980	0.944	0.976	0.825	0.950	0.982
D	Imputed	0.989	0.903	0.988	0.830	0.984	0.991

From the data in Table 7.5, it can be seen that the effect on person separation reliability, and the corresponding ability estimates, is minor. There is a larger effect on the reported accuracy of item estimates, but the effect on the corresponding difficulty estimates is relatively minor. In general, because local dependence is the central assumption of the model, it is appropriate to apply the software correction whenever local dependence is diagnosed. However, due to the systematic nature of the dependence, whether to apply the correction is less clear for the studied datasets. Ultimately, the conservative decision was taken to use the correction for both subjects and items. This is appropriate because a key aim of this evaluation is to explore the limits of the measurement model in a practical setting. Accordingly it was considered better to underestimate reliability, than to risk overstating the level of reliability that could be realistically achieved.

Hypothesis 12 states that the data fit the model adequately. This was rejected in each of the datasets. Lack of fit need not be a problem because the software automatically adjusts the reported standard errors to reflect the degree of misfit. Nevertheless, a detailed investigation of the cause was carried out. A micro-analysis of each dataset determined that in each dataset the cause of the rejection was that an exact binomial test determined that the proportion of cases which failed the chi-squared goodness of fit test at the $p < .05$ level was significantly more than the 5% that would be expected if the response pattern fitted the model. This failure of the goodness of fit tests was traced, in turn, to underuse of the central category. This underuse has been described earlier under the analysis of response set and is not elaborated further here. However, it can be noted that the software responds to

failure of model fit by adjusting the reported standard error. From the analysis of response fit, this adjustment is likely to overestimate the reported error, and thus understate reliability. However, as discussed earlier, the aim of this evaluation is to explore the limits of the measurement model in a practical setting. Consequently, a conservative estimate of the reliability that can be achieved is appropriate.

The results of these hypothesis tests are now summarised. The multidimensionality detected in the S dataset can be considered an anomaly. When a scale is intended to be unidimensional, the recommendation would be to delete the items associated with the second factor from the scale, or to apply the software option for automatic management. In this case, however, the software correction made no noticeable difference to the estimates, and was therefore not applied. The three remaining hypothesis rejections can all be traced to the status reporting patterns used by students. In each case the corrective options were explored. However, more detailed investigation suggested that, except for local dependence, no corrections were needed; the conclusion was that the diagnostics were, in essence, artefacts of the student usage pattern. The broader implications of this are discussed further in section 7.7. For the diagnosed local dependence, the software option was chosen, primarily because this was a more conservative decision.

For the analysis in rest of this chapter, the datasets were used without any information corrections, apart from the correction for local dependence.

7.5. VALIDITY

Three conceptual path analysis models were presented in section 7.2 to address the question of whether the measurement model used for self-efficacy is valid:

- Effect of activity completion on self-efficacy
- Effect of self-efficacy on perseverance with activities
- Effect of self-efficacy on engagement in activities

In addition, a conceptual framework was presented to explore the temporal effects of each of the above. The H dataset was used to examine the effects. The three path analysis models are discussed first.

The first path analysis model tests the proposition that activity completions have a stronger effect on directly reported self-efficacy than activity starts. The associated path diagram and regression plot are shown in Figure 7.8.

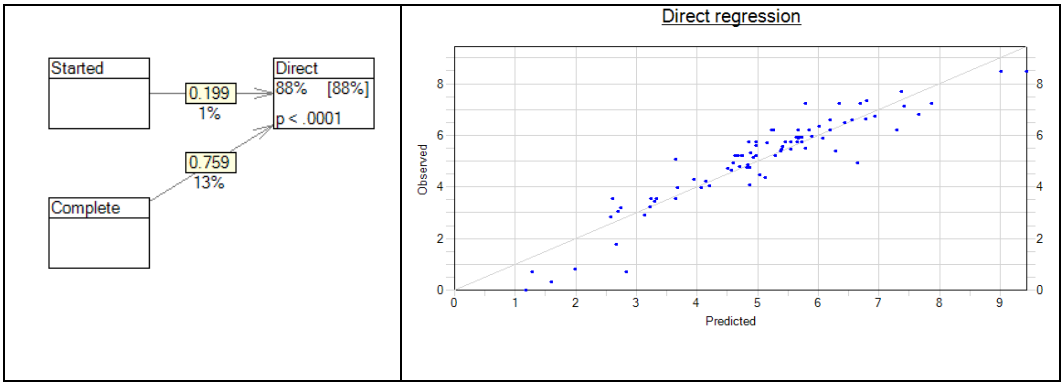


Figure 7.8: Effect of activity completions on perceived self-efficacy

The corresponding regression statistics are shown in Table 7.6

Table 7.6: Regression statistics for sources of self-efficacy

Predictor	Beta	S.E.	t	Sig.	Tolerance	VIF	r_0	r_p	r_{sp}
Started	0.199	0.074	2.70	0.008	0.224	4.468	0.867	0.263	0.094
Complete	0.759	0.074	10.29	0.000	0.224	4.468	0.934	0.721	0.359

This table shows, for each predictor, the estimated path weight (Beta) and its associated standard error, the t-statistic for the contribution made by the predictor, its associated significance level, the variance inflation factor (VIF) which measures common variance among the predictors, and tolerance (the inverse of VIF). Finally, three correlations are shown: the zero-order correlation (r_0), the partial correlation (r_p) and the semi-partial correlation (r_{sp}). This last indicates the unique variance explained by the predictor. The high variance inflation factor (VIF) of 4.468, and the falling pattern among the correlations are indicative of high multicollinearity, as expected from Bandura's theory of reciprocal determinism. With a Beta weight of 0.759, the path from complete to direct is clearly stronger that the path from started (β = 0.199). It explains 14.5 times as much variance ($0.759^2/0.199^2$). This broadly supports the proposition.

The second path analysis model tests the proposition that self-efficacy, as measured by the direct report, has a significant effect on perseverance, as measured by activity completions, after controlling for activity starts. The corresponding path diagram and regression plot are shown in Figure 7.9.

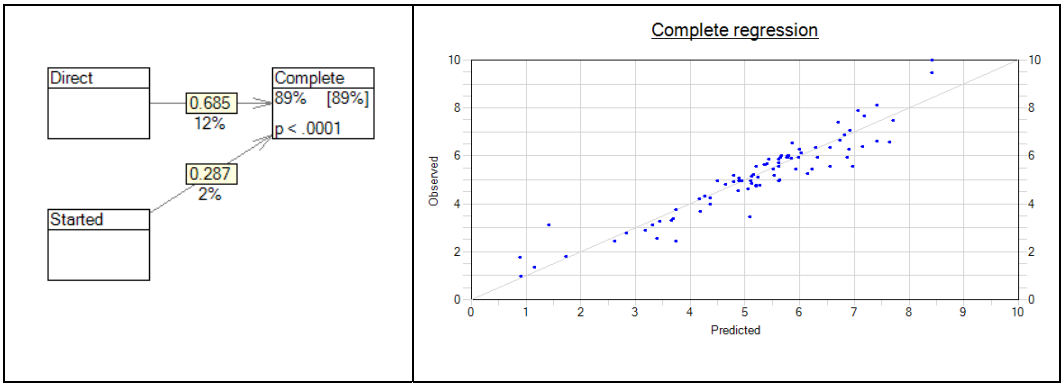


Figure 7.9: Effect of self-efficacy on perseverance

The corresponding regression statistics are shown in Table 7.7.

Table 7.7: Regression statistics for effect of self-efficacy on persistence

Predictor	Beta	S.E.	t	Sig.	Tolerance	VIF	r_0	r_p	r_{sp}
Direct	0.685	0.067	10.29	0.000	0.248	4.034	0.934	0.721	0.341
Started	0.287	0.067	4.31	0.000	0.248	4.034	0.881	0.400	0.143

Again, the high VIF, and falling pattern of correlations, suggest the high proportion of common variance expected from theory. The path analysis clearly supports the proposition that reported self-efficacy is a strong predictor (β = 0.685) of persistence, as measured by activity completions, controlled for activity starts. The unique portion of variance explained is approximately 12%. Again, this broadly supports the proposition.

The third path analysis model tests the proposition that self-efficacy, as measured by the direct report, has a significant effect on starting activities, after controlling for completions. The associated path diagram and regression plot are shown in Figure 7.10.

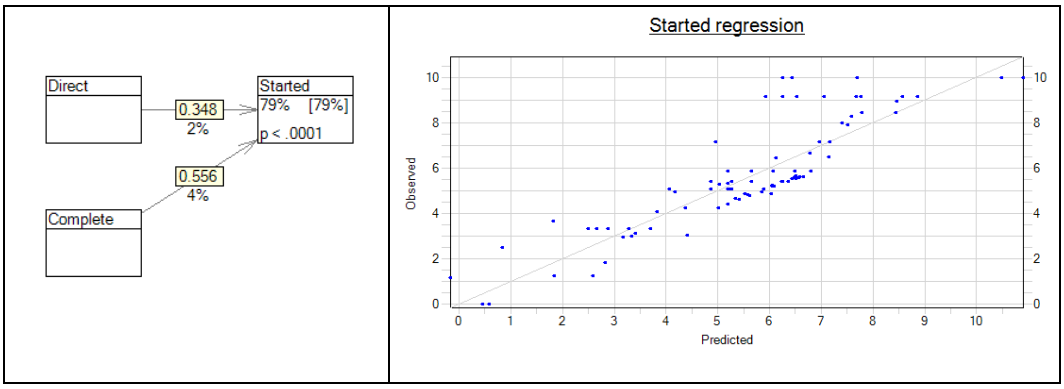


Figure 7.10: Self-efficacy as a predictor of activity starts

A ceiling effect is visible in the regression plot shown in the right of this figure: five subjects had started all activities and were thus estimated at the top of the scale; this ceiling effect may result in underestimation of the effect size. The statistics for the regression are shown in Table 7.8.

Table 7.8: Effect of reported self-efficacy on activity starts

Predictor	Beta	S.E.	t	Sig.	Tolerance	VIF	r_0	r_p	r_{sp}
Direct	0.348	0.129	2.70	0.008	0.128	7.815	0.867	0.263	0.125
Complete	0.556	0.129	4.31	0.000	0.128	7.815	0.881	0.400	0.199

Again, the high variance inflation factor, and falling pattern among the correlations, suggests the expected common variance. The unique contribution (β = .348, r_{sp} = 0.125), although modest, is statistically significant ($p < .01$). This suggests that the proposition is plausible.

Finally three propositions were made about the temporal effects. First, one would expect self-efficacy to have stronger effect on activity starts in the current week, than on starts in later weeks.

Table 7.9 shows the correlations between the direct measure of self-efficacy in each week and activity starts in the current week, the following week, and two weeks ahead.

Table 7.9: Temporal effect of self-efficacy on activity starts

Week:	1	2	3	4	5	6	7	Average
Current week	0.880	0.863	0.847	0.652	0.717	0.674	0.730	0.766
Following week	0.709	0.774	0.273	0.463	0.609	0.647	-	0.579
Two weeks ahead	0.496	0.357	0.450	0.507	0.441	-	-	0.450

In each of the first six weeks, the effect on starts in the current week is larger than in the following week. The probability of this occurring by chance is 2^{-6} ($p \cong 0.016$) and thus this effect is significant at the $p < .05$ level. The difference between the following week and two weeks ahead is not statistically significant, but the falling pattern shown by the averages suggests that the effect diminishes with time. Moreover, the strength of the association in each week is greater that the association with the self-efficacy report in the previous week: this is also statistically significant at the $p < .05$ level.

Second, one would expect self-efficacy to have stronger effect on perseverance and thus on activity completions in the current week, than on completions in later weeks. Table 7.10 shows the correlations between the direct measure of self-efficacy in each week, and activity completions in the current week, the following week, and two weeks ahead.

Table 7.10: Temporal effect of self-efficacy on completions

Week:	1	2	3	4	5	6	7	Average
Current week	0.915	0.909	0.801	0.831	0.905	0.834	0.904	0.871
Following week	0.773	0.791	0.236	0.682	0.809	0.672	-	0.661
Two weeks later	0.641	0.015	0.318	0.645	0.661	-	-	0.456

Again, the same pattern is observed: in each of the first six weeks, the effect on starts in the current week is larger than in the following week ($p < .05$), the difference between the following week and two weeks ahead is not statistically significant, and the falling pattern of the averages suggests that the effect diminishes with time.

Third, one would expect success in activities, as captured by the completed scale, to have a stronger effect on self-efficacy, as reported at the same time as completion, than that reported in future reports. Table 7.11 shows the correlation between activity completions in each week, and reported self-efficacy in that week, the following week, and two weeks later.

Table 7.11: Temporal effect of activity completions on self-efficacy

Week:	1	2	3	4	5	6	7	Average
Current week	0.915	0.909	0.801	0.831	0.905	0.834	0.904	0.871
Following week	0.552	0.767	0.285	0.804	0.777	0.797	-	0.664
Two weeks later	0.496	0.357	0.450	0.507	0.441	-	-	0.554

Yet again, the same pattern is observed: in each of the first six weeks, the effect on starts in the current week is larger than in the following week ($p < .05$), the difference between the following week and two weeks ahead is not statistically significant, and the falling pattern of the averages suggests that the effect diminishes with time.

All of these results are consistent with self-efficacy theory. The proposition that activity completions have a stronger effect on directly reported self-efficacy than activity starts supports the idea that mastery experiences are the strongest source of self-efficacy. The proposition that self-efficacy has an effect on activity completions, after controlling for activity starts, supports the idea that those with higher self-efficacy have greater perseverance. The proposition that self-efficacy has a significant effect on starting activities, after controlling for completions, supports the idea that those with higher self-efficacy are more likely to engage in activities. Moreover, all of these effects diminish over time and stronger associations are found between theoretical causes and effects that are close in time, than those further removed.

All of these findings lend support to the construct validity of the implemented measurement model.

7.6. IMPUTED MEASUREMENT AS PROXY

This section addresses the question of whether the imputed approach to measurement is a suitable proxy for direct measurement. Table 7.12 shows the data used for this discussion.

Table 7.12: Comparison of direct and imputed approaches

	Correlation	Person Separation Reliability		Correlation
Dataset	r (sample)	Direct	Imputed	ρ (population)
S	0.926	0.958	0.917	0.99
H	0.951	0.938	0.976	0.99
D	0.947	0.971	0.988	0.97

In this table, for each of the datasets, the first column shows the observed Pearson correlation between the two self-efficacy estimates under the different approaches. The next two columns show the reliability achieved, as an indication of the level of measurement error. An estimate of the true population correlation is shown in the final column.

From these data, it can be seen that there is a close accord between the two measures. The measures are, however, conceptually distinct. The direct measure estimates a subject's self-efficacy at the time of the report, whereas the imputed measure infers the self-efficacy estimate from the historical pattern of engagement. In essence, then, the imputed measure is *looking back*, and the direct estimate is a *current snapshot*. On the other hand, since mastery experiences are expected to be the major source of efficacy expectations, a person's current self-efficacy is likely to be largely determined by their own subjective assessment of their historical pattern of engagement. Nevertheless, any difference between the measures is an additional source of variation, which will necessarily affect the precision of measurement. This, in turn, will impact the reliability and thus fitness for purpose.

This difference can be quantified by considering the standardised variance. Let σ_A^2 denote the component attributable to the conceptual difference between the two measures. Let σ_B^2 denote the component attributable to measurement error.

Assuming these are uncorrelated, the resultant reliability achieved is given by $R_{obs} = 1 - \sigma_A^2 - \sigma_B^2$. The effect on reliability can thus be estimated as:

$$\mathrm{R}_{obs} = 1 - (1 - \rho^2) - (1 - R_{meas}) \quad (7.1)$$

$$\mathrm{R}_{obs} = \rho^2 + R_{meas} - 1$$

With an estimated population coefficient of 0.97 (i.e. $\rho^2 \approx .94$), this leads to

$$\mathrm{R}_{obs} = R_{meas} - 0.06 \quad (7.2)$$

The consequence of this is that a higher level of measurement reliability, and thus items or cases, is needed to achieve any given fitness for purpose profile. The required reliability (R_{obs}), the corresponding additional number of cases or items, and that number expressed as a percentage, are shown in Table 7.13.

Table 7.13: Increased reliability required for imputed measurement

Purpose	Strata	R_{fit}	R_{imp}	Extra	Prop
No separation required	0	0.000	0.000	0	0%
Minimal separation	2	0.610	0.616	0	2%
Triage	3	0.800	0.806	1	3%
Basic grading or classification	5	0.925	0.931	5	9%
Detailed grading or classification	10	0.981	0.987	97	46%

It can be seen that this has negligible impact (21 rather than 20 items) for the purpose of triage, or less; a larger but manageable impact (58 rather than 53 items) for the purpose of basic classification; and a major impact (308 rather than 211 items) for detailed classification of 10 strata. This suggests that, where detailed classification is required, the *direct* measure should be used. However, when the purpose is triage, or basic classification, the *imputed* approach may be viable.

Finally, the option of using the started and completed scales was investigated, as alternative methods of imputed measurement. The results are shown in Table 7.14.

Table 7.14: Comparison of three methods of imputation (H dataset)

		Reliability		Correlation matrix		
	Basis	Person	Item	1	2	3
1	Combined	0.977	0.863	1.000	**0.99**	**1.00**
2	Started only	0.978	0.868	0.967	1.000	**0.95**
3	Completion only	0.969	0.849	0.977	0.929	1.000

Note: In the correlation matrix, the figures below the diagonal represent the Pearson coefficients observed in the sample. The figures above the diagonal represent the corresponding estimated population correlations calculated by dis-attenuating the sample figures using the person reliability.

In the top row of the correlation matrix, the estimated population coefficients are close to unity. This suggests that all three approaches could be used more or less interchangeably. However, there was a reduced benefit in these datasets from using three-category items because of the systematic underuse of the central category reported earlier. If the data could be sourced in a way that avoided this underuse, three-category should produce noticeably better accuracy. Furthermore, the resulting increased accuracy should offset the error introduced by using the imputed measure as a proxy.

In summary, the imputed measure should be considered as a viable alternative to direct measurement and three category items are appropriate if data can be sourced in a way that avoided the underuse of the central category.

7.7. MEASUREMENT ISSUES

This section addresses practical issues relating to the use of the measurement model in a real setting: whether useful measurement can be achieved in a real classroom setting, and the measurement issues that may arise in such a setting.

A benchmark reliability of 0.8 was set to investigate whether useful measurement can be achieved in a real classroom setting. This corresponds to a purpose of triage, which was judged the practical minimum for useful measurement. All of the datasets studied produced reliabilities well above this level, which suggests that useful measurement is achievable. Moreover, the study was carried out in a

research context: consequently only data from students who had given consent to use their data for research were used. This constraint would not apply in normal educational use, and the additional data could be expected to increase reliability. However, no model can be expected to produce good measurements from poor data. The datasets studied reflected the outcome of a period of continuous improvement in which items that performed poorly were culled, and a better understanding of appropriate use of the model was developed. In particular, the use of simple and consistent language, such as "can do *specific task"* and "understand *specific topic or concept*", was important for scale cohesiveness. The diagnostics produced by the hypothesis tests were particularly helpful throughout this process.

However, some issues still remained at the end of the process, as illustrated by the discussion of the hypothesis tests in section 7.4. In particular, there was a broad tendency for students to defer reporting their learning status in Salsa until they believed they had "something worth reporting". This was discussed openly with the students, who believed this was a good way of using the tool. It is important to note that use of the Salsa tool was voluntary, and students were encouraged to use the tool in whatever way they believed aided their learning. Accordingly, it was not believed appropriate to suggest that the students change their approach to this. However, a number of measurement issues arose as a consequence of this usage.

One aspect found was that students tended to defer reporting their learning status until activities were completed. The effect of this was that there was underuse of the central category compared to what would be expected, based on their ability. The central category represents activities that are started, but not yet complete, or topics that are partly understood, but not understood completely and confidently. This pattern was diagnosed by the response set hypothesis. This did not affect the validity of the estimates, but the impact was that there was less information in the dataset, and consequently less accuracy, than would otherwise be obtained. This was also the cause of a diagnosis of model misfit.

Another aspect found was that students tended to work on a set of topics as a whole, and only report status when all the activities associated with a topic were completed, thus creating the topic effect discussed earlier. The effect of this was to create an illusion of greater local dependency among the items that was really there.

These observations serve to underline both the importance of the integrated hypothesis tests and the need for investigation of the causes and judgement in the interpretation of any findings. The judgement made for the datasets studied was that it was better to accept less accuracy in measurements rather than to try to influence students to change their preferred way of working.

7.8. CHAPTER SUMMARY

The hypothesis tests and measurement model can give assurance that something is being measured systematically, but not what that something is (Loevinger J. , 1957). This chapter has collected evidence that what is being measured is self-efficacy. Establishing the validity of a construct from observed characteristics of empirical data necessarily involves inductive logic. This logic is captured by the everyday phrase, of unknown origin, but usually attributed to the poet James Whitcomb Riley (1849–1916): "when I see a bird that walks like a duck and swims like a duck and quacks like a duck, I call that bird a duck". This phrase highlights both the essence of the inductive reasoning involved, and its shortcomings: the only conclusion that can be drawn from this reasoning is that the inference is *plausible*, not that it is necessarily correct.

The evidence presented in this chapter is consistent with the proposition that self-efficacy is being measured. It shows the expected relationships with activity engagement, perseverance, and mastery experiences. Moreover, the temporal effects are as expected. On this basis, it is reasonable to conclude, subject to the caveat of the limitations of inductive reasoning, that it is self-efficacy that is being measured.

The evaluation has also shown that useful measurement of self-efficacy and challenge is achievable in a real classroom setting. However, it was also found that how the students actually used the Salsa tool differed from initial expectations. The hypothesis tests and framework were particularly helpful in diagnosing and analysing this.

The evaluation has also suggested that self-efficacy could be usefully estimated by inference from the pattern of engagement in activities.

In summary, it is believed, not only that measurement of self-efficacy is possible, but that useful measurement can be achieved in realistic educational settings with the model and techniques presented herein.

Chapter 8.
CONCLUSIONS

This thesis has investigated the possibility of providing educators with objective evidence of students' self-efficacy, and the perceived challenge of the learning activities in a course. Both theoretical and practical perspectives have been considered.

This chapter is organised as follows. The first section presents a brief review of each of the chapters in this thesis. The second section summarises the main findings. The third section highlights the main contributions of the work. The fourth section discusses the limitations of the work and the extent to which the findings might be generalised. The fifth section presents a number of areas where further work would be useful. Finally, conclusions are drawn in the sixth section.

8.1. REVIEW

Chapter Two presented the measurement model, together with its theoretical and conceptual underpinnings. When the assumptions hold, the model produces linear measurement, accompanied by estimates of uncertainty, conjointly for subjects' abilities and item difficulties. These are placed on a common metric, which further allows meaning of subject ability measures to be constructed by reference to item difficulties, and vice-versa. Dichotomous and polytomous items may be freely mixed in the scale and share a common interpretation. The model is generic and applies equally to the measurement of latent or manifest constructs.

Chapter Three presented the perspective of measurement as hypothesis. No model can produce correct measurements from poor data. The quality of outputs depends both on the adequacy and quality of the input data, and on the correctness of the measurement theory and instrument; each measurement exercise is thus a working hypothesis which one actively seeks to disconfirm. A set of 13 hypothesis tests was presented, covering all the assumptions of the model from the theoretical

quantifiability of the constructs, through to fitness for purpose. Taken together, these provide a comprehensive framework for evaluation of the success of the measurement exercise.

Chapter Four described the proof of concept software that was implemented in connection with this thesis. Special attention was paid to the human factors that enable the software to be usable and interpretable by educators in a typical setting. The key elements were the use of multiple output representations, the use of appropriate analogies and metaphors, the mapping of statistical and information theoretic terms and constructs to equivalents that are likely to be more familiar to educators, and the reimagining of reliability as a measure of fitness for purpose. Reference datasets were used to provide basic evidence of convergent and discriminant validity.

Chapter Five presented a theoretical evaluation of the model. Systematic testing of each aspect of the model was carried out by simulation to evaluate accordance with theoretical expectations, robustness to data that do not fit the model, and to derive practical benchmarks that inform use. The close agreement between the results of the simulations and theoretical expectations strongly supports the essential correctness of the information theoretic approach taken in this thesis, and in particular, the contribution made to measurement theory by the theorems in section 5.7. The model was robust under the introduction of random noise and violation of the core assumption of local independence. Relatively few subjects and items were required to achieve useful measurement accuracy, which supports the proposition that useful measurement can be achieved in realistic educational settings.

Chapter Six presented self-efficacy, the central component of Bandura's *Social Cognitive Theory*. It briefly outlined the relevance of self-efficacy to education and discussed some of its implications for educators. Two potential approaches to the measurement of self-efficacy and challenge were introduced: the first based on direct self-reporting of perceived capabilities, and the second based on inference from an observed pattern of engagement in activities.

Chapter Seven presented an empirical evaluation of the software. Evidence was presented for the construct validity of self-efficacy, as measured in this thesis. The evaluation also established that useful measurement of self-efficacy and challenge is achievable in a real classroom setting. It was found that the actual student usage pattern differed from initial expectations, and the hypothesis tests and framework were helpful in diagnosing and analysing this pattern. The evaluation also suggested that self-efficacy could be usefully estimated by inference from the pattern of engagement in activities.

8.2. MAIN FINDINGS

Three broad research questions/objectives were set out for this thesis:

- Is it possible to develop objective measures of challenge and self-efficacy from self-report data? (Objective 1)
- How can these measurements be communicated clearly to educators and learners? (Objective 2)
- Develop a practical computer software implementation that will communicate the measurements in real time (Objective 3)

The first objective was met by use of a formal measurement model. The model produces objective conjoint linear measurement of self-efficacy and challenge from ordinal input data, provided the assumptions of the model hold. No model can produce correct outputs from poor inputs. Thus, any measurement exercise also requires a test of the fit of the data to the model. The assumptions of the model are tested through a comprehensive set of hypotheses, which give assurances ranging from the quantifiability of the construct, through to fitness for purpose. These hypotheses are considered an integral part of the measurement process. In practice, failure of some of the hypotheses is to be expected from time to time: no model will ever fit the data *perfectly*. This suggests that the model should be used in a process of continuous improvement, so that issues found in any administration can be investigated and addressed in future administrations. It also suggests that judgement is required of how much trust can be placed in the measurements

produced in any administration, and thus whether the measurement is useful. Two features of the software support the likelihood of useful measurement. First, the software automatically adjusts the estimated standard error when any misfit is detected. Second, the software provides corrective options for failure of a number of the hypotheses. Although these options should be used judiciously, they allow useful measurement to proceed in many situations where the measurements would otherwise be rejected.

The second objective, that measurements be communicated clearly, led to the development of a simple metaphor for evaluation of success. The traffic light metaphor allows success to be communicated simply as: do not proceed, proceed with caution, or proceed safely. This objective also led to the provision of multiple output representations, such as qualitative labelling and the narrative evaluative summary (see Appendix B). In the software, purpose is specified by asking educators to specify how many statistically distinct strata they need. This is then mapped to the appropriate level of reliability needed for that purpose. Feedback from the group of educators who trialled the software suggested that this conception of purpose is readily understood and intuitive. Reliability indices are used throughout the software as the organising framework for fitness for purpose. This central role led to the development of a conceptual framework that mapped statistical and information theoretic terms and concepts to equivalents that are more familiar to educators. Again, feedback from the group of educators who trialled the software supported the value of this conceptualisation.

The third objective was to develop a practical computer software implementation that could communicate the measurements in real time. Consideration of the need for the implementation to be practical led to its implementation as an expert system. It was also necessary for the software to achieve useful measurement in a typical classroom setting. Theoretical considerations suggest that a minimum of 20 students and items are required for the basic purpose of triage. This is realistic in most educational settings. Moreover, the empirical datasets analysed all achieved accuracy in excess of the level needed to achieve triage. The need for real-time communication required the development of a formal approach to a time-series. It

also required the software to be able to manage calibration drift, as discussed in section 4.3.6.

Broadly speaking, it is believed that each of these objectives has been substantially achieved. However, there are some limitations; these are discussed in section 8.4.

8.3. CONTRIBUTIONS

A number of contributions have been made to the body of knowledge concerning educational measurement. Some of these contributions are theoretical. Others are simply seen as useful conceptualisations that may deserve wider dissemination.

The contributions to measurement theory are discussed first. The approach taken in this thesis has been to reimagine the process of measurement from an information theoretic perspective. Although largely congruent with extant theory and approaches, some differences arise from this approach.

First, analysis of the information content of polytomous items suggests that the natural score cannot be a sufficient statistic for a subject's ability. The essence of the argument is that a respondent's choice of category from those available is not independent of the choice of the other categories: the categories are mutually exclusive. Moreover, the choice of a category in an ordinal scale implies only that the underlying construct has been judged above the lower cut point or boundary of that category, and not above the upper boundary. No additional information is supplied by reference to the other categories. This contribution is encapsulated in the three theorems presented in section 5.7. The assumption that the natural score is a sufficient statistic underpins a number of widely used measurement models. The implications of this will require further work as described in section 8.5.

Second, the information theoretic approach has shown how the Rasch and related models can be extended to a time-series, or to an experimental situation that uses a test before and after treatment. Analysis suggests that application of a correction to information weight in the estimation process is sufficient to accommodate this. Moreover, the approach used to implement the time-series allows the use of incomplete and progressively unfolding data. This is in contrast to conventional

approaches which require all data to be collected before measurement can be established.

Third, it has shed light on the treatment of local dependence, a breach of the central assumption. An approach has been presented that allows one to move beyond diagnosis of local dependence to its automatic correction. Again, analysis suggests that application of a correction to information weight in the estimation process is sufficient to accomplish this correction.

The fourth contribution is the integration of a comprehensive framework for testing the measurement hypotheses within the measurement process. This integration allows a dynamic interaction between the diagnostic and estimation processes in which automatic correction of a number of diagnosed issues is possible.

Two contributions have also been made to the understanding of educational measurement. These are considered useful conceptions, rather than theoretical contributions. The first of these is the reimagining of reliability indices as measures of fitness for purpose; in turn, these reliability indices can be related to the number of statistically distinct strata. It is more natural for an educator to consider purpose as specified by concepts like triage (i.e. strata), than by the value of an index. This approach merits wider dissemination among the educational community.

The second contribution is the mapping of statistical terminology to equivalents which are more likely to be familiar to educators, and the quantification of the relationships between them. Although many educators are aware that factors such as the number of items, the number of categories per item, pass marks, and pass rates, have an effect on measurement accuracy, bringing these together into a single framework, with quantified relationships, is useful.

8.4. LIMITATIONS AND GENERALISATION

The theoretical and empirical evaluations suggest that useful measurement is achievable in real classroom settings. However, achieving this requires that the assumptions of the model hold. In practice, this is likely to require a process of

refinement in which problematic items are identified and then reworded or replaced. This may need a number of iterations.

Measurement requires estimates to be made at an individual level, rather than at the aggregate level that is generally used in research. Moreover, class sizes are often much smaller than are typically used in research. Achieving the level of accuracy required for measurement is therefore challenging, and can only be met if the right conditions apply. The heart of the logic of the model is that greater accuracy can be achieved if there are multiple *independent* estimates of the *same construct*. Independence is a fundamental requirement of the model. Although, a correction can be applied when the degree of dependence can be estimated, this adjusts the estimated standard error to reflect the accuracy achieved, rather than improving the intrinsic accuracy. The requirement that they are estimates of the same construct leads to the need for a scale to be unidimensional.

However, not all courses have a single coherent learning outcome. Consequently, some degree of multidimensionality can be expected. Ideally, when a scale is multidimensional, each component should be estimated separately and then combined into a single overall measure. This approach, however leads to the need for a greater number of items: the minimum number of items to achieve the specified purpose applies to each component. In practice, then, judgement is needed as to whether the larger instrument size required is practical, or whether the components are sufficiently coherent that they can be assessed as a single scale.

Although studied in the context of measurement of self-efficacy and challenge, the model used is generic, and potentially applies to any measurement situation where ordinal data are available. However, further study would be needed to establish specific issues relating to each situation, and whether the level of accuracy achieved in practice in such situations is sufficient for the associated purpose.

8.5. FURTHER WORK

The proof of concept software developed in conjunction with this thesis has been implemented as a generic measurement package. This package could form the basis of a commercial package. However, its use has only been validated at this time within an educational context. Further work is required to establish its validity and usefulness in other contexts, and to customise it for those contexts. One aspect of this relates to the expert system approach in which purpose is specified with relation to pass rates, pass marks etc. For more general use, this needs to be abstracted further and then re-implemented as a customisation of a more abstract form.

The use of narrative form for communicating the evaluation of success was found to be powerful and useful. The use of natural language forms for communicating outputs is likely to become increasingly important in the future of expert systems. Further work is needed to generalise this approach, and to develop the corresponding grammar and data structures for wider application.

The implications of the tension between information theory, and the assumption of the natural score as a sufficient statistic for polytomous items, need further investigation.

The close accord between the direct measure of self-efficacy, and the imputed measure, suggests that the latter could be a suitable proxy. Moreover, the source data for the latter are whether activities are not started, started, or completed. These data are conceptually simple and objective and could potentially be sourced automatically from a learning management system. Furthermore, such an approach should eliminate the issues arising from the reporting pattern adopted by students. Further work could evaluate the potential of this approach.

8.6. SUMMARY

In summary, this thesis has explored the practical measurement of self-efficacy and challenge by a software measurement model in an authentic educational setting. It has presented the model from its theoretical underpinnings, through to its

implementation in software and its evaluation from theoretical and empirical perspectives.

Measurement has been found to be both practical and useful in a real classroom setting. Although issues were found, the hypothesis framework was particularly useful in diagnosing and managing those issues.

REFERENCES

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N. Salkind, *Encylopedia of measurement and statistics.* Thousand Oaks, CA: Sage.

Alchin, L. (2005). *Measure for measure (1604)*. Retrieved from William Shakespeare info (the Complete Works online): http://www.william-shakespeare.info

Aldrich, J. (1997). R. A. Fisher and the making of Maximum Likelihood 1912 – 1922. *Statistical Science, 12*(3), 162-176.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-73.

Andrich, D. (1996). A general hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49*, 347-365.

Andrich, D. (1998). Thresholds, steps and rating scale conceptualization. *Rasch Measurement Transactions, 12*(3), 648-49.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical care, 42*(1), 17-116.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*(3), 181-192.

Armor, D. (1974). Theta reliability and factor scaling. In H. Costner, *Sociological methodology 1973-1974* (pp. 17-50). San Francisco: Jossey-Bass.

Babbage, C. (1864). *Passages from the life of a philosopher.* London: Longman, Green, Longman, Roberts, & Green.

Baker, F. (2001). *The basics of Item Response Theory.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191-215.

Bandura, A. (1978a). Reflections on self-efficacy. *Advances in Behavioural Research and Therapy, 1*, 237-269.

Bandura, A. (1978b). The self system in reciprocal determinism. *American Psychologist, 33*, 344-358.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1989). Social cognitive theory. In R. Vasta, *Annals of child development* (Vol. 6, pp. 1-60). Greenwich, CT: JAI Press.

Bandura, A. (1994). Self-efficacy. In V. Ramachaudrin, & V. Ramachaudrin (Ed.), *Encyclopaedia of human behaviour* (Vol. 4, pp. 71-81). New York: Academic Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: W.H. Freeman.

Bandura, A. (2001). Social cognitive theory: an agentic perspective. *Annual Review of Psychology, 52*, 1-26.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In T. Urdan, & F. Pajares, *Self-efficacy beliefs of adolescents* (pp. 330-337). Greenwich, Conn.: Information Age Publishing.

Bandura, A., & Locke, E. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology, 88*(1), 87-99.

Baumann, C. (2008). Kant and the magnitude of sensation: A neglected prologue to modern psychophysics. *Journal of the History of the Neurosciences, 17*, 1-7.

Baumgartner, H., & Steenkamp, J. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, XXXVII*, 143-156.

Beck, H., Levinson, S., & Irons, G. (2009). Finding Little Albert: A journey to John B. Watson's infant laboratory. *American Psychologist, 64*(7), 605-614.

Biggs, J. (2003). *Aligning teaching and assessment to curriculum objectives (Imaginative Curriculum Project).* LTSN Generic Centre.

Bock, R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bowling, S., Khasawneh, M., Kaewkuekool, S., & Cho, B. (2009). A logistic approximation to the cumulative normal distribution. *Journal of Industrial Engineering and Management, 2*(1), 114-127.

Brady-Amoon, P., & Fuertes, J. N. (2011). Self-efficacy, self-rated abilities, adjustment, and academic performance. *Journal of Counseling and Development, 89*(4), 431-438.

Brennan, R., & Kolen, M. (1987). Some practical issues in equating. *Applied Psychological Measurement, 11*(3), 279-290.

Brogden, W. (1944). Principles of behavior. *Journal of Consulting Psychology, 8*(5), 330-330.

Brown, J. (1996). *Testing in language programs.* Upper Saddle River: Prentice Hall Regents.

Bryman, A., & Hardy, M. (2009). *Handbook of data analysis.* London: Sage.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Caporaso, J. A. (1995). Research design, falsification, and the qualitative-quantitative divide. *The American Political Science Review, 89*(2), 457-460.

Cattell, R. (1966). The scree test for the number of factors. *Multivatiate Behavioral Research, 1*, 629-637.

Chen, C., & Wang, W. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement, 31*(5), 388-411.

Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behaviour. *Language*, 48-63.

Chomsky, N. (1972). Psychology and ideology. *Cognition, 1*(1), 11-46.

Cook, L., & Eignor, D. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45.

Coombs, C. (1964). *A theory of data.* New York: Wiley.

Crombie, A. (1994). *Styles of thinking in the European tradition.* London: Duckworth.

Cronbach, L. (1946). Response set and test validity. *Educational and Psychological Measurement, 6*(Winter), 475-494.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. (1971). Test validation. In R. Thorndike, *Educational measurement* (2nd. ed., pp. 443-507). Washington, DC: American Council on Education.

Dawkins, R. (1976/1989). *The selfish gene.* Oxford, UK: Oxford University Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1), 1-38.

Descartes, R. (1641). *Meditations on First Philosophy (tr. Veitch, J., 1901).* Retrieved from http://www.filepedia.org/files/Descartes'%20Meditations%20on%20First%20Philosophy.pdf

Dougiamas, M., & Taylor, P. (2003). Moodle: Using learning communities to create an open source course management system. In D. Lassner, & C. McNaught (Ed.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2003* (pp. 171-178). Chesapeak, VA: AACE.

Falkenstein, L. (2010). "Étienne Bonnot de Condillac". *The Stanford Encyclopedia of Philosophy* (Fall 2010). (E. Zalta, Ed.) Retrieved from http://plato.stanford.edu/archives/fall2010/entries/condillac/

Fechner, G. (1860/1966). *Elements of Psychophysics (Tr. Adler, H).* (D. Howes, & E. Boring, Eds.) New York: Rinehart & Winston.

Fechner, G. (1887/1997). My own viewpoint on mental measurement (Tr. Scheerer, E.). *Psychological Reseach, 49*, 213-219.

Ferguson, G. (1941). The factorial interpretation of test difficulty. *Psychometrika, 6*(5), 323-329.

Finney, S. J., & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology, 28*, 161-186.

Fischer, G. (1987). Applying the principles of specific objectivity and of generalizability to the measurement of change. *Psychometrika, 4*, 565-587.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A222*, 309-368.

Fisher, R. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society, 22*, 700-725.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*(4), 507-521.

Frontier, S. (1976). Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le modd́le du bâton brisé. *Journal of Experimental Marine Biology and Ecology, 25*(1), 67-75.

Goldstein, K. (1934/1995). *The organism: A holistic approach to biology derived from pathological data in Man.* New York: Zone Books.

Greanleaf, E. (1992). Measuring extreme response style. *Public Opinion Quarterly, 56*, 328-362.

Guilford, J. (1941). The difficulty of a test and its factor composition. *Psychometrika, 6*(2), 67-77.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139-150.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

Guttman, L. (1947). The Cornell technique for scale and intensity analysis. *Educational and Psychological Measurement, 7*(2), 247-279.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika, 19*(2), 149-161.

Harlow, H. (1958). The nature of love. *American Psychologist, 13*(12), 673-695.

Harlow, H., Dodsworth, R., & Harlow, M. (1965). Total social isolation in monkeys. *Proceedings of the National Academy of Sciences of the United States of America, 54*(1), 90-97.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*(1), 35-41.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

Herbart, J. (1877). Possibility and necessity of applying mathematics in Psychology (Tr. Haanel, H). *Journal of Speculative Philosophy, 11*, 251-264.

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information system research. *MIS Quarterly, 28*(1), 75-105.

Hölder, O. (1901/1996). The axioms of quantity and the theory of measurement (Tr. Michell, J & Ernst, C). *Journal of Mathematical Psychology, 40*, 235-252.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.

Hull, C. (1935). The conflicting psychologies of learning - a way out. *Psychological Review, 42*, 491-516.

Hull, C. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.

Hull, C., Hovland, C., Ross, R., Hall, M., Perkins, D., & Fitch, F. (1940). *Mathematico-deductive theory of rote learning: a study in scientific methodology.* Oxford, England: Yale University Press.

Irtel, H. (1995). An extension of the concept of specific objectivity. *Psychometrica, 60*(1), 115-118.

Iverson, G. (2006). Analytical methods in the theory of psychophysical discrimination II: The near-miss to Weber's law, Falmagne's law, the psychophysical power law and the law of similarity. *Journal of Mathematical Psychology, 50*, 283-289.

Jackson, D. (1993). Stopping rules in Principal Components Analysis: A comparison of heuristical and statistical approaches. *Ecology, 74*(8), 2204-2214.

Jansen, P., Van den Wollenberg, A., & Wierda, F. (1988). Correcting unconditional parameter estimates in the Rasch model for inconsistency. *Applied Psychological Measurement, 12*(3), 297-306.

Johnson, N., & Kotz, S. (1972). *Distributions in statistics: Continuous univariate distributions* (Vol. 3). New York: Wiley.

Kaiser, H. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 144-151.

Kant, I. (1781/1998). *Critique of pure reason (Tr. Guyer, P. & Wood, A.).* Cambridge: Cambridge Uiversity Press.

Kant, I. (1783/2004). *Prolegomena to any future metaphysics that will be able to present itself as Science.* Oxford: Oxford University Press.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2*(4), 389-423.

Kelley, T. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*(1), 17-24.

Kelvin, W. T. (1889). Electrical units of measurement. In W. T. Kelvin, *Popular lectures and addresses* (pp. 73-136). London: Macmillan.

Kim, D., De Ayala, R., Ferdous, A., & Nering, M. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement, 35*(6), 447-471.

Knox, H. (1914). A scale, based on the work at Ellis Island, for estimating mental defect. *Journal of the American Medical Association, 62*, 741-747.

Knuth, D. (1974). Computer programming as an Art. *Communications of the ACM, 17*(12), 667-673.

Koch, S. (1944). Hull's Principles of behavior; a special review. *Pschologial Bulletin*, 269-286.

Kolen, M. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-37.

Kolen, M., & Brennan, R. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11*(3), 263-277.

Krathwohl, D., Bloom, B., & Masia, B. (1964). *Taxonomy of educational objectives; the classification of educational goals. Handbook II.* New York: Longman, Green.

Kreuger, L. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences, 12*, 251-230.

La Mettrie, J., & Busey, G. (1912). *Man a machine.* Chicago: The Open court publishing co.

Li, H., & Wainer, H. (1997). Toward a coherent view of reliability in test theory. *Journal of Educational and Behavioral Statistics, 22*(4), 478-484.

Linacre, J. (2002). Facets, factors, elements and levels. *Rasch Measurement Transactions, 16*(2), 880.

Linacre, J. (2011). Winsteps® (Version 3.73.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved from www.winsteps.com

Linacre, J. M. (1997). KR-20 or Rasch reliability: which tells the "Truth"? *Rasch Measurement Transactions, 11*(3), 580-581.

Linacre, J., & Wright, B. (1994). Dichotomous Infit and Outfit mean-square fit statistics. *Rasch Measurement Transactions, 8*(2), 350.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data.* New York: Wiley.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45*, 507-530.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Lopez, M. (2005). Salsa: The Systematic Analysis of Learner Self-Appraisal. Auckland, New Zealand: Manukau Institute of Technology.

Lopez, M. (2007). Estimation of Cronbach's alpha for sparse datasets. *Proceedings of the 20th annual conference of the National Advisory Committee on Computing Qualifications*, (pp. 151-155). Nelson, New Zealand.

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Mahwah, NJ: Erlbaum.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

MacMahon, B., Pugh, T., & Ipsen, J. (1960). *Epidemiologic methods.* Boston: Little Brown.

Marais, I., & Andrich, D. (2008). Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 200-215.

Martin-Löf, P. (1974). The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. *Scandinavian Journal of Statistics, 1*(1), 3-18.

Maslow, A. H. (1943). A theory of human motivation. *Psychological review, 50*(4), 370-396.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Maus, A., & Endresen, J. (1979). Misuse of computer-generated results. *Medical & Biological Engineering & Computing, 17*, 126-129.

Maxwell, A. (1974). The logistic transformation in the analysis of paired-comparison data. *British Journal of Mathematical and Statistical Psychology, 27*, 62-71.

Meele, P. (1945). An examination of the treatment of stimulus patterning in Professor Hull's Principles of Behavior. *Psychological Review, 52*(6), 324-332.

Mehrens, W. (2005). The Consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-18.

Messick, S. (1989). Validity. In R. Linn, *Educational measurement* (3rd. ed., pp. 13-103). Washington, DC: American Council on Education & National Council on Measurement in Education.

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

Michell, J. (2003). Epistemology of measurement: The relevance of its history for quantification in the Social Sciences. *Social Science Information, 43*(4), 515-534.

Miller, W., & Rollnick, S. (2002). *Motivational interviewing: Preparing people for change.* New York: The Guilford Press.

Mokken, R. (1971). *A theory and procedure of scale analysis.* Paris: Mouton.

Mokken, R., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*(4), 417-430.

Mosteller, F., & Tukey, J. (1977). *Data analysis and regression.* Boston: Addison-Wesley.

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counselling Psychology, 38*(1), 30-38.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*(1), 59-71.

Muraki, E. (1992). A generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Nash, M. (1967). *Machine age Maya: The industrialisation of a Guatemalan community* (Vol. 60). Chicago: University of Chicago Press.

Nelson, L., & Rawlings, J. (1983). Ten common misuses of statistics in agronomic research and reporting. *Journal of Agronomic Education, 12*, 100-105.

OECD. (2009). *PISA 2009 technical report.* OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264167872-en

OECD. (2010). PISA Database 2009. Retrieved Jan 12, 2011, from http://pisa2009.acer.edu.au/

Oskamp, S. (1977). *Attitudes and opinions.* Englewood Cliffs, NJ: Prentice Hall.

Overmier, J., & Seligman, M. (1967). Effects of inescapable shock upon subsequent escape and avoidance responding. *Journal of Comparative and Physiological Psychology, 63*(1), 28-33.

Pajares, F. (1996). Self-efficacy beliefs and mathematical problem-solving of gifted students. *Contemporary Educational Psychology, 21*, 325-344.

Paulhus, D. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 36*(3), 598-608.

Paulhus, D. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman, *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Paulhus, D. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D. Jackson, & D. Wiley, *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Lawrence Erlbaum.

Pavlov, I. (1910). *The work of the digestive glands* (2nd. ed.). (W. Thompson, Trans.) London: Charles Griffin and Co.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 50*(302), 157-175.

Peffers, K., Tuuanen, T., Genger, C., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006). The Design Science research process: A model for producing and presenting information systems research. *In Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006) (S. Chatterjee and A. Hevner, Eds)*, (pp. 83-106). Claremont.

Perline, R., Wright, B., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*(2), 237-255.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33-40.

Piper, W. (1930). *The little engine that could.* New York: Platt & Monk.

Plank, M. (1968). *Scientific autobiography, and other papers.* New York: Greenwood Press.

Podani, J., & Miklos, I. (2002). Resemblance coefficients and the horseshoe effect in Principal Coordinates Analysis. *Ecology, 83*(12), 3331-3343.

Popham, W. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Popper, K. (1935/2005). *The logic of scientific discovery.* London: Routledge.

Powers, W. T. (1973). *Behavior: The control of perception.* Chicago: Aldine.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research), expanded*

edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rasch, G. (1968). A mathematical theory of objectivity and its consequences for model construction. *Paper presented at the European meeting on statistics, econometrics and management science.* Amsterdam.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58-94.

Richardson, J. (2005). Knox's cube imitation test: A historical review and an experimental analysis. *Brain and Cognition, 59*, 183-213.

Roberts, J., Donoghue, J., & Laughlin, J. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3-32.

Rogers, C. (1951). *Client-centered therapy: Its current practice, implications and theory.* London: Constable.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology, 21*, 95-103.

Ross, J. (1985). Misuse of statistics in social sciences. *Nature*(318), 514.

Rutter, M. (2007). Proceeding from observed correlation to causal inference. *Perspectives on Psychological Science, 2*(4), 377-395.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, No 17*.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph, No. 18*.

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models. *Psychometrika, 60*, 281-304.

Schunk, D. H., & Ertner, P. A. (1999). Self-regulatory processes during computer skill acquisition: Goal and self-evaluative influences. *Journal of Educational Psychology, 91*(2), 251-260.

Schunk, D. H., & Hanson, A. R. (1989). Self-modeling and children's cognitive skill learning. *Journal of Educational Psychology, 81*(2), 155-163.

Schwarzer, R. (1992). *Self-efficacy: Thought control of action.* Washington, DC: Hemisphere Publishing Corp.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379-423 July; 623-656, October.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13, 24.

Skinner, B. (1953). *Science and human behaviour.* New York: Macmillan.

Skinner, B. (1971). *Beyond freedom and dignity.* New York: Knopf.

Skinner, B. F. (1948). *Walden Two.* Indianapolis: Hackett Publishing Company.

Skinner, B. F. (1950). Are theories of learning necessary? *The Psychological Review, 57*(4), 193-216.

Skinner, B. F. (1974). *About behaviourism.* New York: Random.

Smith, R. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(1), 25-40.

Solberg, V. S., & Villarreal, P. (1997). Examination of self-efficacy, social support, and stress as predictors of psychological and physical distress among Hispanic college students. *Hispanic Journal of Behavioural Sciences, 19*(2), 182-201.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.

Stenner, A. (1997). *The Lexile framework: A map to higher levels of achievement.* Durham, NC: NetaMetrics.

Stevens, S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.

Stevens, S. (1957). On the psychophysical law. *The Psychological Review, 64*(3), 153-181.

Stone, M., & Wright, B. (1983). Measuring attending behavior and short-term memory with Knox's cube test. *Educational and Psychological Measurement, 43*, 803-814.

Taylor, B., & Kuyatt, C. (1994). *Guidelines for evaluating and expressing the uncertainty of NIST measurement results, Technical note 1297.* National Institute of Standards and Technology.

Thorndike, E. (1898). Some experiments on animal intelligence. *Science, 7*(181), 818-824.

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

Thurstone, L. (1926). The scoring of individual performance. *Journal of Educational Psychology, 17*, 445-457.

Thurstone, L. (1927a). Psychophysical analysis. *The American Journal of Psychology, 38*(3), 368-389.

Thurstone, L. (1927b). A Law of comparative judgement. *Psychological Review, 34*, 273-286.

Thurstone, L. (1928). The measurement of opinion. *Journal of Abnormal and Social Psychologyy, 22*, 415-430.

Todes, D. (1997). From the machine to the ghost within: Pavlov's transition from digestive physiology to conditional reflexes. *American Psychologist, 52*(9), 947-955.

Tolman, E. (1922). A new formula for behaviourism. *Psychological Review, 29*, 44-53.

Tolman, E., & Honzik, C. (1930). "Insight" in rats. *Publications in Psychology, 4*, 215-232.

Tolman, E., Ritchie, B., & Kalish, D. (1946). Studies in spatial learning I: Orientation and the shortcut. *Journal of Experimental Psychology, 36*(1), 13-24.

Van der Linden, W., & Glas, C. (2000). *Computerized adaptive testing: Theory and practice.* Boston, MA: Kluwer.

Vasta, R. (1976). Feedback and fidelity: Effects of contingent consequences on accuracy of imitation. *Journal of Experimental Child Psychology, 21*(1), 98-108.

Velleman, P., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47*(1), 65-72.

Vygotsky, L. S. (1978). *Mind in society.* Cambridge, USA: Harvard University Press.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220.

Wang, W., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296-318.

Watson, J. (1913). Psychology as the behaviourist views it. *Psychological Review, 20*(2), 158-177.

Watson, J., & Raynor, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology, 3*(1), 1-14.

Weber, E. (1834/1978). *The sense of touch (Tr. Ross, H & Murr, D).* London: Academic press [original work publised 1834 (De Tactu) and 1836 (Der Tastsinn)].

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*(4), 548-573.

Wood, D., Bruner, J., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*, 89-100.

Wright, B. (1995). Which standard error? *Rasch Measurement Transactions, 9*(2), 436.

Wright, B. (2003). Rack and stack: Time 1 vs. time 2. *Rasch Measurement Transactions, 17*(1), 905-906.

Wright, B., & Linacre, J. (1989). Observations are always ordinal; Measurements, however, must be interval. *Archives of physical medicine and rehabilitation, 70*(12), 857-60.

Wright, B., & Linacre, J. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B., & Masters, G. (1982). *Rating scale analysis.* Chicagi, Illinois: MESA Press.

Wright, B., & Stenner, A. (1999). Using Lexiles. *Popular Measuremet, 2*, 41-42.

Wright, B., & Stone, M. (1979). *Best test design.* Chicago: MESA Press.

Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd. ed.). Wilmington, Delaware: Wide Range inc.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Zumbo, B. D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (Ordinal) item scores.* Ottawa, ON: Directorate of

Human Resources Research and Evaluation, Department of National Defense.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

APPENDIX A

Background information on the OTIS package

As noted in chapter four, the model was implemented in the context of a general purpose data collection and analysis software package written by the author. Although the details of this package are outside the scope of the present thesis, its capabilities are briefly sketched here to provide context for the software implementation.

The package is designed to operate in a Microsoft Windows environment and is written using a combination of the C# and Visual Basic programming languages. It provides for the specification of arbitrary datasets, for the entry of data into the datasets, and for analysis of the data entered. Each dataset is organised into rows and columns with each row corresponding to a case or subject and each column to a variable. The package is declarative in its approach and includes an implementation of the general linear model and a number of conventional statistical procedures such as ANOVA and path analysis.

From the expert systems perspective, it was considered important to provide an integrated experience to the user, rather than merely supplying a number of scripts as plug-in procedures to a statistical package such as SAS or R. The package used was chosen as a basis for the implementation of the measurement model for the following reasons:

- It provided a robust, well tested, framework for the implementation
- It allowed arbitrary data definition and validation, thus simplifying the software development.
- It provided a framework in which all tests and hypotheses could be managed automatically.
- It provided generic import and export capabilities and could readily exchange data with a spread sheet package such as Excel.

APPENDIX B

The following is a sample of a diagnostic report produced by the software. The use of a *narrative* form is intended to support interpretation by those educators who are more comfortable with verbal descriptions than with tables of figures or graphs.

Evaluation of Test1 in *DMP SQL anonymised (DL)*

Summary

The declared measurement purpose was *triage* for items and *triage* for subjects. No corrective options were specified. The results of the measurement hypothesis tests were as follows:

- The hypothesis (H1) that the construct is quantifiable is supported.
- The hypothesis (H2) that response categories are ordinal is supported.
- The hypothesis (H3) that the construct is unidimensional is supported.
- The hypothesis (H4) that there is no differential item functioning (DIF) is supported.
- The hypothesis (H5) that there is no response set is supported.
- **The hypothesis (H6) that there is local independence is REJECTED**.
- The hypothesis (H7) that the measurement model converges correctly is supported.
- The hypothesis (H8) that data are adequate for measurement is supported.
- The hypothesis (H9) that there is a common metric for all measurements is supported.
- The hypothesis (H10) that response patterns are reproducible from measurements is supported.
- The hypothesis (H11) that there are no more outliers than expected is supported.
- The hypothesis (H12) that fit statistics accord with theoretical expectations is supported.
- The hypothesis (H13) that measurement is fit for purpose is supported.

Rejection of hypothesis H6 is not critical because the software can correct for lack of local independence in most cases. Consider setting the software option to do this.

Because not all hypotheses were supported, output measurements should be used with caution.

Overall, the scale can be classified as **productive, excellent quality**.

Items

Item separation reliability was 0.857; this is adequate for the purpose of triage. There were issues with the following items:

Q02

This item exhibits 37% local dependence on Q04 Q05. See footnote 1 for more information.

Q04

This item exhibits 53% local dependence on Q02 Q05.

Q05

This item exhibits 53% local dependence on Q02 Q04.

Q07

This item does not fit the model well. See footnote 2 for more information. It has been classified as

degrading because the *outfit* statistic is too large. See footnote 3 for more information.

Q13

This item has a response pattern that suggests the following categories may be disordinal: 2 to 3 [n1=4, n2=2, t=-2.22, p=0.0903]. See footnote 4 for more information.

Sample

Person separation reliability was 0.908; this is adequate for the purpose of triage. There were issues with the following subjects:

Diana

This subject exhibits 50% local dependence on Xena. See footnote 5 for more information.

Jane

This subject has a response pattern that does not fit the model well. See footnote 6 for more information.

Xena

This subject exhibits 50% local dependence on Diana.

Footnotes

1. Local dependence for an item occurs when the response to the item depends in part on the response to another item. This commonly occurs when a scenario is presented and several questions are asked relating to that item. The main effect of this is to cause the software to understate the estimated standard error of the ability estimates. The software option can correct this understatement.

2. Item misfit is diagnosed when there is a significant difference between the observed pattern of responses and the pattern predicted by the measurement model. This suggests that there are factors other than the ability of the subjects that are influencing responses. The standard error has been adjusted to correct for this, but consider replacing the item in future administrations.

3. Item outfit is diagnosed when there is a significant difference between the observed pattern of responses and the pattern predicted by the measurement model and this difference occurs mostly among responses where there is a large difference between the subjects' abilities and the difficulty of the item. This suggests that there are factors other than the ability of the subjects that are influencing responses. The standard error has been adjusted to correct for this, but consider replacing the item in future administrations.

4. Disordinality in item categories is diagnosed when the average ability of subjects responding in a lower category significantly exceeds the average ability of subjects responding in a higher category. The information provided is the category numbers of the disordinal categories, the number of subjects in each category, the *t statistic* for the difference and the probability of observing this degree of disordinality by chance when the model holds. Disordinality should be carefully investigated, but occasional diagnosis is expected and is not a major concern. Consider whether the item has too many categories to allow clear separation among subject responses.

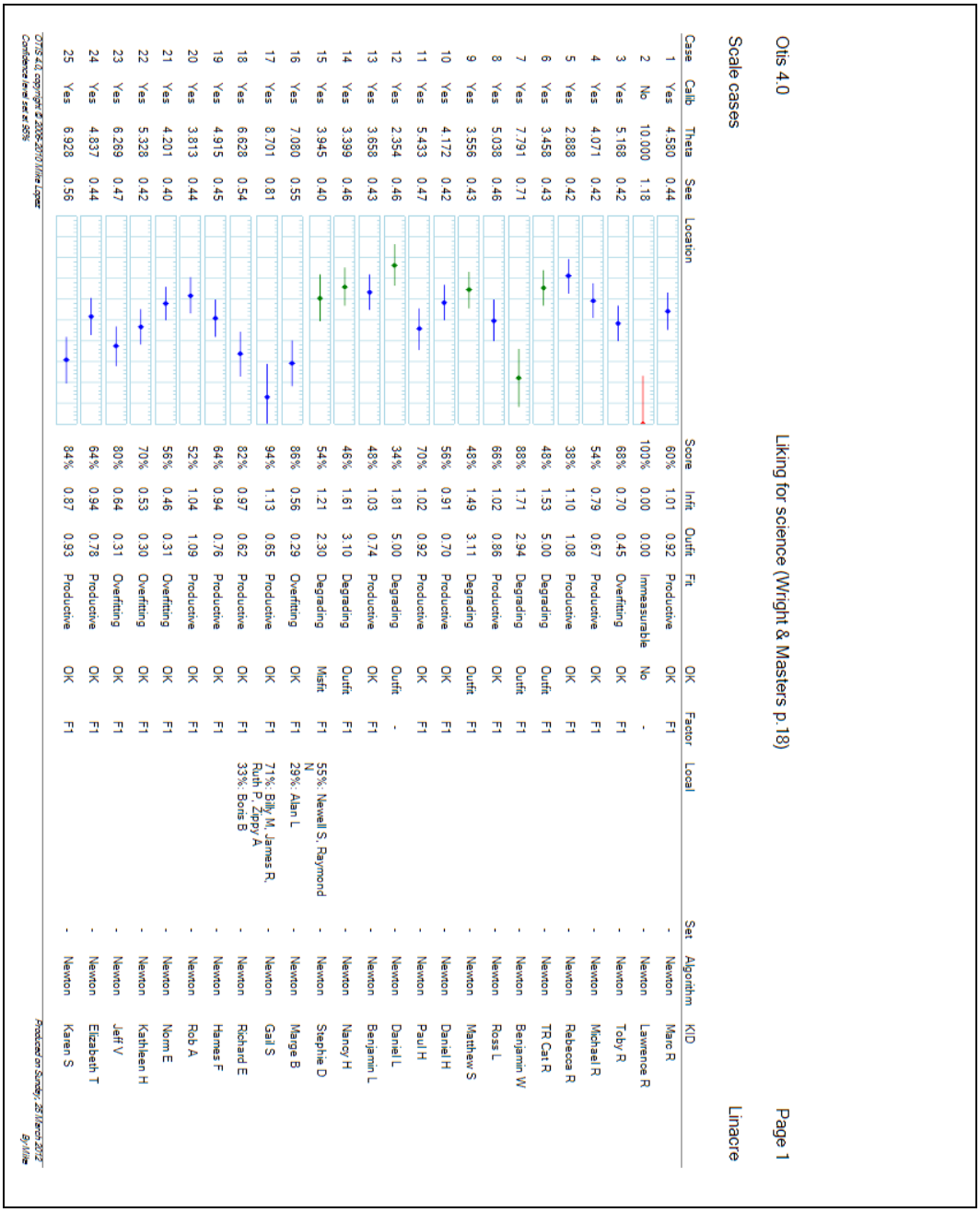
5. Local dependence for a subject occurs when the responses made are associated in some way with the responses made by another subject. This commonly occurs when students study together and thus may share misconceptions or the same pattern of domain knowledge. The main effect of this is to cause the software to understate the estimated standard error of the item and threshold difficulty

estimates. The software option can correct this understatement.

6. Response misfit is diagnosed when there is a significant difference between the observed pattern of responses and the pattern predicted by the measurement model. This suggests that there are factors other than the difficulty of the items that are influencing responses. The standard error has been adjusted to correct for this.

APPENDIX C

The report shown here is a larger version of figure 4.7.



Otis 4.0 Liking for science (Wright & Masters p.18) Page 1

Scale cases Linacre

Case	Calib	Theta	See	Location	Score	Infit	Outfit	Fit	OK	Factor	Local	Set	Algorithm	KID
1	Yes	4.580	0.44		60%	1.01	0.92	Productive	OK	F1		-	Newton	Marc R
2	No	10.000	1.18		100%	0.00	0.00	Immeasurable	No	-		-	Newton	Lawrence R
3	Yes	5.168	0.42		68%	0.70	0.45	Overfitting	OK	F1		-	Newton	Toby R
4	Yes	4.071	0.42		54%	0.79	0.67	Productive	OK	F1		-	Newton	Michael R
5	Yes	2.888	0.42		38%	1.10	1.08	Productive	OK	F1		-	Newton	Rebecca R
6	Yes	3.458	0.43		48%	1.53	5.00	Degrading	Outfit	F1		-	Newton	TR Cat R
7	Yes	7.791	0.71		88%	1.71	2.94	Degrading	Outfit	F1		-	Newton	Benjamin W
8	Yes	5.038	0.46		66%	1.02	0.86	Productive	OK	F1		-	Newton	Ross L
9	Yes	3.556	0.43		48%	1.49	3.11	Degrading	Outfit	F1		-	Newton	Matthew S
10	Yes	4.172	0.42		56%	0.91	0.70	Productive	OK	F1		-	Newton	Daniel H
11	Yes	5.433	0.47		70%	1.02	0.92	Productive	OK	F1		-	Newton	Paul H
12	Yes	2.354	0.46		34%	1.81	5.00	Degrading	Outfit	-		-	Newton	Daniel L
13	Yes	3.658	0.43		48%	1.03	0.74	Productive	OK	F1		-	Newton	Benjamin L
14	Yes	3.399	0.46		46%	1.61	3.10	Degrading	Outfit	F1		-	Newton	Nancy H
15	Yes	3.945	0.40		54%	1.21	2.30	Degrading	Misfit	F1	55% Newell S, Raymond N	-	Newton	Stephie D
16	Yes	7.080	0.55		86%	0.56	0.29	Overfitting	OK	F1	29% Alan L	-	Newton	Marge B
17	Yes	8.701	0.81		94%	1.13	0.65	Productive	OK	F1	71% Billy M, James R, Ruth P, Zippy A	-	Newton	Gail S
18	Yes	6.628	0.54		82%	0.97	0.62	Productive	OK	F1	33% Boris B	-	Newton	Richard E
19	Yes	4.915	0.45		64%	0.94	0.76	Productive	OK	F1		-	Newton	Hamas F
20	Yes	3.813	0.44		52%	1.04	1.09	Productive	OK	F1		-	Newton	Rob A
21	Yes	4.201	0.40		56%	0.46	0.31	Overfitting	OK	F1		-	Newton	Norm E
22	Yes	5.328	0.42		70%	0.53	0.30	Overfitting	OK	F1		-	Newton	Kathleen H
23	Yes	6.269	0.47		80%	0.64	0.31	Overfitting	OK	F1		-	Newton	Jeff V
24	Yes	4.837	0.44		64%	0.94	0.78	Productive	OK	F1		-	Newton	Elizabeth T
25	Yes	6.928	0.56		84%	0.87	0.93	Productive	OK	F1		-	Newton	Karen S

Confidence level set at 95%

Produced on Sunday, 25 March 2007 by Mike