DOCUMENT RESUME

ED 262 046 TM 850 503

AUTHOR Farish, Stephen J.
TITLE Investigating Item Stability: An Empirical Investigation into the Variability of Item Statistics Under Conditions of Varying Sample Design and Sample Size. Occasional Paper No. 18.
INSTITUTION Australian Council for Educational Research, Hawthorn.
REPORT NO ISBN-0-85563-389-1
PUB DATE Nov 84
NOTE 90p.
PUB TYPE Reports - Research/Technical (143) -- Statistical Data (110)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Achievement Tests; *Difficulty Level; *Estimation (Mathematics); Foreign Countries; *Goodness of Fit; *Item Analysis; Junior High Schools; *Latent Trait Theory; Mathematical Models; Mathematics Achievement; Reliability; *Sample Size; Sampling; Statistical Analysis; Statistical Studies; Test Construction; Test Items
IDENTIFIERS Australia; *Rasch Model; Wright (Benjamin D)

ABSTRACT

The stability of Rasch test item difficulty parameters was investigated under varying conditions. Data were taken from a mathematics achievement test administered to over 2,000 Australian students. The experiments included: (1) relative stability of the Rasch, traditional, and z-item difficulty parameters using different sample sizes and designs; (2) effect of different sample types and sizes on the Rasch item fit estimator; (3) effects, on the item fit parameter and the Rasch item difficulty parameter, of removing some less appropriate items from the test; and (4) an examination of Wright's statement that the standard error of the item difficulty parameter is a good estimator of its variance, and that it has an inverse square root relationship to the sample size. Results showed that Rasch and z-item difficulty parameters were similar. Item fit increased as sample size increased. The removal of poorly fitting items improved fit values for the entire test, but worsened them for the remaining items. The Rasch standard error parameter was an appropriate measure of the true error of estimation as calculated from the square root of the sampling variance of the item difficulty index. (Implications for test calibration are concluded, and detailed item analyses are appended.) (GDC)

Occasional Paper No. 18 - November 1984

INVESTIGATING ITEM STABILITY

An empirical investigation into the variability of item statistics under conditions of varying sample design and sample size

INVESTIGATING ITEM STABILITY

An empirical investigation into the variability of item statistics under conditions of varying sample design and sample size

Stephen J. Farish

Occasional Paper No. 18 - November 1984

Australian Council for Educational Research
Radford House, 9 Frederick Street,
Hawthorn, Victoria 3122, Australia

1984

Published by
The Australian Council for Educational Research,
Frederick Street, Hawthorn, Victoria 3122

Printed and bound by Brown Prior Anderson Pty Ltd,
Burwood, Victoria 3125

National Library of Australia Cataloguing-in-Publication data.

Farish, Stephen J., 1955-.
Investigating item stability.

ISBN 0 85563 389 1.

1. Education - Research - Statistical methods.
2. Sampling (Statistics). 3. Educational
statistics. I. Australian Council for Educational
Research. II. Title. (Series: Occasional paper
(Australian Council for Educational Research);
no.18).

370'.7'84

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

A research study requires the assistance of a number of individuals. I would like to express my gratitude to all those who have helped with this study, and, in particular:

Dr John Keeves, Director of the Australian Council for Educational Research on four counts - for suggesting the area of research, for making available the data and the resources of the ACER, for his suggestions and guidance throughout the study and for his patient encouragement at all times.

To colleagues both at ACER and elsewhere who provided advice and assistance when needed. In particular Dr John Izard for overseeing the item selection process; Jeff Clancy for having timely and immediate solutions to problems associated with the computing, and Dr Charles Poole for his comments and thoughtful advice. To Ray Adams and Dr Mark Wilson whose thoughts and suggestions added a wider perspective than one individual can muster.

To Dr Ken Ross for having brought to my attention the problems associated with sampling and for his helpful discussions in this area.

To Mr Mike Plunkett, who made available the data processing resources of Swinburne Institute of Technology, and his staff, in particular, Richard Forster, who was always ready to assist when difficulties arose. Whilst considering the computation I also thank Jeff, John, Jenny, Warren, Molly, Jan and Ray, all of whom found that my data processing requirements caused interruptions to their own work, yet underwent this inconvenience with cheerful forbearance.

This document was competently word-processed by Judith Clark and Carol Shackleton who understood even my most difficult annotations and scrawlings, and bore the changes an author continues making without complaint, for this I am grateful.

Last, but not least, my thanks to Denise for her encouragement and her help in arduous tasks such as proof-reading the text, and worse, the tables.

CHAPTER 1

PROBLEMS IN TEST CALIBRATION

Historical Perspectives

It is now one hundred years since Francis Galton pioneered the development of mental measurement with the publication of his 'Inquiries into Human Faculty and its Development' (Galton, 1883). Galton developed what are generally considered to have been the first mental tests, although it was James Cattell (1890) who introduced the term 'mental test'. By then the process of measurement had started on its way, the techniques of measurement and the standards to be employed for describing the instruments used to measure were emerging (see Suppes, 1976).

Cattell postulated a predictive relationship between his tests and future scholastic success (Cattell and Farrand, 1896). What was reportedly the first substantial use of the technique of correlation by Wissler (1901) illustrated no practical relationship between the scores Cattell obtained and the observed college grades of the students. The next major development occurred when Binet constructed mental tests to discriminate between retarded and normal children (Binet and Simon, 1905). The published tests (Binet and Simon, 1908) were widely used and became the model for later tests developed in other countries.

The involement of the United States in the First World War in 1917 required the selection of recruits as efficiently and effectively as possible. As a consequence, the US Army developed a variety of sub-tests designed to measure various attributes of the incoming recruits. The educational community observed this application of measurement and soon there were many tests emulating the content and format of the army tests. The notions of standardization (meaning the derivation of normative information on sections of a population) and validation (as measured by correlation with some independent measure or measures of the same attributes) were first recognized at about this time. As a result of these developments, it became clear that tests could measure more than just some form of general mental ability; in fact, they began to be applied to the identification of a wide range of somewhat independent dimensions of ability.

The practice of test development and the formulation of theoretical models of test performance grew together in the emerging field of educational and psychological measurement. By the time of the 1940s various models of test performance had been formulated and since then various parallel streams of development have continued.

The first stream was associated with the now classical model of true score and error. This model was practical in that it allowed the formulation of a number of useful relationships. It led to the development of parameters to describe the items comprising

the test. Such parameters were the item difficulty, item reliability and item discrimination, which is now usually measured by the point-biserial correlation between the item responses and the total score obtained from all the other items on the test. Parameters describing the test were forthcoming as well. These included reliability as it is measured in a number of forms, most popularly the Kuder-Richardson formulae (Kuder and Richardson, 1937), and such indices as the standard error of measurement. In addition, the classicial true score and error approach has allowed the exposition of relationships between test length and reliability or precision, such as the well known Spearman-Brown formula (Spearman, 1910). The classical model did, however, have certain problems associated with it, often related to the way the descriptive parameters of the test fluctuated with different samples of people used to estimate such parameters; that is, the parameters used to describe test and item performance were sample-dependent.

By now it was clear that test parameters were related to the selection of people, and also related fairly directly to the selection of items and the parameters of those selected items. The attention of the test developers and those who proposed models for describing test performance was directed at items more than ever before.

The second, and more recent, stream in educational and psychological measurement was initiated by Georg Rasch who formulated a model which focused on a single latent trait (Rasch, 1960). In this model, the function of a test was conceived to be the estimation of an individual's ability not in terms of an observable 'how many' from a domain of tasks, but in terms of 'how much' on a dimension representing the trait to be measured. This model was obviously suited to tasks that varied with respect to a single parameter, namely the difficulty of the task, rather than in the type or content of the task. The amount by which the tasks deviated from the assumed dimension in any way other than their difficulty has generally been measured using factor analytic methods.

Latent Trait Measurement

From the acceptance and use of the notion of a latent trait on which individuals might be measured and placed, there came the task of formulating an effective yet relatively practical model to describe the quantitative parameters of tests, and more specifically, of the items within the tests. Historically, a number of models has been suggested. However, they all have tended to take the approach of specifying a probability function which relates two parameters, the ability of the individual and the difficulty of the item, to the probability of passing the test item, that is, of answering it correctly.

There are in fact three parameters which may easily be demonstrated to differ between items. These are:

1 the location of the item on the trait, namely, the measure of its difficulty;

2 the rate with which the probability of success increases or decreases as one moves up or down the trait in ability, namely, the measure of its discrimination between candidates who differ only slightly in ability; and

3 the lower asymptote of the probability of success - a parameter designed to measure the success rate of candidates for whom the task is nearly impossible, but who have a substantial probability of success because of the item's construction (such as a multiple-choice item), namely, a measure of the probability of guessing correctly.

There are, currently, two main schools of thought regarding the role of these parameters. One either assumes that the second parameter is constant for all items and that the third is zero, or alternatively assumes that although they do not satisfy the above criteria, the amount by which they violate these two assumptions is not great enough to give rise to significant effects. The second approach treats all three parameters within the structure of the model. However, this approach gives rise to problems associated with the complexity of the model and the numbers of candidates necessary to estimate accurately the three parameters for every item.

The model under consideration in this monograph is commonly called the 'one parameter model' because it assumes that the first parameter is the only one which varies, that the second is constant, and that the third is zero. Thus, each item is completely described by only the first parameter, which is the item difficulty value.

Various mathematical formulations for this probability model have been postulated. The two most prominent ones are the normal ogive model and the logistic ogive model. These two are nearly equivalent; the difference between the shape of a logistic ogive and a normal ogive being less than one per cent at the most. The logistic ogive has gained dominance however, because the logistic function is demonstrably easier to manipulate algebraically, and because of the separability of person and item measures. It is this more widely used logistic model which is being investigated in this study.

The simple formulation of the one parameter logistic model is:

$$\mathrm{pr}(X = x \mid b_k, d_i) = \frac{\exp(x(b_k - d_i))}{1 + \exp(b_k - d_i)}$$

where $\mathrm{pr}(X = 1 \mid b_k, d_i)$ is the probability of success for candidate k on item i,

b_k is the estimated ability for candidate k,

and d_i is the estimated difficulty for item i.

It is clear that the parameters describing tests and items depend on the sample of candidates used for calibration, and that the parameters of items may depend on

characteristics of the other items in the test. Because tests are composed of items the test parameters are clearly dependent upon the characteristics of the items.

The Rasch model eliminates the dependence of item (and therefore test) parameters upon the calibration sample when all the assumptions of the model are met. This has not been achieved in practice because these assumptions are not met in real-life testing situations. A test designed to measure the ability of a person, either in the latent trait sense, or in a traditional norm-referenced sense is therefore only as good as the items which it contains. These items must meet a number of criteria. First, the items must be valid. That is, they must exhibit content validity and should be representative of the domain of behaviours on tasks they purport to measure. They must exhibit construct validity, and in so far as is possible, the dimension on which attributes are measured must be unidimensional. It is hoped that the items will have predictive validity: they must be able to be used to predict with some accuracy the success of the candidates, who are tested on tasks drawn from the specified domain. Secondly, the items must also be reliable. They should measure whatever they do measure consistently. If an item is inconsistent in the measures of performance it provides then its usefulness is correspondingly reduced.

Obviously the validity and reliability of a test, and thus also its usefulness, are related to the validity and reliability of the items which make up the test. In latent trait measurement the test is only as sample-independent as are the items from which the test is constructed. It is important that items can be demonstrated to possess certain measurable attributes which do not fluctuate widely under conditions of differing calibration samples. Nor should they fluctuate according to the presence or absence of other items in the test. It is the extent of this fluctuation with different samples and with different item compositions which is under investigation in this study.

CHAPTER 2

RESEARCH INTO THE ONE-PARAMETER MODEL

The Rasch Model Emerges

Gulliksen (1950) remarked that the discovery of item parameters which would remain stable as the analysis group changed would constitute a significant contribution to item analysis theory. Rasch (1960) outlined a model which had the theoretical capability to separate item and person measures, so that the item measures were independent of the person measures, and therefore independent of the sample, or analysis group. It was with Wright's paper 'Sample-free Test Calibration and Person Measurement' (Wright, 1967) that this technique to achieve this stability of item parameters across different analysis groups became more widely investigated by those in the measurement field.

Wright illustrated this technique through a relatively straightforward example of test calibration using two samples of subjects which had been set up to be as different as possible in ability. The resulting test and item parameters were very nearly equal, but not exactly so. Since Wright's initial application of the Rasch model to this problem in educational measurement, the question of the degree to which the test and item parameters approximated invariance across different sample groups has become the issue of many subsequent studies. If the assumptions of the model are met, in theory the item parameters of the test and the person ability parameters are separable, and thus the item parameters are considered sample-free, and consequently invariant, across different samples. In this invariance lay the central feature and the promise of the Rasch model; without it this model would be no more useful than any other model. However, once it had been established that this invariance was only approximated in practice, researchers came to speculate on just how stable the item parameters were. Consequently the model began to be tested for both its robustness across different samples and the degree to which the model would remain robust as the assumptions it demanded were violated.

Hambleton (1969) outlined these assumptions underlying the model. In particular, he showed that the item scaling procedure was insensitive to violation of the assumption that all items had equal discrimination. Panchepakesan (1969) also illustrated this robustness of the model when item discrimination values varied significantly.

Whitely and Dawis (1973) subsequently argued that the Rasch model would not make a significant impact on test development until the technology of latent trait measurement became more sophisticated. History has shown the correctness of their prediction. During the decade from Wright's paper in 1967 and the papers of Hambleton and Panchepakesan in 1969 until the late 1970s little was heard of the Rasch model apart

from isolated research studies. Since then many people have made varying applications of the model and investigated its properties. Others have improved the technology of its estimation procedures so that currently the most frequently adopted procedure is the Unconditional Maximum Likelihood Estimation technique (usually abbreviated to UCON). This method uses successive adjustments (in an iterative manner) of the item difficulty and person ability parameters to generate the closest possible fit of the data to the model. This technique, whilst accurate, is still cumbersome, requiring substantial data processing resources. It is possible that other approaches may, in future, yield equally good parameter estimates with reduced computational requirements.

Sample Size Effects

Forster (1976) investigated the relationship between sample size and the point-biserial discrimination values and the mean square fit values for items. He determined that as the sample size increased the fit values also increased. It was also established that the point-biserial correlation coefficients remained relatively constant, and the average deviation between the theoretical and the true item characteristic curves increased. Close inspection of Forster's tables has indicated that these effects relating item parameters to sample size were not pronounced within the range of sample sizes used (smallest 98, largest 508). Interestingly, Forster did not draw multiple samples to check for the stability of item parameters. He relied on only one sample at each sample size to investigate the trend as sample size was increased.

The conclusions reached by Whitely and Dawis (1973) were similar to those of Forster, who suggested that in order to estimate parameters it would be necessary to have a minimum number of three to five students at each score point, whereas Wright (1967) had contended that accurate estimation was possible even if a number of sequential score points had no students at all. Forster concluded from this perceived necessity to have a minimum number of students at each score point that it was also sufficient and *necessary* in order to obtain stable item difficulty and student achievement estimates. Whitely and Dawis (1973) concluded that even with a group size of 500 or more, some values such as extreme scores could not be estimated accurately.

Haberman (1975) has demonstrated that for a fixed length test the maximum likelihood estimates of the item parameters for the one-parameter model converge to their true value as the sample size tends to infinity. This is, however, an intuitive proposition as all sample parameters converge to the population values as the sample size increases and this effect is seen in the finite population correction of the few known formulae for the sampling variance of many statistics. Conversely, Andersen (1973a, 1973b) showed that when the number of examinees was increased the maximum likelihood procedure did not yield stable estimates of the item difficulty parameters.

Wright (1977) and Whitely (1977) have strongly debated certain issues of the Rasch technique. Wright made clear that there is a direct relationship between the standard error of items and the size of the sample used for calibration. This relationship is claimed to be of the same type as when estimating the sampling error of the mean, namely that the sampling variance of the estimated parameter is inversely proportional to the sample size. This means that the standard error of the item difficulty parameter is inversely proportional to the square root of the sample size. Wright then gave a table relating samples of particular sizes to standard error estimates, and suggested that although sample size does have a relationship to the standard error, that for all practical purposes the item difficulty parameter may be estimated from samples as low as 100. He further stated that sample-invariance depends on a demonstration that the difficulties of the items remain statistically equivalent over the various kinds of persons to be measured using those items, and that this condition is investigated when evaluating the data for fit to the model.

Whitely argued that although difficulty estimation is possible from smaller samples, the estimation of fit becomes more powerful when larger samples are used. She also pointed out that when *differences* in item difficulty were being investigated then larger samples would be necessary.

Forster (1978) examined the issue of sample size by taking five samples at each of four sizes (50, 100, 200, and 300) from a population of approximately 1400. Examination of the correlation between item difficulty estimates based on the samples and the population values led him to suggest that for sample sizes less than 200 the accuracy dropped considerably (as measured by this correlation procedure) and that as sample size increased beyond 200 the increase in accuracy was not substantial. Forster also calculated the standard deviation of the item difficulty estimates for each sample size and compared this with the standard deviation of the item difficulty estimates for the population. This was done in order to compare the equality of the sample difficulty parameter metric to that for the population. The ratio was very nearly unity for all sample sizes, although it did exhibit a slight general decrease as the sample size increased.

Douglass (1980) investigated the stability of the one-parameter (Rasch), two-parameter and three-parameter models across samples of size 200, 600, 800 and 1082. He concluded that the Rasch model was the most useful in that it gave the most consistent calibration of items, particularly for smaller sample sizes. Whitely (1980) has since expressed the opinion, based on a review of earlier studies, that item calibrations for the Rasch model which are sufficiently precise for research applications can easily be obtained from samples of 250 or even less.

Cornish (1983) has investigated empirically the stability of item difficulty estimates for both Rasch and traditional item difficulty parameters. He took 60 samples

each of 120 students from a population of 2342. No inferences could be drawn about the effect of sample size as only one size was used. However, he argued that the Rasch estimates were more stable than the traditional ones because the empirically measured sampling variance of the Rasch estimates was less than the corresponding empirically measured sampling variance for the traditional estimates. Whilst his tables support this conclusion, no account was taken of the different metrics used for the two types of item difficulty estimate. Cornish also used two sample types (simple random and cluster samples) and observed that the *type* of sample did not seem to affect the stability of the Rasch difficulty estimates.

Item Selection Criteria

Not only has the formulation and size of calibration samples come under investigation with mixed results, but item selection criteria have also been investigated and debated at some length. Andersen, Kearney and Everett (1968) investigated the stability of Rasch item difficulty parameters and found that items which fitted the Rasch model well had more stable estimates. Tinsley and Dawis (1975) argued likewise that stability was related to goodness of fit, and went further to suggest that the deletion of poorly fitting items increased the stability of those remaining, although excessive deletion caused a subsequent drop in stability. Tinsley and Dawis also investigated the 'z-item difficulty index' - a standardized form of the traditional item difficulty, and found it to be less stable than the Rasch parameters. Both these research studies used a two-sample design for measuring stability and employed correlation measures to indicate the level of stability in a quantitative manner for comparative purposes.

Forster (1976) investigated the relationship between the point-biserial discrimination values and the mean square fit values for items, and suggested that differences in point-biserial values between items did not affect item difficulty values but did affect item fit values. On a more practical note, Dinero and Haertel (1977) also found that the lack of an item discrimination parameter in the simple logistic (Rasch) model did not result in poor calibration in the presence of varying item discrimination. They therefore suggested that with this in mind test constructors should select items of high discrimination in order to maximize the information available through the use of the test. Forster and Karr (1979) have suggested that neither the point-biserial discrimination value nor the mean square fit value was a satisfactory criterion for the selection of items or the ascription of item quality. They suggested that the item characteristic curve should be consulted in order to select appropriate items for the Rasch model. Similarly George (1979) investigated the standardized residual mean square fit statistic and concluded that it did not detect unacceptable variation in item discrimination. He argued that in order for Rasch model analyses to work in practice the

item discrimination values must be very similar. This was a contrary proposition to those advanced by the earlier researchers who suggested that discrimination values which varied significantly would not substantially affect the stability of the calibrations. However, very few researchers seem to have defined these terms, such as 'similar' or 'varied' in a manner which made clear their meaning in a *quantitative* sense. Forster and Ingebo (1978) successively reduced the number of items in a test from 80 to 15 by excluding those items which were at the extremes of the calibration. They concluded from a correlational procedure that the *range* of item difficulties in a test did not affect the item scaling procedure.

The contextual stability of item parameters was investigated by Yen (1979), who correctly pointed out that the use of the correlation between estimates to indicate stability was not entirely appropriate. Correlation values simply indicated the strength of a linear relationship between two variables but did not indicate the degree of equality of those two variables. Yen argued that since Rasch item difficulty estimates were nonlinear but monotonic transformations of traditional item difficulty values, then the rank order was preserved between the two types of statistic. Yen's study illustrated that contextual effects were greater for item discrimination values than for difficulty estimates. She also investigated the effect of increasing sample size on difficulty parameters, but could not easily interpret these effects because of simultaneous contextual differences. She proposed that if predictions about individual items were of concern to the researcher then it would be wise to use the same context for calibration and the later use of the item. If the same context was not to be used then very large calibration samples, of more than 600, should be employed.

Summary

In retrospect it has been found that no investigation has accurately and systematically quantified the stability of the Rasch item difficulty parameters under a variety of conditions. Such conditions include varying sample size and the deletion of poorly fitting items in such a way as to allow comparisons of stability measures between situations of interest to the researcher. Previous studies have used methods of quantifying stability which take no account of the metric, such as correlation measures. Many previous studies have used as few as two samples at each sample size to investigate the effect of different sample sizes on the stability of item parameters. It seems strange that investigations into the effects of sampling on these parameters have assumed that the effects caused by sampling were sufficiently small between samples of the same size to enable valid comparisons to be made between samples of different sizes. More specifically, it seems strange that the studies which investigated the effects of sampling

on the stability of item calibrations did not seem to take into account the effects of sampling on the parameters used to measure stability.

Wright (1967) outlined the advantages of perfectly invariant item parameters. However, subsequent work has shown that these Rasch model item parameters are not perfectly invariant in practice. The conflicting findings of research into the stability of these parameters under conditions of different sample sizes and different levels of item fit have indicated that the question of just how stable these parameters are under a number of changing conditions has not been fully explored. This study sets out to fill a few of the more fundamental gaps in our knowledge of this area.

CHAPTER 3

THE DESIGN OF THE STUDY

Introduction

This chapter describes the data used in this study, the research questions regarding parameter estimation for which answers are sought, the relationships between parameters which require closer examination, the parameters themselves, and the procedures used to inquire into the issues under investigation.

The Population Data

This study has been made possible through the availability of data on the item responses on a 55-item test of mathematics achievement for a population of Australian students. These data were collected as part of a study which examined the contributions of home, school and peer group environmental factors to changes in the educational achievement of students in the first year of post-primary education (Year 7) in the Australian Capital Territory (Keeves, 1972). Keeves gathered data on the whole population of first-year post-primary students in the Territory, and these students were grouped according to the classes in which they were taught.

In 1969 there were 15 secondary schools in the Australian Capital Territory; nine co-educational Government high schools, four Catholic high schools (two for boys and two for girls), and two Anglican high schools (one for boys and one for girls). The number of students in the target population which was obtained from census data, and the number of students in the achieved population for whom Keeves obtained data are presented in Table 3.1.

The differences between the figures for Keeves' data and those obtained from the census may be ascribed to absenteeism on the day of testing, the movement of population elements between the census date and the date of testing, and the exclusion of one small classroom of children because of the atypical nature of this class.

The Sampling Frame

This frame is described in detail in Table 3.2. Each school has been numbered, and the classes within schools have been numbered from class 01 to class 75 with the numbers of students within each class also given. Square brackets describe classes which were paired into 'pseudoclasses' so that later application of cluster sampling would provide large enough classes for the specified cluster sizes. The bracketed number at the end of each school is the number of students in that school.

Table 3.1 Students Attending Secondary Schools at Form I Level in the Australian Capital Territory during 1969

	Census[a]	Keeves
Government schools	1714	1611
Non-government schools	764	743
Total	2478	2354

[a] From CBCS, 1970a; and CBCS, 1970b.

The Questions Under Investigation

From the previous chapter it is clear that there were many research questions which could be examined; this study confines itself to some of the more important issues.

First, the relative stability of the Rasch item difficulty parameter, the traditional item difficulty parameter, and the z-item difficulty parameter were investigated under conditions of differing sample size and design, and the specific relationship between their stability and the sample size was also examined. Secondly, the effect of different sample types and sizes on the Rasch item fit estimator for items were investigated. In this case the fit estimator used was the one recommended by Wright, obtained through the comparison of multi-group maximum likelihood item response estimates (see Whitely, 1977:230). Thirdly, the effects on the item fit parameter and on the Rasch item difficulty parameter of removing some of the less appropriate items from the calibration were examined. Fourthly, the statements by Wright (1977) that the standard error of the item difficulty parameter is a good estimator of the variance of the item difficulty parameter and the statement that the standard error of the item difficulty parameter has an inverse square root relationship to the sample size also came under investigation.

The Procedures

It is clear that these four questions require that the characteristics of the samples being taken should vary, and that a sufficient number of samples should be taken to determine empirically the effects of such sample types and designs upon the parameters in question. To this end four sample designs were employed: simple random samples, cluster samples with clusters of size 5, cluster samples with clusters of size 10, and cluster samples with clusters of size 20. Each cluster was drawn from one classroom, and classrooms were drawn without replacement. For each of these designs a total of nine sample sizes were employed (40, 60, 80, 100, 120, 160, 200, 240, 320). In addition to analyses carried out on the total test of 55 items, the original 55 items were reduced to 42 by eliminating those items which were not truly appropriate to Rasch calibration.

That is, the items at each extreme of difficulty and those which had poor fit to the item characteristic curve were deleted. These 42 items, considered appropriate for Rasch calibration, were subsequently reduced to 32 by the elimination of items with extreme fit statistics from among the 42. This meant that the effect of deleting poorly fitting items could also be studied. These analyses were repeated for three test lengths (55, 42 and 32 items) giving a total of 108 systematically different combinations of sample size, sample design and test length. For each of these 108 combinations 200 random samples (replications) were drawn, and for each of these 21,600 random samples an estimate of the Rasch item difficulty, the traditional item difficulty, the z-item difficulty, the Rasch fit statistics, the standard error of the Rasch item difficulty and the traditional point biserial discrimination value (unbiased) were calculated for each item. The Rasch statistics were generated by the computer program BICAL (Wright and Mead, 1977). The traditional difficulties were calculated by a small FORTRAN routine written specially for this study. The further computations and statistics later generated and presented in tables and graphs were largely generated by the statistical package SAS (Statistical Analysis Systems). All the computation was performed on a FACOM M180N system under TSS.

The Investigations

Investigation 1: The Stability of Item Difficulty Parameters

In his study of the relative stability of different item difficulty parameters Cornish (1983) made comparisons between the sampling variance of Rasch and traditional item difficulty parameters. To enhance the interpretation of the Rasch parameters Cornish transformed the values obtained by multiplication by 4.551 (equal to the reciprocal of the natural logarithm of 3) and the addition of 50 units. This transformation produced item difficulty estimates which tended to fall between 0 and 100 and so seemed more manageable than the original logits which were centred on zero and had numerically small values. The transformation also produced a neat factor of 3 change to the odds of success on an item every time the difficulty (or person ability) changed by 5 units. Cornish also expressed the traditional difficulty values as percentage values. The problem with such transformations was that any empirically determined variance of an item difficulty is correspondingly expanded or contracted in accordance with the transformation which is used. The traditional and Rasch item difficulties were not on the same metric in the first place in Cornish's study, nor were they after each had been linearly transformed. To overcome this problem it was necessary to find some way to make these variance estimates comparable. A relatively straightforward solution was to divide the individual item variances obtained (in whatever metric) by the variance of the actual item difficulty parameters across the test (in the same metric). This procedure

Table 3.2 The Sampling Frame

SYSTEM 1			
SCHOOL 01	CLASS 01	37	
	CLASS 02	36	
	CLASS 03	39	
	CLASS 04	38	
	CLASS 05	28	
	CLASS 06	10	(183)
SCHOOL 02	CLASS 07	34	
	CLASS 08	33	
	CLASS 09	28	
	CLASS 10	25	
	CLASS 11	30	
	CLASS 12	28	
	CLASS 13	17	(195)
SCHOOL 03	CLASS 14	32	
	CLASS 15	31	
	CLASS 16	23	
	CLASS 17	15	
	CLASS 18	29	(130)
SCHOOL 04	CLASS 19	38	
	CLASS 20	36	
	CLASS 21	36	
	CLASS 22	36	
	CLASS 23	19	(165)
SCHOOL 05	CLASS 24	40	
	CLASS 25	35	
	CLASS 26	35	
	CLASS 27	30	
	CLASS 28	36	(176)
SCHOOL 06	CLASS 29	36	
	CLASS 30	37	
	CLASS 31	30	
	CLASS 32	21	
	CLASS 33	9	(133)
SCHOOL 07	CLASS 34	35	
	CLASS 35	39	
	CLASS 36	37	
	CLASS 37	36	
	CLASS 38	33	
	CLASS 39	21	(201)

SYSTEM 1 (continued)			
SCHOOL 08	CLASS 40	34	
	CLASS 41	36	
	CLASS 42	35	
	CLASS 43	37	
	CLASS 44	27	
	CLASS 45	32	
	CLASS 46	25	
	CLASS 47	27	
	CLASS 48	21	(274)
SCHOOL 09	CLASS 49	32	
	CLASS 50	33	
	CLASS 51	32	
	CLASS 52	31	
	CLASS 53	26	(154)
SYSTEM 2			
SCHOOL 10	CLASS 54	38	
	CLASS 55	40	
	CLASS 56	35	(113)
SCHOOL 11	CLASS 57	35	
	CLASS 58	33	
	CLASS 59	34	
	CLASS 60	31	(133)
SCHOOL 12	CLASS 61	38	
	CLASS 62	37	
	CLASS 63	38	
	CLASS 64	30	
	CLASS 65	18	(161)
SCHOOL 13	CLASS 66	40	
	CLASS 67	44	
	CLASS 68	48	(132)
SYSTEM 3			
SCHOOL 14	CLASS 69	26	
	CLASS 70	26	
	CLASS 71	29	(81)
SCHOOL 15	CLASS 72	32	
	CLASS 73	30	
	CLASS 74	30	
	CLASS 75	31	(123)

eliminated the effect of both transformations and also eliminated the metric used because the variance of the item difficulty parameters were expanded and contracted in exactly the same manner as the individual item variances across the samples under the linear transformation. This process is not unlike transformations used in the analysis of variance, where the ratio of between groups and within groups variability is examined. In this study, the between groups variability is measured by the variance of the item difficulty measures across the test (the means of the difficulty parameter for each item across the samples is used for the item difficulty values). The within groups variability is the variance of the item difficulty parameter across the samples (the mean of the item variance parameter is taken to represent the within groups variability). If a general stability estimate is required for all the items that constitute a test, the average of this within-item between-samples variance may be divided by the between-items variance. This procedure enabled the sampling variance of each item to be expressed as a fraction of the total spread of difficulties encompassed by the items which made up a test. As such, it is a measure of the separability of the items within a test as calibrated using the particular sample. This procedure was applied to the three item difficulty parameters, the Rasch, traditional and z-item difficulty indices. The effect of different sample types and sizes on the ratio of variances just described, and the comparative relationship between the ratios for the three parameters were investigated so that it could be determined which parameter gave the most stable estimates under varying conditions.

Investigation 2: The Effect of Sample Parameters on Item Fit and Point-biserial Discrimination

In this investigation the contentions of Forster (1976) were examined. Namely, that as sample size increased the fit values also increased but the point-biserial discrimination values remained the same. At the same time the effect of different sample sizes on both item fit and point-biserial discrimination were investigated, and various explanations explored. The effects of sample designs on fit and point-biserial discrimination were also examined.

Investigation 3: The Effect of Deleting Items which do not Fit the Rasch Model on Item Fit and Item Variance

Analysis of the items produced clear indications that certain items did not fit the Rasch model, because of extreme facility or difficulty, or because of poor fit values related to variability in discrimination values between items. Items which discriminated either too poorly or too well were eliminated. For this purpose the point-biserial discrimination values were also consulted. Items which were too easy or too difficult were also eliminated, as were poorly fitting items. This procedure was performed twice, yielding the two sub-tests, one of 42 items and one of 32 items. The items which were eliminated and the reasons for doing so are given in Appendix F. The reason for a two-stage

procedure was that the original 55 items were selected on the basis of traditional criteria, particularly for high point-biserial discrimination. Thus the first elimination process produced a test which contained items appropriate to the Rasch model, and the second eliminated those which were less appropriate (in terms of fit value) as members of the larger group of 42 items which fitted Rasch criteria. The effect of this procedure on the fit values of the items and on the stability of the item difficulty parameter for items belonging to the smallest item group were then investigated.

Investigation 4: The Relationship between Rasch Item Variance, Standard Error, and Sample Parameters

In this part of the study a straightforward comparison between the sampling variance of the Rasch item difficulty index and the Rasch standard errors of the items enabled the validity of the standard error, as an estimator of the error associated with an item difficulty value, to be examined, both for individual items and, in general, across all items. The notion that there existed a simple inverse relationship between the square of the standard error (or the sampling variance) and the size of the calibration sample as Wright (1977) had contended, was able to be investigated.

The Finite Population Correction

It is clear that the estimation of sampling variances for most statistics is in error as the size of the sample approaches a significant proportion of the population. In this study the samples ranged from 2 per cent to 14 per cent of the population, and as such it was considered necessary to incorporate a finite population correction to the variance estimates empirically determined from multiple samples. No formulae were available for this correction for the more complex statistics such as the Rasch item difficulty and the z-item difficulty indices. As a first-order approximation the standard form of $(N - n)/(N - 1)$, (where N is the population size and n is the sample size), which is appropriate for the traditional item difficulty statistic, was used. This correction was applied to the variance estimates of the difficulty parameters, where it was considered appropriate.

Summary

The questions asked by these investigations have remained largely unanswered for some years now. This study aims to provide some steps towards a better understanding of the problems associated with the measurement of item difficulty, as well as providing partial answers to some of the uncertainties in our knowledge of this area.

CHAPTER 4

THE RESULTS OF THE STUDY

In this chapter the results of the four investigations previously outlined in Chapter 3 are discussed in turn. It becomes clear that they are not independent investigations but rather, inter-related studies since observations made in any one investigation are associated with, and are, necessarily, consistent with observations in the other investigations.

Investigation 1: The Stability of Item Difficulty Parameters

This investigation covers a number of questions related to the variance of item difficulty indices. In the first stage of this investigation the mean raw item difficulty sampling variances are examined. These are plotted for each of the three item difficulty indices under consideration in Figures 4.1, 4.2 and 4.3, and the actual values are given in Appendix D. It was essential that some form of common metric should be used in the comparisons between the three item difficulty indices. Consequently, the mean raw item difficulty variance across the test length was divided by the variance of the mean item difficulty values across 200 samples for the test. The ratio produced in this manner was unitless in the same way as the coefficient of variation is unitless. This ratio is, specifically:

$$S = \frac{\text{Mean (of 55 item difficulty sampling variances)}}{\text{Variance (of 55 item difficulty sampling means)}}$$

This procedure, as outlined in Chapter 3, provided a measure of the 'separability' of the items, hence the use of the symbol 'S' to indicate the ratio just described. It is a measure of 'separability' because it expresses the mean error surrounding the estimation of item difficulty as a fraction of the total spread of item difficulties across a test. The smaller the value of S, the more clearly the position of each items difficulty value is discernible from amongst the difficulty values of the other items on the test. The ratio might well be described as a 'standardized mean item difficulty variance'. These ratio values are shown plotted for each sample design and test length in Figures 4.4, 4.5 and 4.6. The actual values used in the drawing of these figures are given in Appendix D. The four sample designs are designated, for convenience, as 'SRS-1' (Simple Random Sample), 'CLS-5' (Cluster Sample - 5 persons per cluster), 'CLS-10' (Cluster Sample - 10 persons per cluster), and 'CLS-20' (Cluster Sample - 20 persons per cluster). These designations indicate both the type of sample - simple random or clustered, and the size of the primary sampling unit - 1, 5, 10 or 20 persons.

Table 4.1 Design Effect (Deff) Values for the Variance of the Three Item Difficulty Indices when estimated by the Three Clustered Sample Designs**

Test length	Sample type	Rasch difficulty index: Sample size 40	60	80	100	120	160	200	240	320
55	CLS-5	1.10	1.11	1.06	1.08	1.03	1.02	1.03	0.99*	0.98*
	CLS-10	1.22	1.26	1.21	1.19	1.17	1.13	1.17	1.11	1.13
	CLS-20	1.58	1.52	1.50	1.44	1.50	1.48	1.46	1.37	1.39
42	CLS-5	1.10	1.14	1.06	1.09	1.05	1.04	1.01	1.00	0.94*
	CLS-10	1.23	1.29	1.22	1.24	1.23	1.16	1.16	1.15	1.09
	CLS-20	1.65	1.67	1.58	1.55	1.61	1.49	1.47	1.43	1.40
32	CLS-5	1.11	1.15	1.08	1.10	1.06	1.05	1.02	1.01	0.97*
	CLS-10	1.24	1.30	1.23	1.29	1.25	1.15	1.18	1.20	1.17
	CLS-20	1.66	1.70	1.61	1.57	1.65	1.53	1.50	1.47	1.50
		Traditional difficulty index: Sample size 40	60	80	100	120	160	200	240	320
55	CLS-5	1.34	1.44	1.29	1.33	1.20	1.20	1.15	1.04	0.87*
	CLS-10	2.02	2.03	1.90	1.78	1.89	1.95	1.89	1.71	1.52
	CLS-20	3.47	3.18	3.21	2.88	3.18	3.08	3.27	3.08	3.21
42	CLS-5	1.38	1.51	1.32	1.38	1.24	1.25	1.09	1.07	0.86*
	CLS-10	2.14	2.17	1.99	1.88	2.02	2.08	1.86	1.83	1.58
	CLS-20	3.78	3.48	3.50	3.12	3.50	3.33	3.34	3.37	3.48
32	CLS-5	1.41	1.57	1.35	1.41	1.24	1.25	1.18	1.06	0.85*
	CLS-10	2.25	2.30	2.08	1.93	2.09	2.14	2.07	1.88	1.64
	CLS-20	3.95	3.70	3.66	3.20	3.67	3.46	3.71	3.48	3.67
		z-item difficulty index: Sample size 40	60	80	100	120	160	200	240	320
55	CLS-5	1.08	1.12	1.08	1.06	1.04	1.04	1.05	1.03	0.96*
	CLS-10	1.20	1.27	1.23	1.19	1.22	1.19	1.25	1.20	1.16
	CLS-20	1.55	1.55	1.53	1.49	1.56	1.50	1.59	1.55	1.54
42	CLS-5	1.07	1.12	1.07	1.06	1.04	1.03	1.00	1.02	0.93*
	CLS-10	1.17	1.27	1.21	1.18	1.21	1.16	1.17	1.15	1.09
	CLS-20	1.49	1.54	1.50	1.46	1.52	1.44	1.47	1.45	1.41
32	CLS-5	1.09	1.12	1.10	1.07	1.04	1.04	1.03	1.04	0.97*
	CLS-10	1.19	1.26	1.23	1.23	1.25	1.16	1.23	1.21	1.17
	CLS-20	1.53	1.59	1.56	1.51	1.57	1.49	1.54	1.51	1.52

* See Appendix B for a discussion of these figures.

** See Appendix F for a warning regarding the use of this table to 'correct' variance values.

The Sample Effects

An examination of the raw item difficulty sampling variance plots (see Figures 4.1 to 4.3) for the three difficulty indices showed a number of interesting sample effects. In the case of all three difficulty indices there was a systematic reduction in sampling variance, that is, a trend towards greater stability in the estimates of the indices, as the sample size increased. In terms of the sample types used, it was clear that simple random samples produced the most stable estimates of the difficulty indices, followed by the cluster sample designs of cluster size 5 and 10 in that order. Trailing behind these and noticeably inferior as a sample design for accurately estimating item difficulty values was the cluster sample design with a cluster size of 20.

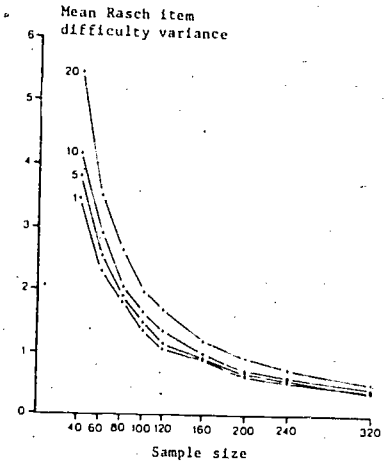
In order to consider the relative effectiveness of different sample types it was useful to apply the notion of a design effect (Deff), defined by Kish (1965) as: '. . . the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements' (Kish, 1965:258). That is, for a statistic such as the Rasch item difficulty:

$$\text{Design Effect (Deff)} = \frac{V(R)_{complex}}{V(R)_{srs}}$$

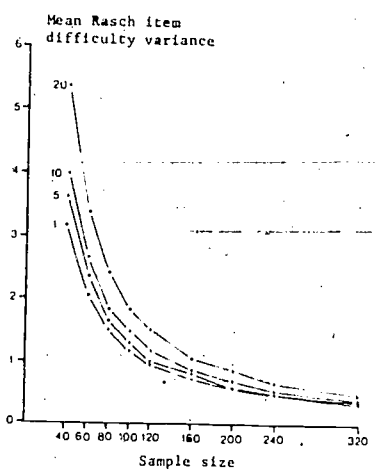
where R is the Rasch item difficulty index,
$V(R)_{complex}$ is the variance of the Rasch item difficulty for a complex, or non-simple random sample,
and $V(R)_{srs}$ is the variance of the Rasch item difficulty for a simple random sample.

This value of Deff was also calculated for the traditional and z-item difficulty indices, designated T and Z respectively. The Deff values for the three item difficulty indices under investigation are given in Table 4.1.

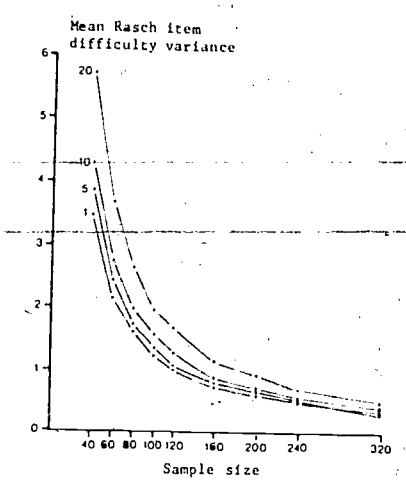
The design effect is a measure of the proportional increase in the variance of a statistic which has been derived for a sample other than a simple random one. This proportional increase, (or, in some less frequent cases, decrease) indicates and quantifies the increase (or decrease) in the error associated with the statistic being measured. Large design effect values indicate that far more caution is needed in the interpretation of statistical tests and other comparative techniques. The Rasch model of measurement is in a sense a comparative technique, as the essential element is the *difference* between an item difficulty measure and the ability measure for a person. Any increase in the variance of the item difficulty measure produces a correspondingly larger uncertainty about this difference, and is therefore associated with a decrease in the confidence with which the ensuing interpretations (and any implications) are held.



(a) Test length 55 items

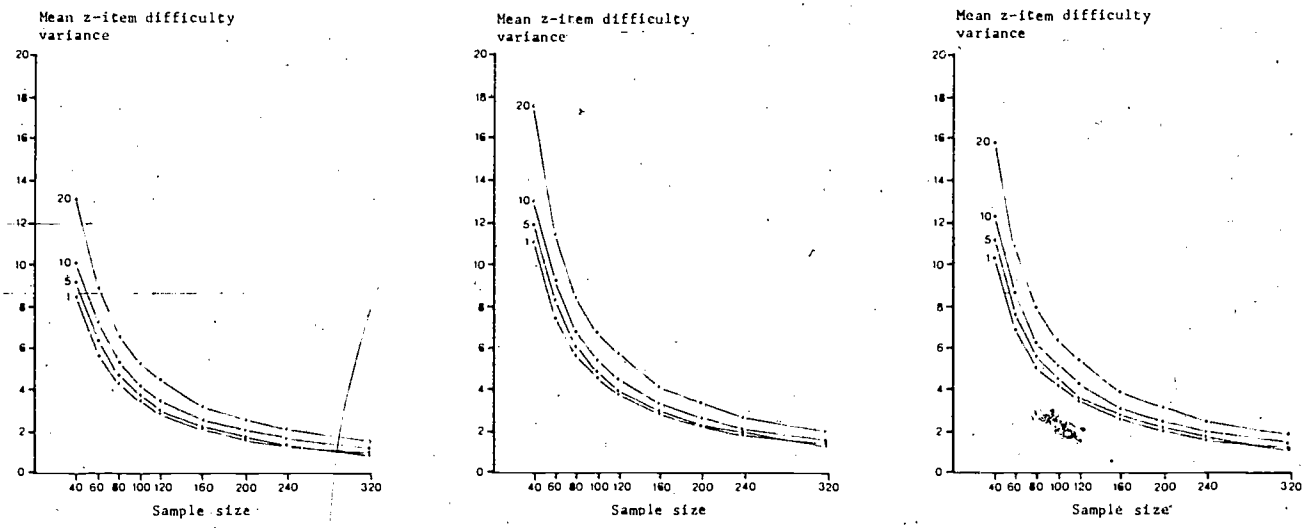


(b) Test length 42 items



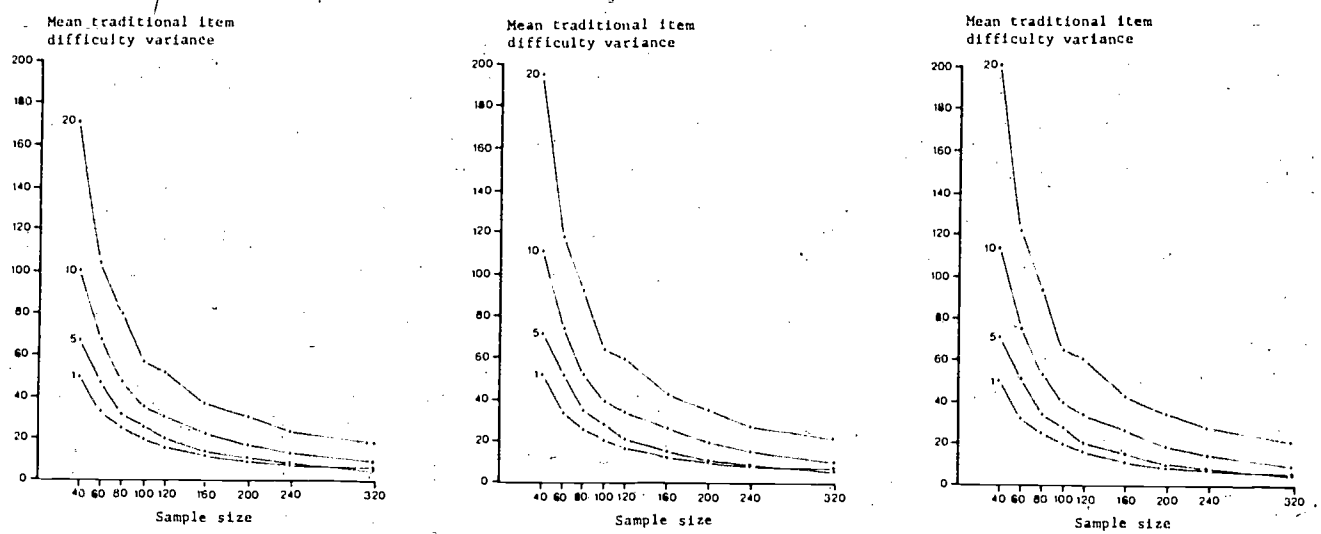
(c) Test length 32 items

Figure 4.1 Raw Mean Item Difficulty Variance for the Rasch Item Difficulty Index



(a) Test length 55 items (b) Test length 42 items (c) Test length 32 items

Figure 4.2 Raw Mean Item Difficulty Variance for the z-item Difficulty Index



(a) Test length 55 items

(b) Test length 42 items

(c) Test length 32 items

Figure 4.3 Raw Mean Item Difficulty Variance for the Traditional Item Difficulty Index

The size of the Deff values given in Table 4.1 indicate the inferiority of cluster samples compared to simple random samples for estimating accurately any of the three item difficulty indices. It is also apparent that larger clusters for a given sample size lead to greater variance of the item difficulty measures. Thus the stabilities of the item difficulty measures decrease as cluster sizes increase. The consistent exception to this statement in this study is for sample size 320, where Deff is less than unity for the CLS-5 design. This exception is a peculiarity of the sample design and population structure used in this study and the possible reasons for such an exception are provided in Appendix C.

Given that many educational surveys in which item and test statistics are examined have often used intact classes or even intact within-school year level groups as the cluster size, this finding has implications for the stability, and imputed error, of whatever parameters might have been estimated in such studies. These findings again offer strong support for the argument presented by Kish (1957:156) that in the social sciences the use of simple random sample formulae on data from complex samples remains the most frequent source of gross mistakes in the construction of confidence intervals and tests of hypotheses.

In the light of the effect of different sample types on the sampling variance of the Rasch item difficulty index it would seem clear that previous statements which attempted to suggest an appropriate sample size for stable item difficulty estimation have failed to take into account the sample types and associated design effects. Estimates which are 'stable', as defined by some measure, for one sample size may be more or less stable for other samples of the same size but of different design. The conclusions based on earlier research, and some of the arguments which ensued, are seen to have been based on the oversimplified notion that sample size was the only relevant criterion when determining the attributes of the sample necessary for stable estimates. It is also clear that stability is a relatively regular function of the sample size and type, and of the statistic in question. This means that the level of stability desired may be obtained through an examination of the relationships presented, and the subsequent selection of a sample of the best design and size to generate the required level of stability. This study does indicate that very low sample sizes, such as less than 100, yield poor estimates in terms of stability, and that for sample sizes beyond 200 the increase in stability is not necessarily economical in terms of the large number of additional sample elements (people) required to gain additional stability. This finding is in general agreement with that of Forster (1978) who, with reference to the Rasch item difficulty index, concluded that for sample sizes less than 200 the accuracy of the item difficulty dropped considerably, and that as the sample size increased beyond 200 the increase in accuracy was not substantial. These statements do, of course, need to be

tempered by design effect considerations, for if economy permits, a better design may improve accuracy more readily than a large increase in the number of sample elements.

The present study also indicates that simple random samples generate the most stable estimates, and that for cluster samples *small* clusters generate more stable estimates than large clusters for a given sample size. Of particular concern is the noticeable relative instability for clusters of size 20 (CLS-20 design). As suggested earlier, it is commonplace for educational surveys to sample from class or year-level clusters which are considerably larger than 20, and, as such, some concern arises from the observation that for such samples the traditional error values may grossly underestimate how large the variance of the item difficulty estimates might actually be. Table 4.1 suggests by extrapolation that the Deff values of such samples would be at least 1.5, and as such, samples of one and a half times the given size would be required to generate estimates as stable as a simple random sample of given size.

Comparison of the Three Difficulty Indices

So far only the raw item difficulty variance measures have been considered. Cornish (1983) sought to compare the stability of two types of difficulty index, the Rasch and the traditional indices. This study incorporates a third, the z-item difficulty index. The z-item difficulty index is obtained by calculating the traditional difficulty values across all items on a test and then converting the traditional values by linear transformation in such a way that the new difficulty values have a fixed mean and a fixed standard deviation.

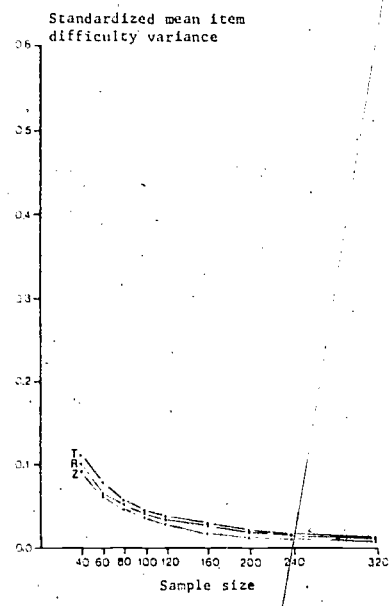
Each of the three indices being examined in this study, was measured using a different metric. The Rasch item difficulty was transformed by multiplication by 4.5512 and subsequent addition of 50 units. The traditional difficulty values were expressed as percentages. The z-item difficulty comprised traditional values which were transformed so that the mean difficulty across the test was 50 units and the standard deviation of difficulty values was 15 units. These procedures meant that at all times there was no occasion when any of the three indices had a value lower than zero or greater than 100. Although these procedures made the three indices appear similar, in order to compare these three difficulty indices in a meaningful way, it was still necessary, as indicated earlier, to place them on some form of common metric. For this purpose the following procedure was adopted for each of the three item difficulty measures.

The mean difficulty values for each item over 200 samples, for each combination of sample size, sample design and test length were calculated. The variance of these mean item difficulty values across the test length is directly related to the metric used to determine the difficulty values. If each item difficulty value, or if the mean of the sampling variance of the item difficulty values is divided by this variance across the test of mean item difficulty values then the resulting ratio is freed from the units of

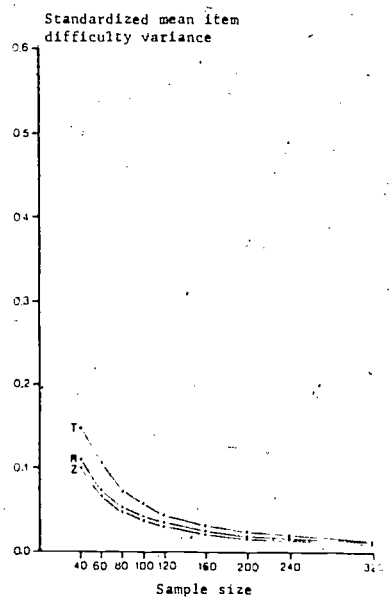
measurement. The measure obtained is equivalent to the raw mean item difficulty sampling variance which would be obtained if the variance, and standard deviation, of the test items across the test were made equal to unity. This ratio is also a measure of the error variance associated with item difficulty values expressed as a fraction of the spread of the item difficulty values across a test. This 'standardized mean item difficulty variance' is thus a measure of how clearly defined is the position of the difficulty of an item from amongst the difficulty values of the other items on the test. This ratio is a measure of the ability of any form of item difficulty index to indicate difficulty values which are clearly differentiated from the other items of a test. In addition, this ratio, in so far as the variance of the item difficulty values is low, is also a measure of item stability. It is suggested that the term 'item separability variance ratio' may be an appropriate label for such a measure. One interesting feature of this variance ratio was that the denominator was extremely stable in value across samples of different size and design. This therefore meant that the shape of the plotted standardized mean item variance ratio values was the same as the shape of the plotted raw item variance values. The advantage was that the three different difficulty indices could be plotted on the same axes. These relationships have been presented in Figures 4.4, 4.5 and 4.6. The labels 'T', 'R' and 'Z' were used to indicate the plots for the traditional, Rasch, and z-item difficulty indices respectively.

From the graphs of the standardized mean item variance ratio a number of features are evident. The first is the substantial inferiority of the traditional item difficulty measure as the cluster size increases. Given that most educational surveys use a clustered sample design, this inferiority of the traditional measure suggests that it is unsatisfactory as a measure of item difficulty when compared to the other two item difficulty indices.

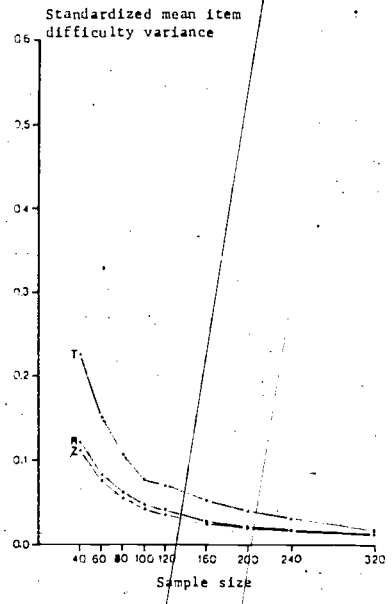
One reason for the instability of the traditional measure compared to the other two difficulty indices is that the traditional index has no fixed mean value. Both the Rasch and the z-item difficulty values are constrained to a fixed mean (of 50 in this study) for each occurrence of calibration, that is, for each sample. This is not the case for the traditional index. Consequently, the mean item sampling variance for the traditional index has an additional component associated with the average difference of the mean traditional difficulty values between samples. This component does not arise for the other two indices. As the sample size increases this difference between samples will tend to become less, and as a consequence this additional component of the variance will also be reduced, meaning that larger samples are less affected by this component. It should be noted that this explanation is consistent with the trends found for the traditional index in Figures 4.4 to 4.6, where the greatest instability of the traditional index compared to the other two indices is exhibited at the smallest sample sizes.



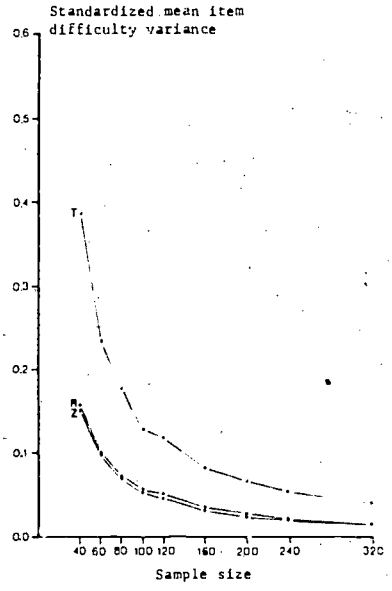
(a) Sample design SRS-1



(b) Sample design CLS-5



(c) Sample design CLS-10



(d) Sample design CLS-20

Figure 4.4 Standardized Mean Item Difficulty Variance of the Traditional (T), Rasch (R) and z-item Difficulty (Z) Indices, on the 55 Item Test

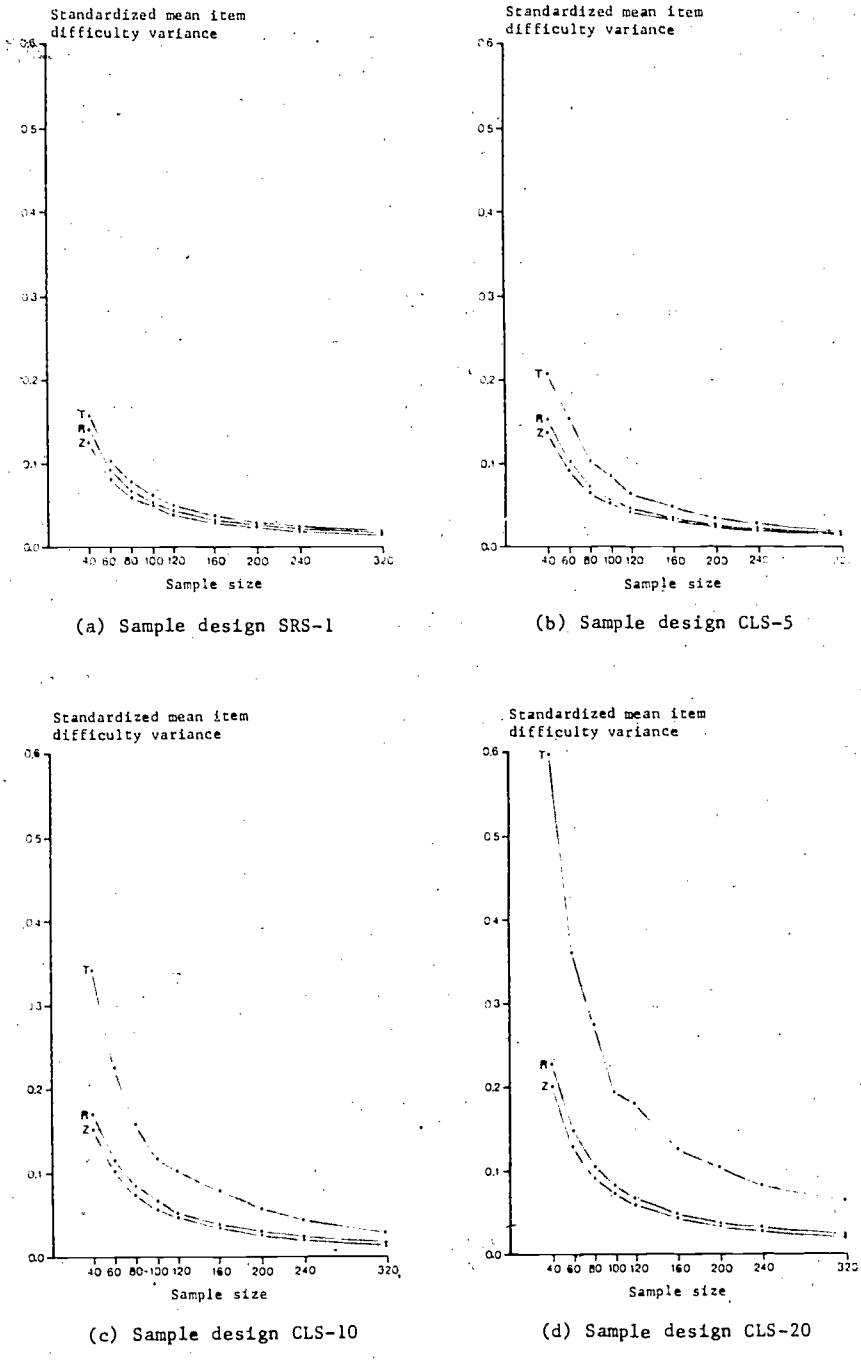
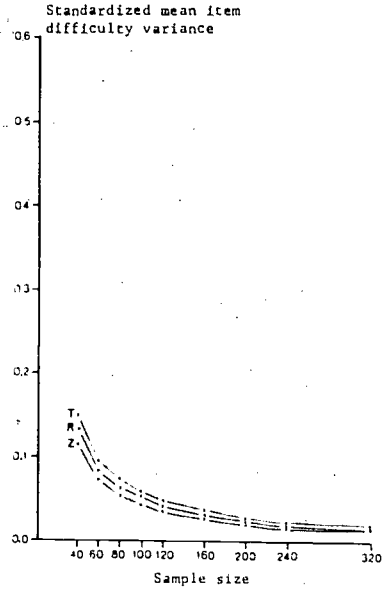
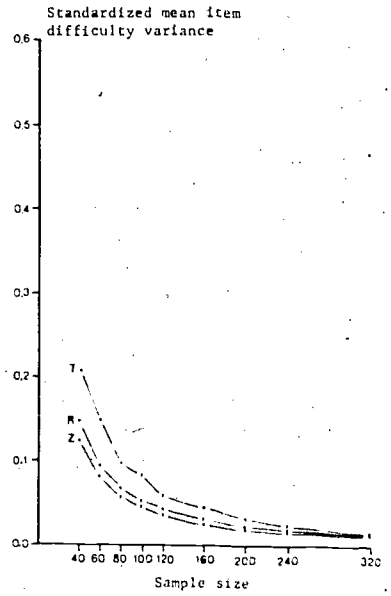


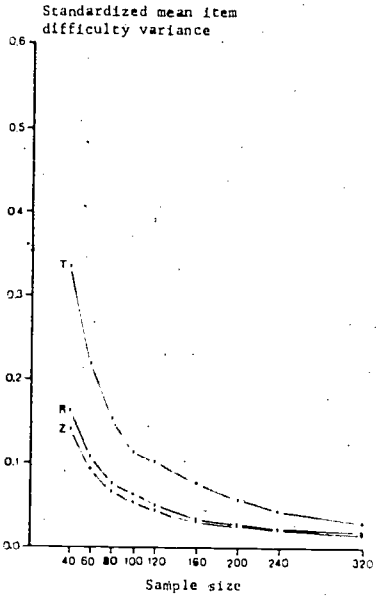
Figure 4.5 Standardized Mean Item Difficulty Variance of the Traditional (T), Rasch (R) and z-item Difficulty (Z) Indices, on the 42 Item Test



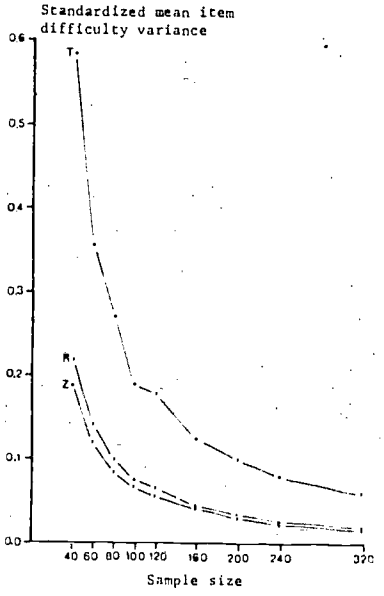
(a) Sample design SRS-1



(b) Sample design CLS-5



(c) Sample design CLS-10



(d) Sample design CLS-20

Figure 4.6 Standardized Mean Item Difficulty Variance of the Traditional (T), Rasch (R) and z-item Difficulty (Z) Indices, on the 32 Item Test

The second feature of note is the obvious consistency in the order of size of these standardized mean item variance ratios for any given combination of sample size, sample design and test length. The traditional difficulty index produces the highest value, and the z-item difficulty index produces the lowest. Within this study there were no exceptions to this phenomenon. This feature, however, contradicts the findings of Tinsley and Dawis (1975) that the z-item difficulty index was less stable than the Rasch item difficulty index. Associated with this comparison of the Rasch and the z-item difficulty indices is the additional and obvious feature shown in Figures 4.4 to 4.6, namely that these two difficulty measures are closely related at all times. The consistent superiority of the z-item difficulty indexing in terms of stability as measured by the standardized mean item variance is countered by the very small size of the improvement in stability obtained through the use of the z-item difficulty index rather than the Rasch item difficulty index. This leads to questions regarding the usefulness of the z-item difficulty index compared to the Rasch difficulty index. Indeed, both indices owe much of their stability to the fixing of the mean item difficulty at the time of calibration. Certainly, at present, the z-item difficulty index is far simpler to calculate, but it does not have the probability model features of the logistic model. In addition, the possibility of a common scale for item difficulty and person ability does not arise with the z-item difficulty index. History may prove that the Rasch item difficulty index is superior for reasons of practicality and usefulness, even though it is slightly inferior in terms of item stability as indicated by the standardized mean item variance ratio.

The Underlying Structure of Variability

Up to this point this study has investigated these difficulty indices in relation to the findings of earlier studies such as those of Whitely and Dawis (1973), Forster (1976), Wright (1977) and Cornish (1983). It would be useful if the major part of the variability of such complex measures as these difficulty indices could be explained by features of the calibration samples. To this end the standardized mean item variance ratio was further modified. One additional transformation was made.

To understand this, let us first consider the sampling variance of a statistic such as the mean.

$$V(\bar{X}) = \frac{N - n}{N - 1} \quad \frac{V(X)}{n}$$

where $V(\bar{X})$ = sampling variance of the mean,

$V(X)$ = variance of X across the population,

N is the population size,

n is the sample size,

and $\frac{N - n}{N - 1}$ is the finite population correction.

This equation may be transposed to give:

$$V(X) = V(\bar{X}).n.\frac{N-1}{N-n}$$

In this study has been measured the sampling variance of the item difficulty, the term equivalent to $V(\bar{X})$. We also know N and n for each sample size and design. For the term $V(\bar{X})$ we have substituted the measured standardized mean item difficulty variance. The resulting value, in the place of $V(X)$, is difficult to describe algebraically. However, conceptually it is a measure of the underlying variability of a particular item difficulty index as measured using a particular sampling design. For this reason it has been labelled the 'structure value'. The degree to which it remains constant across sample size indicates how well the above equation may be applied to explain the total variability of the item difficulty index. If the structure factor is constant with relation to other variables then the above equation is adequate in explaining the variability of the item difficulty index in terms of sample size, sample design, finite population correction and the structure value. The use of the standardized mean item difficulty variance, rather than the raw item difficulty sampling variance, allows comparisons to be made between the three different indices.

The equation:

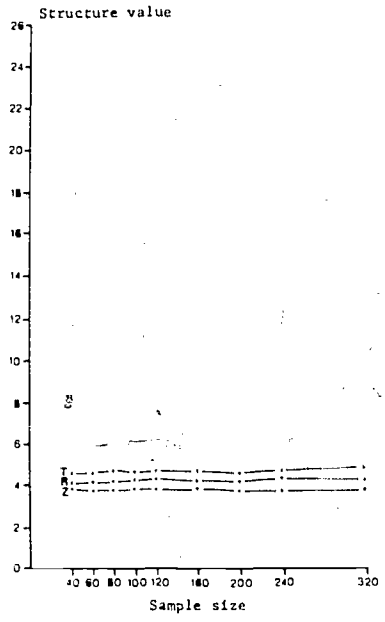
$$F = S.n.\frac{N-1}{N-n}$$

where F is the structure value,

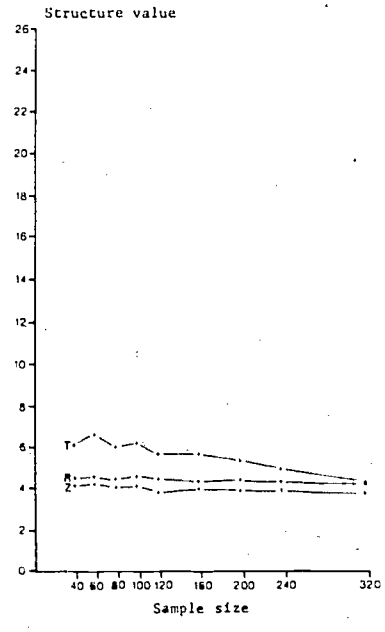
and S is the standardized mean item difficulty variance,

is derived directly from the earlier formula for $V(X)$ and is based therefore on the assumption of simple random sampling. The degree to which increases in the standardized mean item variance cause an increase in the structure value correspond to the degree to which non-simple random designs cause an increase in the sampling variance. As such, the structure value for simple random samples may be considered to be the base against which the structure values for other sample designs are measured. The structure values are plotted in Figures 4.7 to 4.9. The values plotted in these Figures are given in Appendix D.

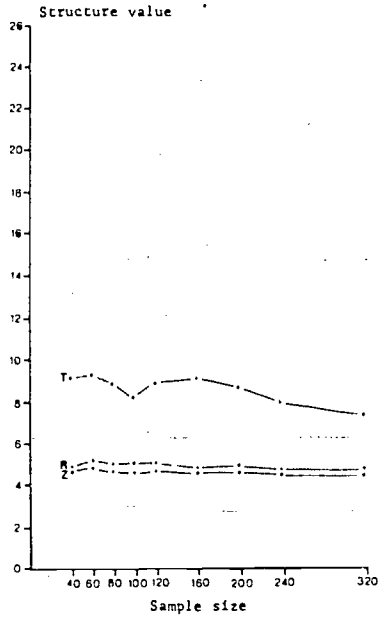
These graphs of the structure values indicate some useful features of the sampling variability of the three item difficulty indices. The horizontal nature of these plots suggest, at least for the Rasch and z-item difficulty indices, that the major contributing factors to the variability have been accounted for. It seems that the finite population correction and the relationship to sample size are the two most important single effects determining the variability of item difficulty measures. The traditional difficulty value is again somewhat different. In this case there appears to be an interaction between the sample size and the cluster size which is particularly apparent in the CLS-5 design. In all cases, the lowest structure values are, in general, found for the simple random design,



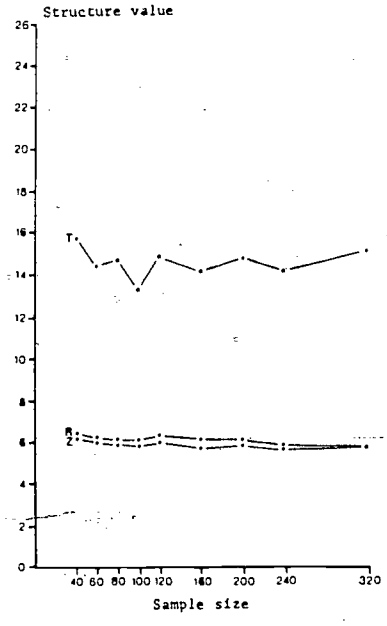
(a) Sample design SRS-1



(b) Sample design CLS-5



(c) Sample design CLS-10



(d) Sample design CLS-20

Figure 4.7 Structure Values for Rasch (R), Traditional (T), and z-item Difficulty (Z) Indices on the 55 Item Test

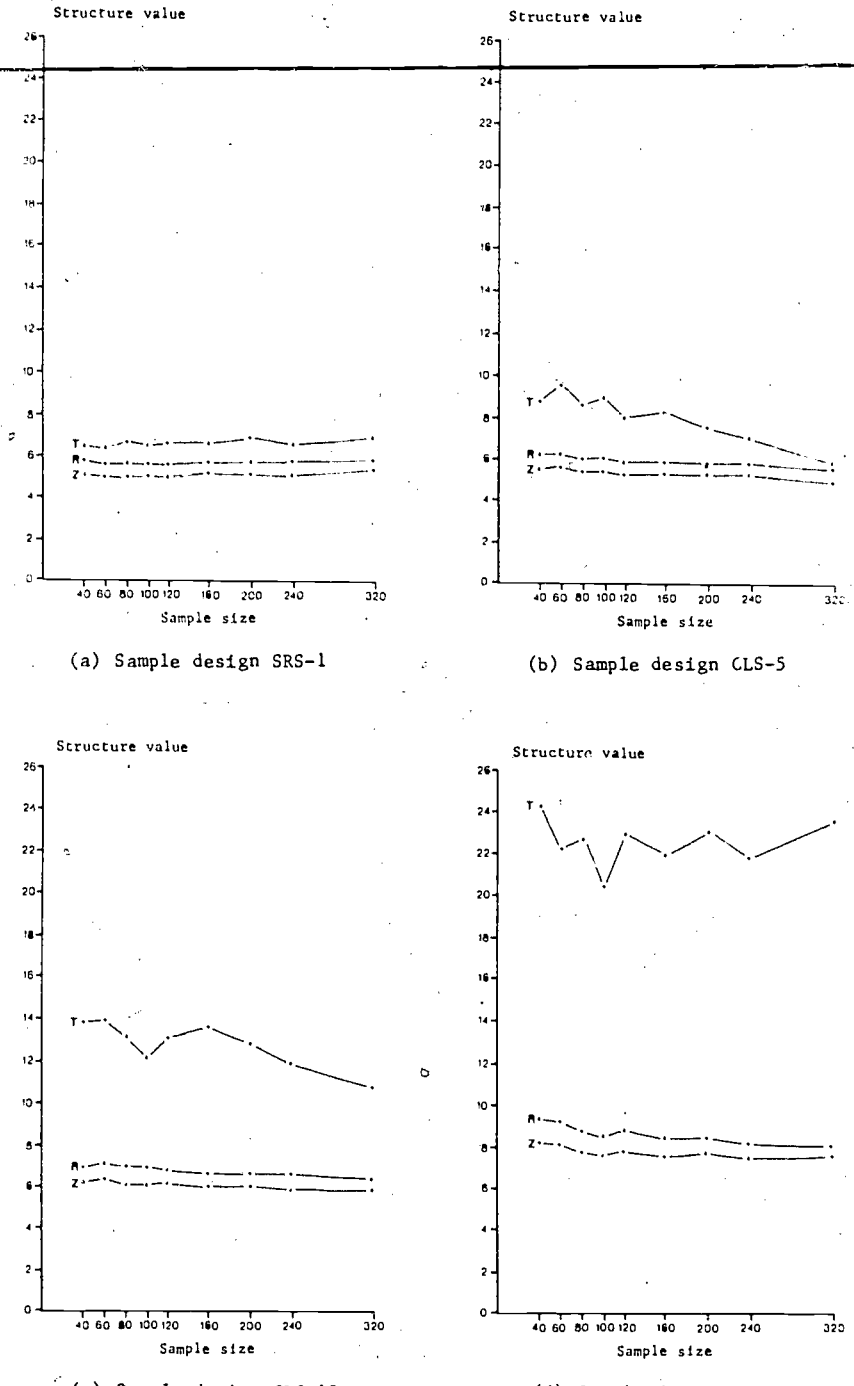
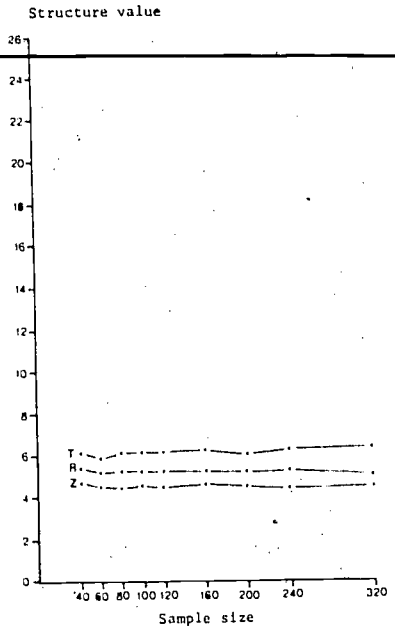
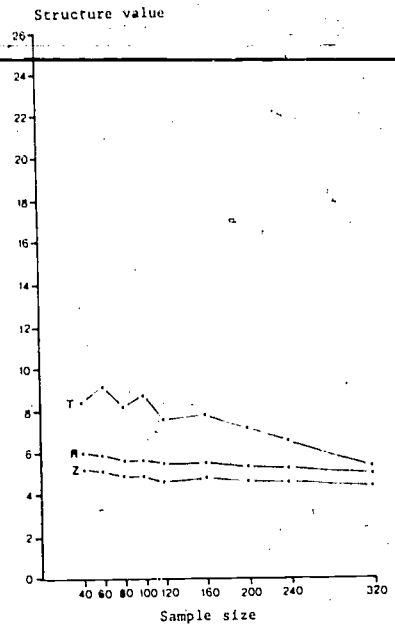


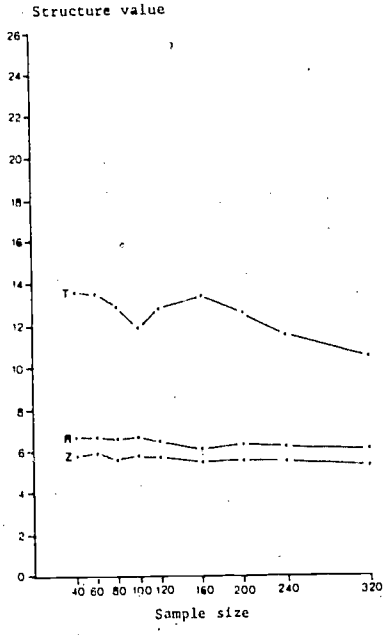
Figure 4.8 Structure Values for Rasch (R), Traditional (T), and z-item Difficulty (Z) Indices on the 42 Item Test



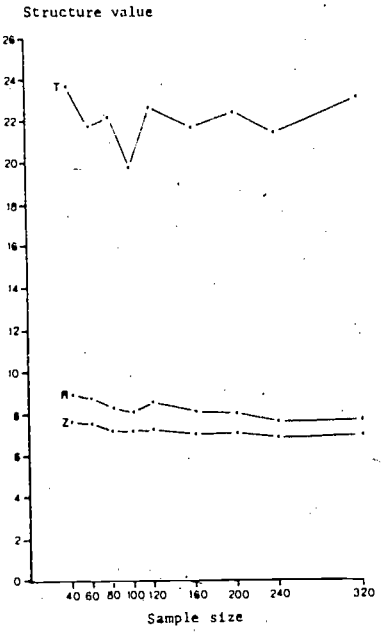
(a) Sample design SRS-1



(b) Sample design CLS-5

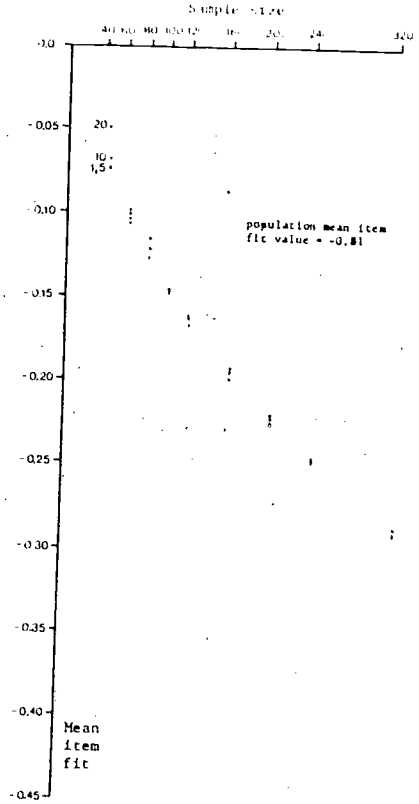


(c) Sample design CLS-10

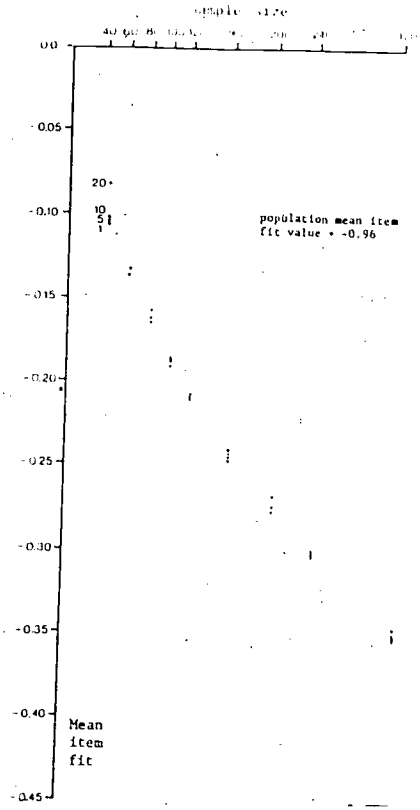


(d) Sample design CLS-20

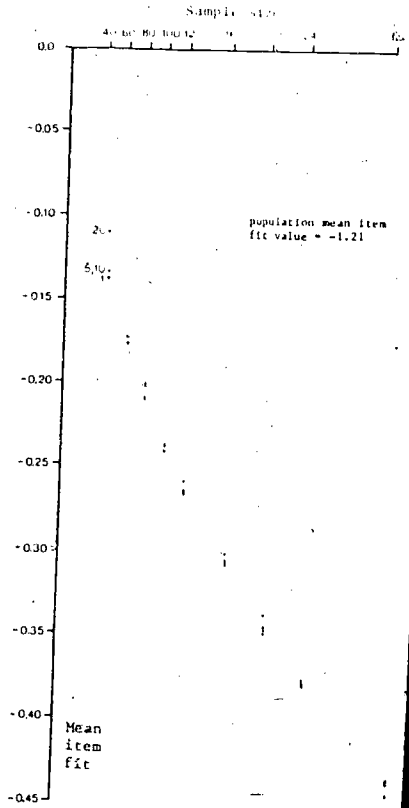
Figure 4.9 Structure Values for Rasch (R), Traditional (T), and z-item Difficulty (Z) Indices on the 32 Item Test



(a) Test length 55 items



(b) Test length 42 items



(c) Test length 32 items

Figure 4.10 Mean Rasch Item Fit Values for the Three Test Lengths

and as the cluster size increases, so too the structure values increase. The variability also depends upon the sample design and upon the length of the test, which in this situation is related to the fit of the items. However, there is no obvious way of accounting for this in a quantitative manner. Here too, the parallel nature of the behaviour of the Rasch and z-item difficulty indices is apparent, with even some degree of association evident in the smallest fluctuations from the horizontal.

Summary

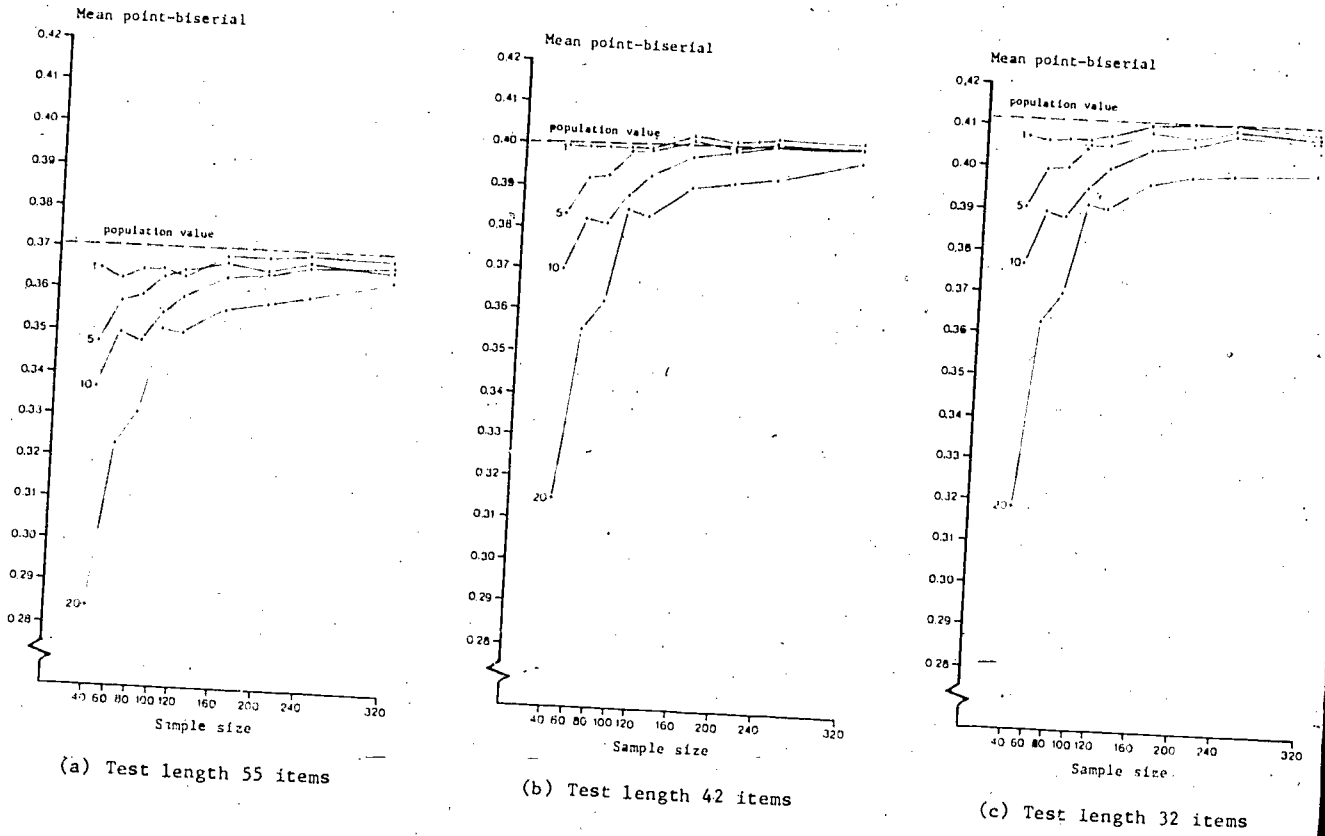
In the first phase of the investigation reported above it was found that the sampling variance of the item difficulty values was clearly related to sample size in a systematic and quantifiable manner. This is consistent with the claims of earlier researchers. The design effect was also a major contributing factor to item difficulty sampling variance when non-simple random samples were used. The traditional difficulty index appeared to be inferior to the other two indices in many respects. The standardized mean item difficulty variance of the z-item difficulty index indicated that it is marginally superior to the Rasch index. However, other considerations make this advantage small. The close correspondence in behaviour of the Rasch item difficulty index and the z-item difficulty index also suggested that there was little difference for all practical purposes between the two.

Investigation 2: The Effect of Sampling on Item Fit and Point-biserial Discrimination Values

Forster (1976) has suggested that Rasch item fit values increased as the sample size increased, but that point-biserial values remained the same. To test this contention, plots were made of the mean Rasch item fit values for the items of a test and for the mean point-biserial values. These results are presented in Figures 4.10 and 4.11 respectively.

The Effect of Sample Size on Item Fit Values

Figure 4.10 indicates quite clearly that there exists a distinct relationship between sample size and the fit values of the items. The fact that all the mean item fit values plotted were negative was indicative of the good fit of the items as a group to the Rasch model, perhaps better than might have been expected for a test originally constructed using traditional procedures. The mean item fit values increased rapidly (in the negative direction) as the sample size increased. However, two effects are apparent with respect to test length. Since the estimation of item fit values requires candidates, the fewer the number of candidates the more the item fit values tend towards zero. Conversely, the 'population' values for mean item fit, which are recorded in Figure 4.10, are well above even the mean item fit values for the largest sample size of 320. It would seem that



(a) Test length 55 items

(b) Test length 42 items

(c) Test length 32 items

Figure 4.11 Mean Point-biserial Discrimination Values for the Four Sample Designs

whereas item difficulty values can be estimated relatively accurately with small samples, say around 200, and only slightly less accurately with smaller samples, the estimation of item fit values becomes markedly weaker as the sample size decreases, and quickly becomes stronger as the number of subjects increases. This is not unexpected. The analysis of item fit is dependent upon the response patterns of candidates. The ability to detect unusual, or mis-fitting items, requires accurate information about the expected behaviour of individual items, and such information cannot be gained with confidence from a small number of subjects. Wright and Stone (1979:74-77) have stated that the item fit statistic follows a t-distribution with degrees of freedom equal to one less than the number of candidates. For smaller sample sizes this test becomes less powerful for detecting extreme, or discrepant, item fit values. This means that as the sample size decreases the fit values will also decrease in absolute size. As the sample size is increased towards the population size, so too the item fit values move closer to the 'population' values.

It should be noted that the 'population' value of the mean item fit is somewhat different from many other population statistics, such as, for example, the population mean or the population variance for some parameter. If it were possible to increase the size of the population, as might be simulated, for example, by counting each person twice, most population statistics such as the mean and variance, would remain constant. This is not true for fit however, because, as mentioned above, the fit value is dependent upon the number of subjects. This means that the notion of a 'population' value for the mean item fit is less absolute than for most other statistics.

Thus, in general agreement with Forster (1976), it is apparent that as sample size increases, so too the item fit values increase from small values, in absolute terms, towards the population item fit values, whether these population values are positive or negative. It is also of particular interest to note that the sample design had a negligible effect upon the estimation of item fit values, with any effect which might exist being most apparent for small sample sizes. It should be noted, however, that the amount of fluctuation in fit values between similar samples has not been considered as part of this study.

The Effect of Sample Size on Point-biserial Discrimination Values

For each sample size and design, and each test length, the mean point-biserial item discrimination value was calculated. These values are plotted in Figure 4.11, where the empirically determined relationship between sample size and the mean point-biserial discrimination value is seen. The population value of the mean point-biserial discrimination index is also shown for each test length.

In contrast to item fit estimation, the sample estimates quickly approach the population values. It would seem that, unlike item fit estimation, the major factor in

determining the size of estimated point-biserial discrimination values is not the sample size, but the interaction between sample size and sample design. The simple random samples estimated the point-biserial discrimination values well, even at the smallest sample size of 40, whereas this accuracy was not achieved by the cluster sample with 20 students per cluster (CLS-20 design) even at a sample size of 320. The statements of Forster concerning the consistency of the point-biserial discrimination value over varying sample sizes is only true for simple random samples, and not true for the cluster samples examined in this study. A possible explanation of this effect would appear to be associated with the consistency of class groups in responding to items. The formula[1] for the point biserial is given by

$$r_{pbi} = \frac{\bar{X}_1 - \bar{X}_0}{s_x}\sqrt{pq}$$

where $\bar{X}_1$ is the mean test score of students who were correct on the item,
$\bar{X}_0$ is the mean test score of students who were incorrect on the item,
s_x is the test standard deviation,
p is the proportion of students answering the item correctly,
and q is the proportion of students who answered the item incorrectly
(i.e. $p + q = 1$).

The $\sqrt{pq}$ component incorporates the traditional difficulty of the item. Investigation 1 has shown that the traditional difficulty value is the one most susceptible to sample size and design effects, and the major effect is a noticeably larger sampling variance for the traditional difficulty index under conditions of a combination of complex sample design and smaller sample size. These are exactly the same conditions which are associated with low mean point-biserial discrimination values. It seems likely that the more extreme values of the traditional item difficulty index which occur under such conditions cause a reduction in the $\sqrt{pq}$ value, thereby decreasing the associated point-biserial discrimination values.

Of note also was the general increase in mean discrimination value as the test length was reduced. This is an artifact of the procedure whereby, in general, the items which were deleted were non-fitting items, and many did not fit because of poor discrimination.

Summary

Item fit increases, in absolute value, as the sample size increases. However, there appears to be a reason why this increase should be expected, namely the increased power

---

1 Note that $\bar{X}_1$ and $\bar{X}_0$ are calculated on the test excluding the item in question, so that r_{pbi} is an unbiased value.

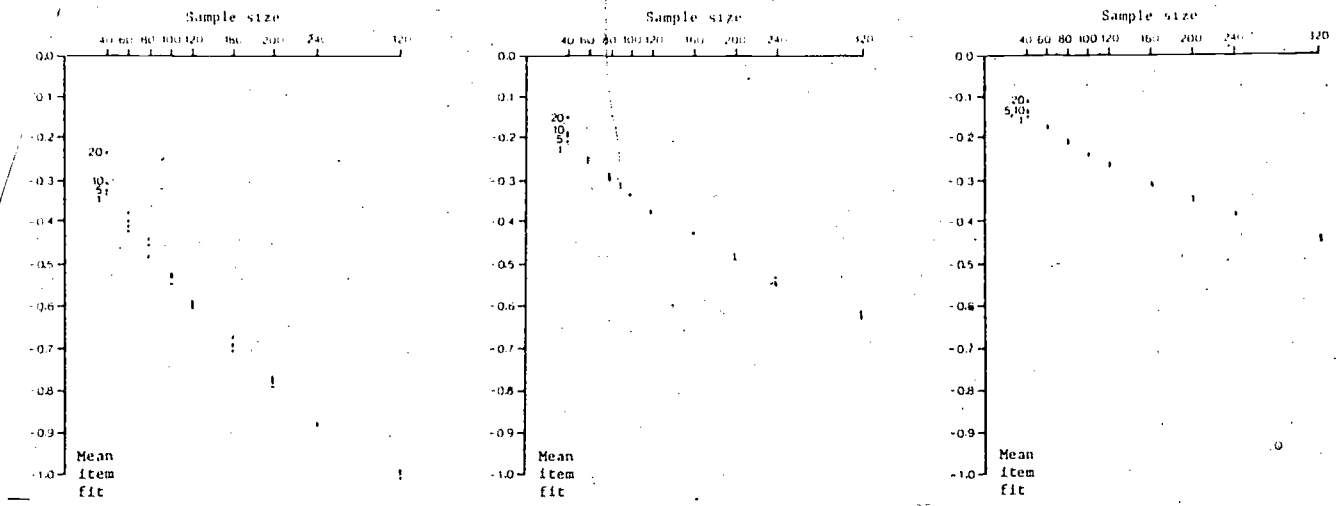
of the model to detect inconsistent item behaviour through the increased degrees of freedom as the sample size become larger. Although Forster (1976) was correct in this respect, his statements concerning the constancy of point-biserial discrimination values across different sample sizes are seen to be true only for simple random samples. For clustered samples the mean point-biserial discrimination decreased as the sample size decreased.

Investigation 3: The Effect of Deleting Items which do not Fit the Rasch Model on Item Fit and Item Variance

Items which did not fit the Rasch model well initially were deleted from the 55 item test to produce a test of 42 items, which were deemed to satisfy the Rasch calibration procedures. From these 42 items, further items with poor or extreme item fit values were removed to produce a test of 32 items. The items deleted and the reasons for doing so are presented in Appendix F. The effect of these procedures on the item fit values and the item variance of the core of 32 items was then investigated.

The Effect of Item Fit

The examination of item fit is a comparative process. That is, the fit of any one item is dependent upon the characteristics of the other items around it. This occurs because the latent trait against which item fit values are calculated can be viewed as corresponding approximately to the first, or principal component of a factor analysis when applied to the items of a test. Thus it is the complete group of items, namely the combined item characteristics which define the latent trait. Each item contributes to this trait in part, and thus one component of an item fit value involves a comparison between that item and the other items which comprise the test. The other component of the fit value is associated with a comparison between the Rasch model and the item, and involves a quantification of the degree to which the item conforms to the Rasch model. That the item group in this study does fit the Rasch model well is shown by Figure 4.10, where the mean item fit is negative, indicating good fit. Because the items which were deleted in the first reduction of test length were the ones not truly appropriate for Rasch calibration, that is, because they were items which did not fit the Rasch model well, the effect was, in general, to take away from the mean fit value those items which had poor, or, high positive fit values. This meant that those remaining had a better, or more negative, mean fit value. Similarly, it was again non-fitting items which were deleted from the 42 item test to produce the 32 item test. Again the same effect applied, and so the mean item fit value became more negative. This effect is seen in Figure 4.10, where the mean item fit becomes better (more negative) as the test length is reduced by the deletion of poorly fitting items. The consequences of this reduction on the core of 32 items are shown in Figure 4.12.



(a) Test length 55 items

(b) Test length 42 items

(c) Test length 32 items

Figure 4.12 Mean Rasch Item Fit Values for the Core of 32 Items for the Tests of Different Length

Again it is clear that as the sample size increases the fit values increase in absolute terms. However, the effect of reducing the test length by deleting poorly fitting items on the core of 32 items is to *worsen* their fit values, that is, to make them more positive. Wright and Stone (1979:80) point out that as misfitting items are removed from a test the fit values of those items which remain will tend to become worse, that is, more positive, particularly in the case of those items which do not fit as well as most others. It would seem from the trend in Figure 4.12 that this is consistently the case. (It should be noted that Figure 4.10(c) and Figure 4.12(c) are identical, but on a different vertical scale.)

While, in general terms, the deletion of misfitting items improves the mean fit value of the test, which is to be expected, contrarily it also makes worse the mean fit value for those items which remain, or for any particular subgroup of well-fitting items amongst those remaining.

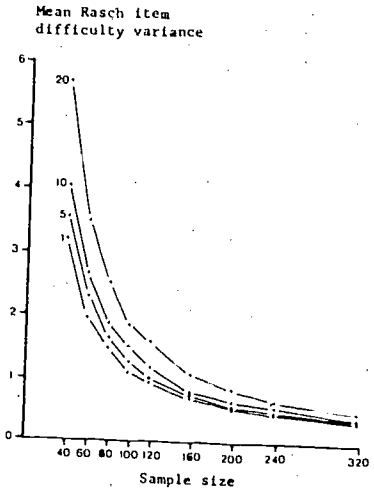
The Effect on Rasch Item Variance

It may be of more value to know the effect of the deletion of poor-fitting items on the stability of the difficulty index than to know the effect on fit values. When the tests are taken as a whole, as shown in Figure 4.1, the effect, even though slight, is apparent. The effect on the item variance of the reduction of items from 55 to 42 and then to a core of 32 items is seen in Figure 4.1 (with actual values being given in Appendix D, Table D.1). The initial effect of the reduction of test length of 55 items to 42 by deleting those possibly unsuitable to Rasch calibration was to decrease the mean item variance. However, the subsequent deletion of further items on the basis of mis-fit, which reduced the test from 42 items to 32 items, has actually increased the mean item variance back to the same values, if not higher, than for the 55 item test. The small general increase in the variance of the 32 core items as the test length is reduced should be noted in Figure 4.13.

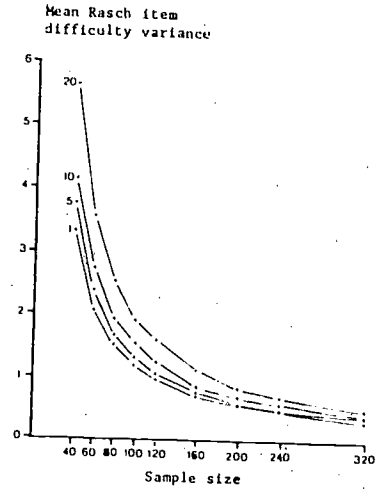
As far as item stability is concerned, it would appear that the mean Rasch item difficulty variance may be reduced by an initial deletion of inappropriate or poorly fitting items, but that any further deletion produces an increase in the item variance which counters the gain in stability obtained through the initial exclusion of poor items.The size of these effects was however, quite small, and as such was not of great consequence in this study.

Summary

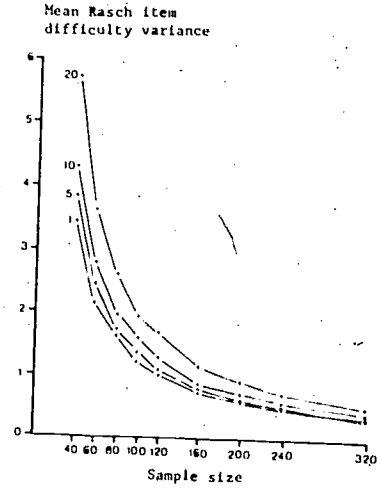
The features of the items comprising a test, in general, act to influence the Rasch item fit and Rasch item variance. Exclusion of items which do not fit well causes fit values to improve for the test as a whole, whilst actually making worse the fit values for the core of items remaining. Item variance is also affected by the deletion of poor items.



(a) Test length 55 items



(b) Test length 42 items



(c) Test length 32 items

Figure 4.13 Mean Rasch Item-Variance for the Core of 32 Items for the Tests of Different Length

The variance of the core of remaining items is not substantially affected. However, as a whole, the mean item variance for a test is reduced after only a first deletion of poorly-fitting items. Subsequent deletion of more items causes the gain in stability obtained from the first deletion of items to be lost.

Investigation 4: Measurement of the Rasch Errors

This investigation covers two aspects of the measurement of the standard error of the Rasch item difficulty index. The first is its appropriateness, or accuracy, as compared with the empirical measures available from this study. The second aspect is a brief discussion of the relationship between the standard error and the size of the calibration sample.

The Accuracy of the Standard Error

The standard error measure considered here is the one defined by Wright and Stone (1979) and produced by the computer program BICAL (Wright and Mead, 1977), namely an asymptotic estimate of the standard error of the maximum likelihood difficulty estimate.

For each item the variance of the Rasch item difficulty was calculated and converted to a standard deviation. Also for each item the mean of the standard error measure across the 200 samples was also obtained. The ratio of this standard error to the standard deviation was calculated for each item for all the different combinations of test length, sample size and sample design. Within each such combination the ratio was summarized by descriptive statistics, tables of which are found in Appendix E, Tables E.1 to E.12. To reduce these tables to a form in which they could be readily comprehended the two most extreme values found in Tables E.1 to E.12 are given in Table 4.2. It should be noted that if persons wish to use the above table (4.2) and the information in Appendix E to 'correct' values of the Rasch Standard Error, that the values in these tables should be *divided into* the values output by programs such as BICAL.

The first feature noted in Table 4.2 is that the ratio (as expressed in the 'mean' column) is indeed close to unity, with systematic deviations from unity according to the sample type. Thus it appears that the standard error is a good estimator of the variance of item difficulty values for items in general. In order to determine how much variation of this ratio was occurring across items, the standard deviation of the ratio values across the test length was also calculated. From the standard deviation values in Table 4.2 it was clear that these ratio values had a very narrow distribution, the maximum standard deviation over all combinations of test length, sample size and sample type being 0.144 and the minimum being 0.058, with typical values for the standard deviation of the ratio being of the order of 0.1. The narrowness of these distributions indicated that, in

Table 4.2 Distributional Attributes of the Ratio of Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty

No. of items	Sample design	Ratio distribution			
		Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
55	SRS-1	1.02-1.09	0.066-0.098	0.85-0.93	1.18-1.46
	CLS-5	0.97-1.11	0.066-0.097	0.69-0.94	1.16-1.37
	CLS-10	0.93-1.03	0.086-0.128	0.56-0.71	1.12-1.52
	CLS-20	0.86-0.95	0.113-0.144	0.40-0.48	1.08-1.39
42	SRS-1	1.01-1.08	0.063-0.074	0.88-0.97	1.13-1.29
	CLS-5	0.97-1.12	0.066-0.083	0.68-0.95	1.11-1.33
	CLS-10	0.94-1.05	0.084-0.103	0.56-0.70	1.06-1.19
	CLS-20	0.85-0.95	0.110-0.125	0.40-0.48	1.00-1.20
32	SRS-1	1.01-1.12	0.058-0.072	0.89-1.01	1.12-1.24
	CLS-5	0.97-1.14	0.064-0.084	0.68-1.01	1.11-1.32
	CLS-10	0.94-1.05	0.093-0.106	0.56-0.70	1.07-1.19
	CLS-20	0.85-0.96	0.122-0.139	0.40-0.48	1.02-1.19

general, the ratio did not differ from unity by a substantial amount, thus confirming as appropriate the use of the standard error parameter to indicate the error associated with the Rasch item difficulty index.

Further investigation of the trends in Table 4.2 showed that as the sampling method became more clustered, that is, as clusters became larger, the standard error tended to underestimate the true error associated with the difficulty estimates. For simple random samples the standard error slightly overestimated the true error. Again, the effect of different sample designs upon the variance of the Rasch item difficulty measure and therefore the standard error are seen. The larger cluster size is associated with an underestimation of the true error by the standard error estimate. This is consistent with the earlier findings where larger cluster sizes resulted in an increase of the variance of the difficulty index, as measured by the design effect (Deff) shown in Table 4.1, where large clusters are associated with high Deff values, the result of them having increased the variance of the item difficulty value.

Finally, an examination of the complete tables of distributional attributes of this ratio of standard error to standard deviation (Tables E.1 to E.12 in Appendix E) shows that for all sample types except the simple random samples the ratio increased slightly as the sample size increased. This systematic increase, although slight, is largely eliminated if the finite population correction is applied to the standard deviation of the difficulty values (the denominator of the ratio).

The Inverse Relationship

Earlier researchers, such as Wright (1977) have contended that there is a simple inverse relationship between the square of the standard error (or the sampling variance) of the Rasch item difficulty index and the size of the calibration sample.

The first investigation in this chapter has indicated clearly the very close relationship between the variance of the Rasch item difficulty index and the size of the calibration sample. This relationship is just as Wright had contended, that there was a simple inverse relationship between the item variance and the sample size. The earlier part of this fourth investigation has illustrated the appropriateness of the standard error as a measure of the square root of the sampling variance of the item difficulty. These two clear and inter-related findings combine to confirm Wright's contention.

Summary

The Rasch standard error parameter was found to be appropriate as a measure of the true error of estimation as calculated from the square root of the empirically determined sampling variance of the item difficulty index. This finding, coupled with the simple inverse relationship between sample size and item variance, as discussed in Investigation 1, shows the relationship between the Rasch standard error and the sample size to be likewise simple; namely, that there is a simple inverse relationship between the square of the standard error and the size of the calibration sample. Systematic deviations in the ratio of the standard error to the true error are explained by the finite population correction and the design effect, both of which are discussed in Investigation 1.

CHAPTER 5

ISSUES IN THE CALIBRATION OF TEST ITEMS

Introduction

In this study several issues relating to the calibration of items in educational tests were investigated. Primarily, the study sought to both compare and describe quantitatively the variance of the three item difficulty measures as the size and type of calibration sample was varied. At the same time, it was possible to investigate issues related to the Rasch techniques of item calibration by linking these questions to the major investigations, so that the maximum amount of information could be gained from the one study, without enlarging its scope of reference beyond manageable limits. In Chapter 4 the results of four investigations are reported in detail. This concluding chapter brings forward the major findings in nine propositions, which, it is hoped, will provide answers to some past doubts by clarifying our knowledge, and will give direction for further investigation where they do not fully complete the picture. No proposition is seen as more important than the others. Finally, these propositions are projected into ideas regarding the implications for theory, practice and further research which stem from this study.

The Propositions

Proposition 1

The Rasch item difficulty is less susceptible to design effects and to variations from sample to sample, particularly at lower sample sizes, than is the traditional difficulty index.

The standardized mean item difficulty variance for the Rasch index was, without exception, lower than that for the traditional measure of difficulty (see Figures 4.4, 4.5 and 4.6). This might be largely attributed to the setting of the mean item difficulty value at the time of calibration. The Rasch item difficulty index is also less inclined to wide variation at low sample sizes than the traditional item difficulty index. The Rasch index has considerably lower design effect values than the traditional index when estimating the item difficulty from non-simple random sample designs. Furthermore, the Rasch index now has the advantage of a body of techniques associated with it which allow, for example, the linking of two tests with different mean item difficulties onto one difficulty scale through common items. These advantages over the traditional index combine with the previously mentioned greater stability to make the use of the Rasch index a more favourable proposition than the use of the traditional item difficulty index.

Proposition 2

The z-item difficulty index is superior to the Rasch item difficulty index with respect to the stability of estimation.

The z-item difficulty index not only has the mean item difficulty set at a fixed value, but, in addition, the spread of item difficulties across the test, as measured by the standard deviation, is also set at a fixed value. Whether this feature is the major reason for the greater stability of the z-item difficulty index is in need of investigation. Nevertheless, the z-item difficulty index has been demonstrated empirically in this study to be very closely related to the Rasch item difficulty index in its behaviour under the varying conditions imposed by these investigations. The standardized mean item difficulty variance (see Figures 4.4, 4.5 and 4.6) for the z-item difficulty index was, without exception, slightly lower than for the Rasch item difficulty index. This indicates that it was slightly more stable under all the conditions examined in this study. Unfortunately not enough is known about the practical applicability of the z-item difficulty index, compared with the now widely used Rasch and traditional item difficulty indices. The z-item difficulty index does not have the various advantages offered by the Rasch index. For this reason the slight advantage it has in terms of stability will not cause it to be used in preference to the Rasch index.

Proposition 3

The major part of the variance of the Rasch and the z-item difficulty indices, and a lesser part of the variance of the traditional item difficulty index are explained quantitatively through (1) the sample size and (2) the finite population correction; and qualitatively by reference to (3) the sample design and (4) the level of item selection.

As shown by the graphs of the parameter given the name 'structure value' (see Figures 4.7, 4.8 and 4.9), almost all of the variations in item difficulty variance can be explained by the dependence on sample size and the finite population correction. Together these two features reduce the standardized mean item difficulty variance to a near constant value for each combination of sample design, test length and difficulty index. Paradoxically, the one index for which we may easily calculate a theoretical variance measure is the traditional one. Whilst the relationships shown for all three indices are those which are theoretically correct for the traditional index, yet the traditional index fits this formulation least well of the three indices in the conditions examined in this study. Nevertheless, it is apparent that for a known variance we could easily extrapolate to a new expected variance for a different sample size quite accurately, by applying this knowledge of the way sample and population size are related to the variance of item difficulties.

That is it appears that we may write:

$$V_i = \frac{F}{n} \cdot \frac{N-n}{n-1} \cdot K_T \cdot deff$$

where V_i is the mean item variance for item difficulty index i,

F is the structure factor in the simple random case,

n is the sample size,

N is the population size,

K_T is a constant associated with the particular test,

and $deff$ is the appropriate design effect value.

Proposition 4

The variance of the Rasch item difficulty index is inversely proportional to the sample size, as are the variances of the z-item difficulty index and, more approximately, the variance of the traditional item difficulty index.

The plots of the structure value (see Figures 4.7, 4.8 and 4.9) show that after adjusting the standardized mean item difficulty variance by the finite population correction, the product of this variance and the sample size produces a straight line plot for the Rasch index, that is, a near constant value. Thus, as was discussed in Investigation 1, it is clear that the variance of the Rasch item difficulty index is inversely proportional to the sample size. The z-item difficulty index also produces a straight line plot thereby indicating that the same relationship also holds. Although the same trend is true for the traditional item difficulty index, the graph demonstrates some additional perturbations not explained by the sample and population size, suggesting that it is also dependent upon other factors which were not identified. One possibility does lie in the fact that no fixing of the mean item difficulty occurs for the traditional index.

Proposition 5

The Rasch standard error is a good estimator of the variability of the Rasch item difficulty index and is inversely proportional to the square root of the sample size.

The relationship between the standard error and the square root of the item variance is self-evident from the relationship described in Proposition 4, given that the standard error does estimate well the square root of the item variance. This is shown to be the case by the properties of the ratio of the standard error to the square root of the empirically determined item variance. This ratio is always close to unity, and deviations away from unity are small. Whilst the calculated standard error best estimates the true error for simple random samples, there is a trend for the standard error to slightly underestimate the true error at large sample sizes for non-simple random sample designs. One possible explanation involves the finite population correction applied to the

empirically determined errors, but this explanation does not hold in the case of simple random samples. Nevertheless, the Rasch standard error is a worthwhile and practical estimator of the error associated with the measurement of Rasch item difficulty values.

Proposition 6

Rasch item fit values increase from zero towards the population value as the sample size increases from zero towards the population size, and the estimation of these fit values is not affected by sample design except to a very small extent at very low sample sizes.

The trends observed in item fit values as the sample size increased were very clear and confirmed the contentions of Forster (1976) regarding the increase of fit values with sample size (see Figure 4.10). The observed effect does, however, have an underlying basis in theory, associated with the reduced ability to detect statistically significant effects for small samples. In fact, it would be better to say that the detection of good or poorly fitting items is less powerful for small samples. The overall effect of increasing sample size is to 'inflate' the individual item fit values. Whatever their value for one sample size, item fit values will move towards zero for smaller samples sizes and away from zero for larger samples. Thus the entire distribution of item fit values for a particular test is expanded or contracted about the zero fit point, reflecting the differing ability of the sample size to detect significant effects at different sample sizes. One interesting effect is that unlike many other estimated item parameters (such as the three difficulty indices and the point-biserial discrimination index which were all well estimated using small samples) the fit value continues to increase and decrease as the sample size increases and decreases. This means that even for the largest sample size used in this study, of 320 candidates, the mean fit value could have been increased had the sample size been increased. The closeness of the fit values for different sample designs at any given sample size (see Figure 4.10) indicates that the sample design has little effect on the estimation of fit, except for very small samples (less than 60). This hardly need cause concern, because fit is so poorly estimated for such sample sizes as to render this small design effect inconsequential in comparison to the inability of the fit statistic to detect mis-fitting items at such low sample sizes. The fit statistic is not an asymptotic estimate of a population parameter as are most of the other item statistics estimated in this study.

Proposition 7

The deletion of items from a test which do not fit the Rasch model well causes an improvement in overall fit and a reduction in mean item variance, however, although the additional deletion of poorly fitting items improves overall fit further, the mean item variance increases again.

It is apparent that the deletion of poorly fitting items should improve the overall fit of the test, and this is borne out by the results of this study (see Figure 4.10). However, the effect of this deletion process on those items which remain is to reduce their fit values (see Figure 4.12). This, too, is to be expected. The internal consistency of a subgroup of items is most apparent when badly fitting items are also present. Removal of poorly fitting items reduces the apparent internal consistency of those items remaining, as measured by the fit statistic. This effect has been discussed by Wright and Stone (1979). The other effect of deleting poor items is to reduce initially the item variance. However, this is not a large effect (see Figures 4.4, 4.5 and 4.6), and further deletion of poor items in this study caused an increase in item variance comparable with the previous decrease. Ultimately, the conclusion to be drawn is that for constructing Rasch calibrated tests the excessive deletion of items is to be avoided, since quite satisfactory results were obtained through only one process of item deletion, with this improvement being small over the initial test..

Proposition 8

Point-biserial discrimination values remain near constant over varying sample size in the case of simple random samples, otherwise, in the case of non-simple random samples they decrease as the sample size decreases below approximately 200 subjects, with sharper decreases occurring for sample designs with large cluster sizes.

Forster (1976) had contended that point-biserial discrimination values remained constant over varying sample sizes. The results of this study indicate this only to be true for simple random samples. Reference to Figure 4.11 shows this effect quite plainly, and also indicates that the sample point-biserial discrimination values obtained estimate the population values well, provided that the sample size is larger than 200.

Proposition 9

The effect of cluster sampling on the estimation of a variety of item parameters is substantial, particularly at the large cluster sizes often used in educational research and surveys.

A substantial effect of using non-simple random samples was evident for nearly every parameter which was estimated in this study. Throughout, one sample design, the cluster sample design with clusters of size 20, has produced the most deviant outcomes compared to the other designs. Clearly, the magnitude of the deviation from simple random sample estimates is consistently related to the size of the clusters used in the sample design. In terms of deviation from the results of the simple random sample design, increasing cluster size causes an increase in the deviation of estimates from those obtained through the use of simple random samples. This effect was readily apparent in all investigations except those regarding Rasch item fit, which would seem

relatively insensitive to sample design. This study has only touched on the magnitude of this problem of design effect. First, the sample design must also affect the estimation of other parameters not examined here, including those related to item and test reliability. Secondly, this study has used a maximum cluster size of 20. This is smaller than the cluster sizes commonly used in educational surveys, such as intact classes or even within school year-level cohorts. Thirdly, this effect would be less of a problem were it not for the fact that large numbers of educational surveys use cluster sample designs. If cluster sampling were a rarity, rather than commonplace, then the design effect problem would appear less often. Although not a major part of this study, the design effect problem has been evident throughout.

Implications for Theory

Clearly the Rasch and the z-item difficulties conform to a structure which is empirically satisfactory to describe the sampling variability of these two indices. This structure is that which is known to be theoretically correct for the traditional difficulty index. It could be useful if it were possible to express mathematically the item difficulty sampling variances for the Rasch and the z-item difficulty indices, at least in the case of simple random samples. The empirical evidence suggests that even if these formulae are complex, then at least they should still approximate quite well the structure examined in this study. An algebraic description of the sampling distributions of these two indices would assist in the understanding of their expected properties under a variety of conditions, provided these conditions may also be expressed mathematically. In particular, a better mathematical understanding of the properties of the z-item difficulty index may indicate why this index has very similar stability to the Rasch index, and whether this may be the case in all circumstances. Given that both indices show distinct advantages over the traditional index, the z-item difficulty index is worthy of further theoretical study.

The Rasch fit statistic appears to be very susceptible to sample size effects. This makes the interpretation of any particular fit value difficult, for the same items on the same test will exhibit different fit values for different sample sizes. What appears to be an item with poor fit for one sample may be considered quite satisfactory for another sample size. This is particularly important given that many test developers use a 'rule of thumb' cutoff level when examining item fit values with a view to deleting poorly fitting items from a test. The type of fit statistic which is needed, and which may be developed as a useful feature of the Rasch item analysis techniques, is one which exhibits two components. Consider a statistical procedure such as the detection of differences between group means. In this procedure numerical parameters must be considered. The first of these is a measure of the size of the difference detected, and its value may

be largely independent of the sample size. The second is clearly dependent on sample size, and is a measure of the statistical significance we may apply to the first measure. This type of technique is commonly used in a variety of well-established statistical procedures, such as chi-squared tests, or analysis of variance where two measures are examined; the first is concerned with the size of the effect, the second with the statistical significance of the first. If such an approach could be applied to item fit, the first measure would involve the *degree* to which an item appeared to belong, or otherwise, among the others on the test, and this measure would hopefully be largely independent of sample size. The second measure would be an indication of the *significance level* which could be attributed to the first measure.

One of the practical advantages of such a procedure would be the comparability of fit statistics obtained from different sample sizes. It would not matter that the significance levels were different, provided that both were acceptable. Thus the earlier problem of the same item on the same test having different fit values for different sample sizes would reduce to a position where the first measure, that concerned with the degree of fit, was largely unchanged in value. However, the second measure, indicating the significance level of this degree of fit, might well be much larger for one sample size than for another.

Implications for Practice

Item Variance and Sample Size

The knowledge of the structure of variability, whether determined theoretically or empirically, allows the prediction of item variance under some circumstances. If item variances are known from an early calibration on a small sample, the expected error variance to be obtained when a larger sample is used may be estimated. Conversely, the necessary sample size to obtain a maximum allowable item variance may also be estimated. In practice this feature of prediction of the effects under changed circumstances allows more systematic planning when attempting to obtain a certain accuracy, or stability, of item calibration when wise use is made of the knowledge of the relationships between sample size and item variance.

Effects of Cluster Samples

The effect of cluster sampling on the item variances of the three indices has been apparent in this study. Research workers and test developers need to be very cautious in the selection of sample designs and in the way in which they interpret or use the variance of any statistic based on a non-simple random sample design. The observation that both the Rasch and the z-item difficulty indices were less susceptible to design

effects suggests that their use may be preferable to the traditional index when non-simple random sample designs are used.

Deletion of Items

The excessive deletion of items may cause the loss of any gain in precision obtained through the initial deletion of poorly fitting items. This indicates that sufficient caution needs to be used in the item selection process lest one becomes over-enthusiastic in the search for a uni-dimensional subset of items, and in the process produce a test where the mean item variance is no better than when more items were included. Considerations such as test reliability and the need to separate candidates along an ability scale suggest that a longer test may be preferable provided that the stability of the item estimates is comparable with a shorter one.

Implications for Future Research

The z-item difficulty index is in need of more study. Whilst theory may help in understanding some of its features, research into its operational properties may help to illustrate whether the slight advantage of stability it has exhibited over the Rasch index is in fact offset by disadvantages or properties as yet unknown.

The process of item selection deserves more attention. This study has used only three different levels of item selection criteria, one of which was simply to leave intact a test which was sound by traditional criteria. Even at this coarse level of item scrutiny, it was clear that there was an initial advantage gained at the first stage of judicious item deletion, but there was also a later loss of this advantage through continued deletion of items. The point at which the maximum advantage is gained needs to be found. That is, the point where the mean item variance falls to a minimum. Then, if possible, the relationships between this optimum level of item deletion and the criteria which are used for selection should be investigated, even if only empirically. If this is done, it may then be possible to describe criteria by which the maximum item stability may be obtained systematically for all tests able to be scaled with the Rasch model.

REFERENCES

Andersen, E.B.
1973(a) Conditional Inference in multiple choice questionnaires. British Journal of Mathematical and Statistical Psychology, 1973, 26, 31-44 (a).

Andersen, E.B.
1973(b) A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 123-140 (b).

Andersen, J., Kearney, G.E. and Everett, A.V.
1968 An evaluation of Rasch's structural model for test items. British Journal of Mathematical and Statistical Psychology, 1968, 21, 231-238.

Archie, G.
1979 Theoretical and Practical Consequences of the use of Standardised Residuals as Rasch Model fit statistics. Paper presented at the Annual Meeting of the American Educational Research Association. (63rd, San Francisco, California, April, 1979) ED 191 915.

Binet, A. and Simon, T.
1905 Methodes nouvelles pour le dianostic du niveau intellectuel des anormaux. L'Annee Psychologique, 1905, 11, 191-244.

Binet, A. and Simon, T.
1908 Le developpement de l'intelligence chez les enfants. L'Annee Psychologique, 1908, 14, 1-94.

Cattell, J.M. and Farrand, L.
1896 Physical and mental measurements of the students of Columbia University. Psychological Review, 1896, 3, 618-648.

Cattell, J.M.
1890 Mental Tests and Measurement, Mind, 1890, 15, 373-381.

Commonwealth Bureau of Census and Statistics (C.B.C.S,)
1970a Australian Capital Territory Statistical Summary 1970. Canberra, The Bureau, 1970.

Commonwealth Bureau of Census and Statistics (C.B.C.S.)
1970b Australian Capital Territory Statistical Summary 1970. Canberra, The Bureau, 1970.

Commonwealth Bureau of Census and Statistics (C.B.C.S.)
1970a Schools 1969. Canberra, The Bureau, 1970.

Cornish, G.B.
1983 In Search of Stability in Educational Measurement. Unpublished Master's Dissertation; University of Melbourne, 1983.

Dinero, T.E. and Haertel, E.
1977 Applicability of the Rasch Model with varying item discrimination. Applied-Psychological Measurement, 1977, 1, 581-592.

Douglass, J.B.
1980 Applying Latent Trait Theory to Classroom Examination System: Model Comparison and Selection. Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, Massachusetts, April, 1980), ED 189 105.

Forster, F.
1976 The Rasch Item Characteristic Curve and Actual Item Performance. Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, 1976), ED 129 840.

Forster, F.
1978 Everything you wanted to know about the Rasch Model (But were afraid to ask). Portland Public Schools Occasional Papers in Measurement, No. 17, 1978, ED 189 099.

Forster, F. and Ingebo, G.
1978 Rasch Model Monograph Series. Portland Public Schools Occasional Papers in Measurement, No. 20, 1978, ED 189 101.

Forster, F. and Karr, C.
1979 Using the Rasch Model to Increase the Power of Item Analysis. Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, 1979) ED 191 915.

Galton, F.
1883 Inquiries into Human Faculty and its Development. London, Macmillan, 1883.

George, A.A.
1979 Theoretical and Practical Consequences of the use of Standardised Residuals as Rasch model fit statistics. Texas University, Austin. Research and Development center for Teacher Education, April 1979. ED 191 915.

Gulliksen, H.
1950 Theory of Mental Tests, New York, John Wiley and Sons, 1950.

Haberman, S.
1975 Maximum likelihood estimators in exponential response models. (Technical Report). University of Chicago, 1975.

Keeves, J.P.
1972 Educational Environment and Student Achievement. Melbourne, Australian Council for Educational Research, 1972.

Hambleton, R.K.
1969 An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Kish, L.
1957 Confidence intervals for clustered samples. American Sociological Review, 22, 1957, 154-165.

Kish, L.
1965 Survey Sampling. New York: John Wiley.

Kuder, G.F. and Richardson, M.W.
1937 The Theory of Estimation of Test Reliability. Psychometrika, 1937, 2, 151-160.

Panchepakesan, N.
1969 The Simple Logistic Model and Mental Measurement. Unpublished Doctoral Dissertation, University of Chicago, 1969.

Rasch, G.
1960 Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedogogiske Institut, 1960.

Ross, K.N.
1976 Searching for Uncertainty. Melbourne, Australian Council for Educational Research, 1976.

Spearman, C.
1910 Coefficient of Correlation Calculated from Faulty Data. British Journal of Psychology, 1910, 3, 271-295.

Tinsley, H.A.E. and Dawis, R.V.
1975 An investigation of the Rasch simple logistic model: sample free item and test calibration. Educational and Psychological Measurement, 1975, 35, 325-339.

Suppes, P. (Ed)
1978 Impact of Research on Education: Some case studies. National Academy of Education, Washington, DC, 1978.

Whiteley, S.E. amd Dawis, R.V.
1973 The Nature of Objectivity with the Rasch Model Center for the Study of Organizational Performance and Human Effectiveness. Minnesota University, Minneapolis, 1973, ED 075 484.

Whiteley, S.E.
1977 Issues in Applying Rasch's Theory. Journal of Educational Measurement, 1977, 14, 227-235.

Whiteley, S.E.
1980 Latent Trait Models in the Study of Intelligence. Intelligence, 1980, 4, 97-132.

Wissler, C.
1901 The correlation of mental and physical traits. Psychological Monographs, 1901, 3, 1-62.

Wright, B.D.
1967 Sample-free Test Calibration and Person Measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. New York. ETS Princeton, New Jersey.

Wright, B.D.
1977 Misunderstanding the Rasch Model. Journal of Educational Measurement, 1977, 14, 219-226.

Wright, B.D. and Mead, R.J.
1977 BICAL: Calibrating Items and Scales with the Rasch Model Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1977.

Wright, B.D. and Stone, M.
1979 Best Test Design. MESA press, Chicago, 1979.

Yen, W.M.
1979 The extent, causes and importance of context effects on item parameters for two latent-trait models. CTB/McGraw-Hill, Monterey, California, 1979, ED 183 567.

APPENDIX A

POPULATION VALUES OF THE SIX ITEM PARAMETERS FOR THE THREE TEST LENGTHS

Table A.1 The Population Values of the Six Item Parameters Estimated on the 55 Item Test

	Traditional Statistics			Rasch Statistics		
Item number	Traditional difficulty	z-item difficulty	Point-biserial	Rasch difficulty	Standard error	Item fit
01	4.06	29.74	0.1667	34.63	0.482	-0.747
02	48.80	51.02	0.3001	50.76	0.209	5.800
03	42.06	47.81	0.2759	49.26	0.214	6.112
04	29.80	41.98	0.4536	46.30	0.228	-4.456
05	22.59	38.55	0.2784	44.28	0.246	2.235
06	33.05	43.53	0.4735	47.12	0.223	-5.156
07	30.23	42.19	0.3625	46.41	0.228	0.416
08	7.09	31.18	0.2387	37.43	0.382	-0.529
09	21.65	38.10	0.3226	44.00	0.246	-0.343
10	25.75	40.05	0.3316	45.20	0.237	0.551
11	31.85	42.96	0.4378	46.82	0.223	-3.246
12	40.82	47.23	0.4713	48.97	0.214	-4.809
13	10.89	32.99	0.3582	39.77	0.319	-2.236
14	84.59	68.05	0.3398	60.31	0.282	-1.988
15	56.58	54.72	0.3702	52.51	0.214	1.594
16	15.88	35.36	0.4205	41.99	0.278	-3.618
17	27.24	40.77	0.4324	45.62	0.232	-3.478
18	36.21	45.02	0.4811	47.89	0.218	-5.522
19	25.15	39.77	0.4986	45.03	0.237	-6.597
20	46.16	59.66	0.4936	50.17	0.209	-6.061
21	39.40	46.56	0.3418	48.64	0.214	2.515
22	53.63	53.32	0.3565	51.84	0.209	2.479
23	52.65	52.85	0.4350	51.62	0.209	-2.197
24	35.48	44.69	0.3987	47.72	0.218	-0.900
25	17.51	36.13	0.2210	42.60	0.269	1.933
26	55.89	54.40	0.3991	52.35	0.209	-0.265
27	53.12	53.08	0.3670	51.73	0.209	1.540
28	53.59	53.30	0.3576	51.83	0.209	2.210
29	47.57	50.44	0.4166	50.49	0.209	-1.157
30	75.28	63.62	0.2308	57.18	0.241	4.054
31	54.83	53.89	0.3798	52.11	0.209	1.161
32	31.38	42.74	0.4607	46.70	0.223	-4.656
33	26.43	40.38	0.4843	45.39	0.232	-6.039
34	7.43	31.34	0.3057	37.68	0.373	-1.826
35	56.66	54.76	0.3101	52.53	0.214	5.078
36	41.50	47.55	0.4858	49.12	0.214	-5.783
37	39.11	46.41	0.3359	48.57	0.214	2.903
38	41.25	47.43	0.4552	49.06	0.214	-3.837
39	67.34	59.81	0.4813	55.05	0.223	-6.037
40	56.49	54.68	0.5122	52.49	0.214	-7.304
41	57.26	55.05	0.4542	52.66	0.214	-3.712
42	61.57	57.10	0.3227	53.66	0.214	3.149
43	70.67	61.42	0.3913	55.91	0.228	-2.094
44	60.29	56.49	0.2123	53.36	0.214	9.477
45	57.52	55.17	0.4377	52.72	0.214	-2.776
46	55.64	54.27	0.5526	52.29	0.209	-10.148
47	63.11	57.83	0.3564	54.02	0.218	1.042
48	64.69	58.58	0.2060	54.40	0.218	8.790
49	68.28	60.29	0.3626	55.29	0.223	-0.286
50	84.97	68.23	0.1889	60.47	0.282	1.423
51	73.48	62.77	0.3901	56.67	0.237	-2.016
52	75.75	63.84	0.2298	57.32	0.241	4.049
53	64.48	58.48	0.4535	54.35	0.218	-3.867
54	81.04	66.36	0.2353	59.00	0.259	-1.651
55	80.15	65.93	0.3351	58.69	0.255	-1.227

Table A.2 The Population Values of the Six Item Parameters Estimated on the 42 Item Test

	Traditional Statistics			Rasch Statistics		
Item number	Traditional difficulty	z-item difficulty	Point-biserial	Rasch difficulty	Standard error	Item fit
01						
02						
03						
04	29.80	41.56	0.4554	46.40	0.232	-3.812
05	22.59	37.57	0.2775	44.35	0.250	2.895
06	33.05	43.36	0.4806	47.24	0.223	-4.563
07	30.23	41.80	0.3613	46.52	0.228	1.454
08						
09	21.65	59.66	0.3262	44.04	0.250	0.249
10	25.75	39.32	0.3314	45.27	0.241	1.537
11	31.85	42.69	0.4362	46.96	0.228	-2.442
12	40.82	47.65	0.4778	49.14	0.218	-3.891
13	10.89	31.10	0.3586	39.72	0.323	-2.088
14						
15	56.58	56.37	0.3691	52.79	0.214	3.229
16	15.88	33.86	0.4221	41.98	0.278	-3.444
17	27.24	40.14	0.4350	45.69	0.237	-2.777
18	36.21	45.10	0.4856	48.04	0.223	-4.761
19	25.15	38.99	0.5004	45.11	0.241	-6.280
20	46.16	50.61	0.5015	50.38	0.214	-5.174
21	39.40	46.87	0.3407	48.81	0.218	3.930
22	53.63	54.74	0.3552	52.12	0.214	4.061
23	52.65	54.20	0.4344	51.88	0.214	-0.512
24	35.48	44.70	0.3916	47.85	0.223	0.733
25	17.51	34.76	0.2258	42.63	0.269	2.324
26	55.89	55.99	0.4037	52.64	0.214	0.967
27	53.12	54.46	0.3620	52.00	0.214	3.112
28	53.59	54.72	0.3509	52.10	0.214	4.112
29	47.57	51.38	0.4186	50.71	0.214	0.320
30						
31	54.83	55.40	0.3880	52.39	0.214	2.281
32	31.38	42.43	0.4666	46.82	0.228	-4.133
33	26.43	39.69	0.4821	45.46	0.237	-5.237
34						
35						
36	41.50	48.03	0.4882	49.30	0.218	-4.558
37	39.11	46.71	0.3328	48.73	0.218	4.491
38	41.25	47.89	0.4562	49.24	0.218	-2.490
39	67.34	62.32	0.4787	55.43	0.228	-5.033
40	56.49	56.32	0.5083	52.79	0.214	-5.722
41	57.26	56.75	0.4537	52.96	0.218	-2.258
42						
43	70.67	64.16	0.3824	56.32	0.232	-0.771
44						
45	57.52	56.89	0.4334	53.02	0.218	-1.097
46	55.64	55.85	0.5447	52.58	0.214	-8.324
47	63.11	59.98	0.3515	54.38	0.223	2.379
48						
49	68.28	62.84	0.3520	55.68	0.228	1.305
50						
51	73.48	65.72	0.3842	57.12	0.241	-1.142
52						
53	64.48	60.74	0.4443	54.70	0.223	-2.097
54	81.04	69.90	0.2251	59.53	0.264	2.864
55	80.15	69.41	0.3226	59.21	0.261	-0.306

Table A.3 The Population Values for the Six Item Parameters Estimated on the 32 Item Test

	Traditional Statistics			Rasch Statistics		
Item number	Traditional difficulty	z-item difficulty	Point-biserial	Rasch difficulty	Standard error	Item fit
01						
02						
03						
04	29.80	42.59	0.4542	46.57	0.232	-3.296
05						
06	33.05	44.36	0.4776	47.43	0.228	-4.079
07	30.23	42.82	0.3597	46.73	0.232	1.704
08						
09	21.65	38.14	0.3249	44.17	0.255	0.639
10	25.75	40.38	0.3286	45.43	0.241	1.965
11	31.85	43.71	0.4339	47.14	0.232	-1.894
12	40.82	48.59	0.4767	49.39	0.218	-3.528
13	10.89	32.28	0.3562	39.78	0.328	2.114
14						
15						
16	15.88	35.00	0.4200	42.05	0.287	-3.244
17	27.24	41.19	0.4338	45.85	0.241	-2.315
18	36.21	46.08	0.4865	48.25	0.223	-4.431
19						
20	46.16	51.50	0.4908	50.65	0.218	-4.180
21						
22						
23	52.65	55.04	0.4277	52.18	0.218	0.292
24	35.48	45.68	0.3751	48.08	0.223	1.879
25	17.51	35.49	0.2282	42.74	0.273	2.644
26	55.89	56.81	0.3926	52.95	0.218	1.983
27						
28						
29	47.57	52.27	0.4223	50.97	0.218	0.573
30						
31	54.83	56.23	0.3853	52.70	0.218	2.772
32	31.38	43.45	0.4614	47.03	0.232	-3.709
33	26.43	40.75	0.4711	45.62	0.241	-4.354
34						
35						
36	41.50	48.97	0.4890	49.55	0.218	-4.351
37						
38	41.25	48.83	0.4508	49.47	0.218	-1.613
39	67.34	63.04	0.4704	55.79	0.232	-4.582
40	56.49	57.13	0.5081	53.09	0.218	-5.435
41	57.26	57.55	0.4450	53.28	0.218	-1.397
42						
43	70.67	64.86	0.3791	56.71	0.237	-0.568
44						
45	57.52	57.69	0.4280	53.33	0.218	-0.524
46						
47	63.11	60.74	0.3450	54.75	0.223	2.858
48						
49	68.28	63.56	0.3383	56.06	0.232	2.139
50						
51	73.48	66.39	0.3771	57.53	0.241	-0.829
52						
53	64.48	61.49	0.4346	55.06	0.228	-1.469
54						
55	80.15	70.03	0.3179	59.66	0.264	-0.348

APPENDIX B

A WARNING REGARDING THE USE OF DEFF VALUES

A Warning Regarding the Use of Deff Values

It should be noted that the Deff values given in Table 4.2 are the ratios of two *empirically* obtained variances. In general, persons who wish to 'correct' for design effects have a value of the item variance which has been *calculated*, for example, the Rasch Standard Error output by the program BICAL. If this is the case, the values in Table 4.2 should NOT be used. Instead, reference should be made to Table 4.2 and to Tables E.1 to E.12 in Appendix E.

The values given in Appendix E are appropriate to 'correct' the Rasch Standard Error for design (and other) effects. Two points should be made. First, the values in Tables E.1 to E.12 and in Table 4.2 are associated with deft values and not with deff values; that is, they should not be applied to variances but to standard errors. Secondly, they 'correct' the Standard Error by being *divided* into it, *not* by being multiplied by it.

APPENDIX C

DISCUSSION OF THE DEFF VALUES LESS THAN UNITY

Discussion of the Deff Values Less Than Unity

There is a number of instances where Deff values less than unity are encountered. Although not always expected, such values are not exceptional, and are a function of the sample design and the sampling frame. In the case of the CLS-5 design, an increase in the sample size quickly increases the number of clusters required, because the cluster size is small. The largest sample size of 320 requires 64 clusters. The sampling frame contains only 67 'pseudoclasses' or possible clusters. Thus, each CLS-5 sample taken has a minimum overlap with each other CLS-5 sample of 61 pseudoclasses out of a total of 67 pseudoclasses. This overlap means that a fairly 'representative' cross-section of the total population is taken for large sample sizes under the CLS-5 design. That this cross-section is more representative than the SRS-1 design, in terms of the stability of the parameters estimated from the sample, is shown in the Deff values less than unity. Such effects are most likely to occur under just the conditions described above, namely, an almost complete coverage of the primary sampling units (classes or clusters), coupled with some degree of conformity within classes compared to the whole population.

APPENDIX D

TABLES OF THE VALUES WHICH ARE PLOTTED IN FIGURES 4.1 to 4.13

Table D.1 (Raw) Mean Item Variance of the Rasch Item Difficulty Index (Plotted as Figure 4.1) (All values rounded)

Test length	Sample design	Sample size								
		40	60	80	100	120	160	200	240	320
55 items	SRS-1	3.45	2.31	1.73	1.38	1.14	0.82	0.63	0.53	0.37
	CLS-5	3.81	2.56	1.83	1.50	1.18	0.83	0.65	0.52	0.36
	CLS-10	4.20	2.91	2.09	1.64	1.34	0.93	0.74	0.58	0.42
	CLS-20	5.46	3.51	2.59	1.99	1.71	1.21	0.93	0.72	0.51
42 items	SRS-1	3.29	2.08	1.55	1.20	0.97	0.72	0.57	0.46	0.34
	CLS-5	3.63	2.36	1.64	1.30	1.03	0.75	0.58	0.46	0.32
	CLS-10	4.03	2.67	1.89	1.48	1.20	0.84	0.66	0.53	0.37
	CLS-20	5.44	3.47	2.44	1.85	1.56	1.08	0.84	0.66	0.48
32 items	SRS-1	3.47	2.14	1.60	1.23	1.01	0.74	0.58	0.47	0.33
	CLS-5	3.85	2.45	1.72	1.35	1.07	0.77	0.59	0.47	0.32
	CLS-10	4.30	2.78	1.97	1.58	1.26	0.85	0.69	0.56	0.39
	CLS-20	5.75	3.63	2.58	1.93	1.67	1.13	0.87	0.69	0.50

Table D.2 (Raw) Mean Item Variance of the Traditional Item Difficulty Index (Plotted as Figure 4.2) (All values rounded)

Test length	Sample design	Sample size								
		40	60	80	100	120	160	200	240	320
55 items	SRS-1	49.66	32.96	25.10	19.71	16.47	11.97	9.24	7.71	5.71
	CLS-5	66.77	47.47	32.27	26.27	19.84	14.42	10.67	8.03	4.97
	CLS-10	100.5	66.86	47.75	35.12	31.15	23.33	17.44	13.16	8.71
	CLS-20	172.5	104.9	80.55	56.70	52.43	36.83	30.24	23.72	18.36
42 items	SRS-1	51.99	34.21	26.37	20.55	17.06	12.55	10.34	7.98	6.02
	CLS-5	71.61	51.72	34.82	28.40	21.20	15.66	11.28	8.55	5.16
	CLS-10	111.4	74.20	52.53	38.72	34.49	26.15	19.25	14.61	9.52
	CLS-20	196.3	118.9	92.15	64.02	59.65	41.81	34.53	26.90	20.97
32 items	SRS-1	51.40	33.27	26.04	20.55	16.91	12.53	9.58	7.98	5.90
	CLS-5	72.20	52.23	35.17	28.98	21.02	15.69	11.35	8.43	4.99
	CLS-10	115.6	76.43	54.19	39.65	35.34	26.77	19.86	15.00	9.71
	CLS-20	203.2	123.1	95.40	65.70	62.13	43.35	35.50	27.77	21.67

Table D.3 (Raw) Mean Item Variance of the z-item Difficulty Index (Plotted as Figure 4.3) (All values rounded)

Test length	Sample design	Sample size 40	60	80	100	120	160	200	240	320
55 items	SRS-1	8.56	5.72	4.31	3.54	2.89	2.17	1.65	1.36	1.01
	CLS-5	9.22	6.43	4.67	3.77	3.01	2.25	1.73	1.41	0.97
	CLS-10	10.29	7.27	5.30	4.22	3.54	2.59	2.06	1.64	1.16
	CLS-20	13.28	8.90	6.62	5.26	4.50	3.25	2.62	2.11	1.54
42 items	SRS-1	11.22	7.54	5.69	4.66	3.83	2.92	2.31	1.87	1.42
	CLS-5	12.02	8.45	6.10	4.96	3.98	3.02	2.31	1.90	1.31
	CLS-10	13.12	9.33	6.89	5.52	4.62	3.40	2.69	2.15	1.55
	CLS-20	16.77	11.58	8.51	6.81	5.81	4.21	3.40	2.72	2.00
32 items	SRS-1	10.42	6.90	5.13	4.27	3.48	2.66	2.04	1.66	1.21
	CLS-5	11.34	7.75	5.66	4.58	3.61	2.77	2.10	1.72	1.17
	CLS-10	12.42	9.72	6.33	5.23	4.34	3.09	2.51	2.01	1.42
	CLS-20	15.92	10.95	8.01	6.45	5.48	3.96	3.14	2.52	1.84

Table D.4 (Raw) Mean Item Difficulty Variance of the Three Difficulty Indices on the 55 Item Test (Plotted as Figure 4.4) (All values rounded)

Sample design	Difficulty index	Sample size 40	60	80	100	120	160	200	240	320
SRS-1	R	.1004	.0672	.0511	.0409	.0340	.0247	.0190	.0160	.0113
	T	.1126	.0743	.0569	.0444	.0371	.0272	.0208	.0175	.0130
	Z	.0938	.0608	.0451	.0367	.0298	.0222	.0168	.0138	.0102
CLS-5	R	.1105	.0744	.0535	.0440	.0352	.0252	.0197	.0158	.0111
	T	.1497	.1067	.0722	.0593	.0447	.0389	.0241	.0182	.0112
	Z	.1017	.0688	.0490	.0320	.0311	.0230	.0176	.0143	.0098
CLS-10	R	.1213	.0842	.0606	.0485	.0398	.0279	.0226	.0177	.0127
	T	.2277	.1511	.1071	.0789	.0702	.0530	.0397	.0297	.0198
	Z	.1149	.0785	.0560	.0441	.0367	.0265	.0210	.0167	.0118
CLS-20	R	.1583	.1014	.0739	.0584	.0503	.0360	.0279	.0220	.0156
	T	.3889	.2361	.1790	.1288	.1188	.0839	.0686	.0537	.0416
	Z	.1534	.0978	.0710	.0556	.0472	.0336	.0269	.0215	.0157

Table D.5 Standardized Mean Item Difficulty Variance of the Three Difficulty Indices on the 42-Item Test (Plotted as Figure 4.5) (All values rounded)

Sample design	Difficulty index	Sample size 40	60	80	100	120	160	200	240	320
SRS-1	R	.1424	.0911	.0688	.0540	.0446	.0332	.0261	.0212	.0157
	T	.1597	.1039	.0804	.0624	.0522	.0386	.0314	.0244	.0184
	Z	.1267	.0816	.0604	.0490	.0399	.0301	.0236	.0191	.0144
CLS-5	R	.1554	.1031	.0727	.0581	.0465	.0345	.0264	.0216	.0148
	T	.2173	.1568	.1050	.0863	.0645	.0484	.0344	.0264	.0158
	Z	.1370	.0925	.0650	.0522	.0415	.0311	.0236	.0194	.0133
CLS-10	R	.1707	.1160	.0841	.0668	.0539	.0382	.0303	.0244	.0173
	T	.3422	.2266	.1596	.1176	.1049	.0802	.0591	.0446	.0292
	Z	.1514	.1031	.0741	.0585	.0485	.0353	.0276	.0220	.0157
CLS-20	R	.2291	.1497	.1052	.0822	.0699	.0493	.0388	.0305	.0219
	T	.5989	.3621	.2756	.1969	.1827	.1291	.1063	.0824	.0642
	Z	.2023	.1312	.0931	.0732	.0617	.0440	.0352	.0279	.0204

Table D.6 Standardized Mean Item Difficulty Variance of the Three Difficulty Indices on the 32-Item Test (Plotted as Figure 4.6) (All values rounded)

Sample design	Difficulty index	Sample size 40	60	80	100	120	160	200	240	320
SRS-1	R	.1361	.0858	.0645	.0506	.0421	.0310	.0242	.0197	.0140
	T	.1504	.0965	.0754	.0594	.0492	.0367	.0278	.0234	.0172
	Z	.1167	.0743	.0542	.0446	.0361	.0273	.0208	.0169	.0123
CLS-5	R	.1496	.0970	.0692	.0552	.0438	.0324	.0247	.0200	.0136
	T	.2084	.1505	.1006	.0842	.0607	.0462	.0329	.0247	.0145
	Z	.1284	.0842	.0600	.0481	.0375	.0285	.0215	.0175	.0119
CLS-10	R	.1649	.1090	.0798	.0645	.0516	.0354	.0287	.0234	.0164
	T	.3382	.2209	.1568	.1141	.1024	.0782	.0581	.0435	.0284
	Z	.1423	.0958	.0677	.0553	.0455	.0320	.0258	.0205	.0144
CLS-20	R	.2216	.1430	.1010	.0779	.0678	.0473	.0365	.0288	.0208
	T	.5887	.3585	.2717	.1920	.1810	.1275	.1036	.0808	.0630
	Z	.1904	.1233	.0873	.0690	.0581	.0413	.0325	.0258	.0188

Table D.7 Structure Values for the Three Difficulty Indices on the 55-Item Test (Plotted as Figure 4.7) (All values rounded)

Sample design	Difficulty index	Sample size 40	60	80	100	120	160	200	240	320
SRS-1	R	4.086	4.137	4.233	4.274	4.302	4.237	4.147	4.288	4.203
	T	4.579	4.574	4.709	4.633	4.688	4.663	4.545	4.677	4.802
	Z	3.814	3.739	3.735	3.836	3.769	3.814	3.670	3.670	3.766
CLS-5	R	4.496	4.582	4.428	4.599	4.448	4.328	4.305	4.230	4.103
	T	6.091	6.568	5.981	6.188	5.653	5.645	5.261	4.870	4.163
	Z	4.139	4.236	4.058	4.094	3.926	3.956	3.844	3.817	3.618
CLS-10	R	4.933	5.183	5.016	5.064	5.035	4.795	4.932	4.730	4.713
	T	9.264	9.299	8.870	8.237	8.871	9.094	8.671	7.948	7.318
	Z	4.675	4.832	4.638	4.605	4.636	4.557	4.591	4.451	4.359
CLS-20	R	6.438	6.238	6.116	6.099	6.366	6.183	6.100	5.879	5.787
	T	15.82	14.53	14.82	13.45	15.02	14.40	15.00	14.36	15.40
	Z	6.240	6.018	5.876	5.802	5.964	5.766	5.886	5.752	5.814

Table D.8 Structure Values for the Three Difficulty Indices on the 42-Item Test (Plotted as Figure 4.8) (All values rounded)

Sample design	Difficulty index	Sample size 40	60	80	100	120	160	200	240	320
SRS-1	R	5.791	5.609	5.696	5.639	5.623	5.703	5.708	5.659	5.835
	T	6.498	6.397	6.658	6.517	6.598	6.619	6.863	6.510	6.826
	Z	5.152	5.025	5.000	5.113	5.041	5.171	5.166	5.107	5.32
CLS-5	R	6.324	6.346	6.015	6.067	5.879	5.917	5.781	5.761	5.495
	T	8.840	9.650	8.692	9.012	8.150	8.307	7.509	7.044	5.857
	Z	5.571	5.691	5.383	5.449	5.250	5.342	5.161	5.187	4.923
CLS-10	R	6.943	7.137	6.964	6.971	6.820	6.552	6.621	6.523	6.391
	T	13.92	13.95	13.22	12.28	13.26	13.77	12.92	11.93	10.83
	Z	6.160	6.344	6.134	6.110	6.135	6.053	6.044	5.888	5.821
CLS-20	R	9.320	9.212	8.714	8.588	8.840	8.458	8.483	8.163	8.130
	T	24.36	22.29	22.82	20.56	23.10	22.17	23.23	22.03	23.80
	Z	8.227	8.079	7.710	7.643	7.804	7.557	7.697	7.470	7.571

Table D.9 Structure Values for the Three Difficulty Indices on the 32-Item Test (Plotted as Figure 4.9) (All values rounded)

Sample design	Difficulty index	Sample size								
		40	60	80	100	120	160	200	240	320
SRS-1	R	5.538	5.279	5.343	5.284	5.317	5.327	5.284	5.261	5.199
	T	6.117	5.941	6.239	6.200	6.223	6.294	6.078	6.247	6.360
	Z	4.747	4.571	4.484	4.661	4.568	4.691	4.552	4.527	4.551
CLS-5	R	6.086	5.970	5.726	5.762	5.543	5.562	5.391	5.335	5.036
	T	8.477	9.266	8.331	8.788	7.668	7.929	7.183	6.612	5.380
	Z	5.222	5.180	4.972	5.022	4.740	4.886	4.692	4.689	4.403
CLS-10	R	6.708	6.712	6.608	6.731	6.528	6.073	6.277	6.248	6.089
	T	13.76	13.60	12.98	11.92	12.95	13.43	12.70	11.63	10.52
	Z	5.790	5.897	5.603	5.770	5.747	5.487	5.633	5.488	5.344
CLS-20	R	9.013	8.802	8.359	8.136	8.575	8.116	7.974	7.694	7.715
	T	23.95	22.07	22.49	20.04	22.89	21.88	22.65	21.59	23.34
	Z	7.743	7.592	7.231	7.207	7.340	7.087	7.095	6.904	6.966

Table D.10 Mean Rasch Item Fit Values for the Three Test Lengths (Plotted as Figure 4.10) (All values rounded)

Test length	Sample design	Sample size								
		40	60	80	100	120	160	200	240	320
55 items	SRS-1	-0 073	-0.104	-0.127	-0.147	-0.161	-0.199	-0.224	-0.246	-0.287
	CLS-5	-0.073	-0.105	-0.121	-0.148	-0.167	-0.194	-0.220	-0.247	-0.286
	CLS-10	-0.067	-0.101	-0.117	-0.146	-0.169	-0.199	-0.222	-0.249	-0.288
	CLS-20	-0.049	-0.099	-0.115	-0.146	-0.163	-0.194	-0.226	-0.247	-0.292
42 items	SRS-1	-0.106	-0.138	-0.165	-0.186	-0.208	-0.245	-0.274	-0.304	-0.351
	CLS-5	-0.103	-0.138	-0.162	-0.187	-0.210	-0.240	-0.269	-0.300	-0.346
	CLS-10	-0.101	-0.137	-0.157	-0.188	-0.211	-0.247	-0.274	-0.303	-0.350
	CLS-20	-0.082	-0.133	-0.157	-0.191	-0.208	-0.244	-0.278	-0.302	-0.354
32	SRS-1	-0.138	-0.177	-0.210	-0.237	-0.260	-0.306	-0.348	-0.380	-0.439
	CLS-5	-0.134	-0.176	-0.209	-0.238	-0.267	-0.302	-0.339	-0.377	-0.436
	CLS-10	-0.134	-0.176	-0.202	-0.238	-0.266	-0.308	-0.346	-0.379	-0.439
	CLS-20	-0.110	-0.172	-0.201	-0.241	-0.265	-0.309	-0.350	-0.381	-0.445

Table D.11 Mean Point-biserial Discrimination Values for the Three Test Lengths (Plotted as Figure 4.11) (All values rounded)

Test length	Sample design	Sample size 40	60	80	100	120	160	200	240	320
55 items	SRS-1	0.3652	0.3635	0.3659	0.3667	0.3650	0.3695	0.3689	0.3696	0.3690
	CLS-5	0.3484	0.3579	0.3590	0.3642	0.3646	0.3677	0.3658	0.3676	0.3665
	CLS-10	0.3372	0.3497	0.3478	0.3554	0.3590	0.3644	0.3652	0.3672	0.3674
	CLS-20	0.2846	0.3235	0.3309	0.3517	0.3496	0.3567	0.3579	0.3600	0.3642
42 items	SRS-1	0.3995	0.3993	0.4002	0.4002	0.4005	0.4042	0.4031	0.4041	0.4037
	CLS-5	0.3838	0.3926	0.3933	0.3993	0.3994	0.4026	0 4005	0.4025	0.4014
	CLS-10	0.3705	0.3830	0.3813	0.3884	0.3934	0.3984	0.3999	0.4015	0.4015
	CLS-20	0.3155	0.3561	0.3632	0.3852	0.3835	0.3909	0.3921	0.3936	0.3986
32 items	SRS-1	0.4079	0.4068	0.4072	0.4072	0.4085	0.4107	0.4117	0.4119	0.4105
	CLS-5	0.3908	0.4002	0.4007	0.4060	0.4065	0.4099	0.4082	0.4106	0.4095
	CLS-10	0.3768	0.3897	0.3889	0.3954	0.4010	0.4055	0.4066	0.4093	0.4085
	CLS-20	0.3182	0.3633	0.3701	0.3917	0.3907	0.3970	0.3991	0.4008	0.4061

Table D.12 Mean Rasch Item Fit Values of the Core of 32 Items for the Three Test Lengths (Plotted as Figure 4.12) (All values rounded)

Test length	Sample design	Sample size 40	60	80	100	120	160	200	240	320
55 items	SRS-1	-0.332	-0.422	-0.481	-0.535	-0.596	-0.693	-0.797	-0.876	-0.998
	CLS-5	-0.325	-0.411	-0.480	-0.546	-0.608	-0.706	-0.787	-0.878	-1.014
	CLS-10	-0.303	-0.395	-0.457	-0.528	-0.602	-0.693	-0.781	-0.869	-0.995
	CLS-20	-0.238	-0.376	-0.440	-0.525	-0.585	-0.678	-0.773	-0.843	-0.999
42 items	SRS-1	-0.203	-0.253	-0.294	-0.333	-0.376	-0.422	-0.491	-0.549	-0.631
	CLS-5	-0.195	-0.254	-0.300	-0.332	-0.374	-0.431	-0.485	-0.545	-0.630
	CLS-10	-0.188	-0.249	-0.288	-0.335	-0.375	-0.433	-0.478	-0.541	-0.616
	CLS-20	-0.145	-0.242	-0.282	-0.333	-0.371	-0.418	-0.484	-0.533	-0.627
32 items	SRS-1	-0.138	-0.177	-0.210	-0.237	-0.260	-0.306	-0.348	-0.380	-0.439
	CLS-5	-0.134	-0.176	-0.209	-0.238	-0.267	-0.302	-0.339	-0.377	-0.436
	CLS-10	-0.134	-0.176	-0.202	-0.238	-0.266	-0.308	-0.346	-0.379	-0.439
	CLS-20	-0.110	-0.172	-0.201	-0.241	-0.265	-0.309	-0.350	-0.381	-0.445

Table D.13 Mean Rasch Item Difficulty Variance of the Core of 32 Items for the Three Test Lengths (Plotted as Figure 4.13) (All values rounded)

Test length	Sample design	Sample size								
		40	60	80	100	120	160	200	240	320
55 items	SRS-1	3.19	1.95	1.46	1.14	0.94	0.68	0.53	0.43	0.31
	CLS-5	3.56	2.27	1.59	1.25	0.97	0.71	0.55	0.44	0.30
	CLS-10	4.06	2.64	1.86	1.48	1.16	0.80	0.64	0.52	0.36
	CLS-20	5.68	3.49	2.49	1.83	1.57	1.07	0.82	0.65	0.47
42 items	SRS-1	3.32	2.05	1.54	1.19	0.97	0.72	0.57	0.47	0.34
	CLS-5	3.72	2.36	1.65	1.30	1.02	0.74	0.57	0.46	0.31
	CLS-10	4.15	2.69	1.91	1.53	1.21	0.83	0.67	0.54	0.37
	CLS-20	5.65	3.54	2.53	1.89	1.61	1.11	0.85	0.67	0.48
32 items	SRS-1	3.47	2.14	1.60	1.23	1.01	0.74	0.58	0.47	0.33
	CLS-5	3.85	2.45	1.72	1.35	1.07	0.77	0.59	0.47	0.32
	CLS-10	4.30	2.78	1.97	1.58	1.26	0.85	0.69	0.56	0.39
	CLS-20	5.75	3.63	2.58	1.93	1.67	1.13	0.87	0.69	0.50

APPENDIX E

COMPLETE TABLES OF THE RATIO OF CALCULATED RASCH STANDARD ERROR TO EMPIRICALLY DETERMINED SAMPLING STANDARD DEVIATION OF THE RASCH ITEM DIFFICULTY
(Summarized as Table 4.2)

Table E.1 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 55 Item Test and Sample Design SRS-1 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	1.017	0.098	0.862	1.456
60	1.025	0.074	0.861	1.245
80	1.031	0.066	0.891	1.175
100	1.025	0.073	0.851	1.175
120	1.036	0.072	0.873	1.269
160	1.046	0.067	0.899	1.222
200	1.066	0.068	0.929	1.206
240	1.077	0.069	0.923	1.243
320	1.094	0.069	0.931	1.205

Table E.2 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 55 Item Test and Sample Design CLS-5 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	0.977	0.097	0.690	1.370
60	0.973	0.077	0.747	1.236
80	0.997	0.082	0.746	1.275
100	0.998	0.076	0.725	1.155
120	1.017	0.072	0.842	1.179
160	1.037	0.066	0.844	1.191
200	1.049	0.070	0.884	1.272
240	1.067	0.071	0.904	1.217
320	1.114	0.084	0.939	1.314

Table E.3 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 55 Item Test and Sample Design CLS-10 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	0.949	0.128	0.561	1.523
60	0.928	0.093	0.567	1.212
80	0.951	0.099	0.563	1.201
100	0.954	0.093	0.614	1.122
120	0.960	0.089	0.646	1.183
160	0.992	0.092	0.606	1.175
200	0.991	0.099	0.640	1.222
240	1.008	0.086	0.707	1.197
320	1.034	0.084	0.709	1.179

Table E.4 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 55 Item Test and Sample Design CLS-20 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	0.858	0.144	0.440	1.389
60	0.866	0.122	0.403	1.240
80	0.871	0.114	0.432	1.173
100	0.890	0.125	0.421	1.089
120	0.877	0.113	0.402	1.081
160	0.904	0.114	0.431	1.102
200	0.904	0.116	0.481	1.117
240	0.926	0.115	0.464	1.106
320	0.946	0.120	0.467	1.205

Table E.5 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 42 Item Test and Sample Design SRS-1 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	1.010	0.065	0.878	1.133
60	1.029	0.065	0.912	1.144
80	1.036	0.067	0.886	1.158
100	1.038	0.066	0.902	1.171
120	1.052	0.071	0.940	1.233
160	1.050	0.062	0.972	1.222
200	1.060	0.073	0.902	1.203
240	1.073	0.063	0.954	1.283
320	1.080	0.074	0.929	1.288

Table E.6 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 42 Item Test and Sample Design CLS-5 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	0.971	0.073	0.681	1.115
60	0.972	0.070	0.742	1.111
80	1.005	0.074	0.733	1.126
100	1.010	0.080	0.711	1.153
120	1.027	0.067	0.833	1.161
160	1.038	0.066	0.836	1.192
200	1.056	0.071	0.865	1.252
240	1.073	0.072	0.900	1.223
320	1.122	0.083	0.954	1.330

Table E.7 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 42 Item Test and Sample Design CLS-10 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	0.945	0.097	0.563	1.116
60	0.936	0.088	0.560	1.058
80	0.959	0.103	0.561	1.146
100	0.961	0.097	0.607	1.100
120	0.965	0.089	0.641	1.189
160	0.998	0.098	0.603	1.176
200	1.002	0.099	0.637	1.185
240	1.012	0.084	0.702	1.179
320	1.046	0.088	0.699	1.181

Table E.8 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 42 Item Test and Sample Design CLS-20 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	0.845	0.114	0.441	1.070
60	0.854	0.110	0.398	0.999
80	0.871	0.115	0.427	1.104
100	0.893	0.125	0.421	1.098
120	0.882	0.114	0.398	1.087
160	0.913	0.121	0.428	1.094
200	0.911	0.119	0.477	1.123
240	0.933	0.118	0 461	1.126
320	0.953	0.125	0.464	1.202

Table E.9 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 32 Item Test and Sample Design SRS-1 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	1.008	0.065	0.885	1.117
60	1.034	0.060	0.917	1.146
80	1.045	0.072	0.898	1.172
100	1.045	0.064	0.903	1.188
120	1.056	0.059	0.935	1.150
160	1.060	0.066	0.973	1.240
200	1.076	0.065	0.926	1.188
240	1.095	0.068	0.928	1.237
320	1.118	0.058	1.005	1.202

Table E.10 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 32 Item Test and Sample Design CLS-5 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	0.967	0.081	0.681	1.118
60	0.979	0.071	0.750	1.112
80	1.004	0.078	0.731	1.133
100	1.015	0.084	0.708	1.153
120	1.035	0.071	0.829	1.157
160	1.046	0.072	0.834	1.206
200	1.063	0.071	0.864	1.222
240	1.084	0.064	0.935	1.225
320	1.137	0.070	1.006	1.316

Table E.11 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 32 Item Test and Sample Design CLS-10 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	0.943	0.106	0.565	1.107
60	0.939	0.093	0.563	1.067
80	0.964	0.105	0.565	1.149
100	0.955	0.103	0.612	1.099
120	0.966	0.098	0.643	1.181
160	1.015	0.100	0.606	1.164
200	1.006	0.105	0.642	1.137
240	1.010	0.093	0.700	1.179
320	1.051	0.095	0.699	1.187

Table E.12 Ratio of the Calculated Rasch Standard Error to Empirically Determined Sampling Standard Deviation of Rasch Item Difficulty for 32 Item Test and Sample Design CLS-20 (All values rounded)

Sample size	Mean ratio	Standard deviation	Minimum ratio	Maximum ratio
40	0.848	0.126	0.440	1.041
60	0.863	0.122	0.399	1.016
80	0.876	0.130	0.432	1.110
100	0.897	0.133	0.426	1.113
120	0.886	0.126	0.401	1.061
160	0.919	0.134	0.430	1.081
200	0.922	0.129	0.480	1.143
240	0.947	0.133	0.463	1.135
320	0.964	0.139	0.466	1.188

APPENDIX F

RASCH ITEM ANALYSES AND OUTLINE OF THE ITEMS DELETED
(on microfiche)