DOCUMENT RESUME

ED 241 540 TM 830 275

AUTHOR Skaggs, Gary; Lissitz, Robert W.
TITLE Test Equating: Relevant Issues and a Review of Recent Research.
PUB DATE Mar 82
NOTE 62p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Educational Research; *Equated Scores; *Latent Trait Theory; Literature Reviews; Standardized Tests; Tables (Data); Test Construction; *Testing; Testing Problems
IDENTIFIERS One Parameter Model; *Rasch Model; *Three Parameter Model

ABSTRACT

Equating studies using item response theory (IRT) are reviewed. The most well-known papers, as well as a sampling of lesser-known studies, are included. Accompanying tables list the papers and classify them according to the test used, models used, test length and type, sample size and type, method of assessment, equating design, and kinds of comparisons made. A majority of the equating research has focused on the Rasch, or one parameter logistic, model. Initial Studies using the Rasch model investigated the invariance properties of the model: person-free item calibration and item-free person measurement. With tests of similar difficulty and samples of comparable ability, the research suggests that Rasch horizontal equating provides reasonable results. Research on other IRT models has focused on comparing different strategies using the same data set. With regard to vertical equating, most of the research has demonstrated the superiority of the three-parameter model over the Rasch model. Additional equating studies using Monte Carlo methods are reviewed. Finally, four issues relevant to test equating are discussed: assessing the adequacy of equating, sources of equating error, multidimensionality, and out-of-level testing. (PN)

TEST EQUATING: RELEVANT ISSUES AND A REVIEW OF RECENT RESEARCH

Gary Skaggs & Robert W. Lissitz

University of Maryland

Paper presented at the Annual Meeting
of the American Educational Research Association
Los Angeles, 1981

TEST EQUATING: RELEVANT ISSUES AND A REVIEW OF RECENT RESEARCH

One of the most important problems in measurement is to find a precise means of comparing different assessment procedures. Two such procedures, or tests, may be intended to measure the same ability. Both may deemed to be useful in terms of their concurrent or predictive validity. However, an important question is, how can scores on one instrument be compared to scores on another instrument? That is, can the scores on the two instruments be linked in any meaningful way?

This issue manifests itself in many measurement applications. One of these is concerned with the development of an interconnected series of tests. This can be in one of two forms. Tests can be interchangeable, alternate forms designed to have identical psychometric properties. Or, tests can indicate varying degrees of intensity of a trait so that the tests can measure a single dimension across a wide range of ability. The first situation reflects horizontal test equating, and the second reflects vertical test equating.

More formal definitions of test equating have been proposed. According to Angoff (1971, p.562), to equate two tests is "to convert the system of units of one form to the system of units of the other -- so that scores derived from form two after conversion will be directly equivalent." Scores will be equivalent, therefore, if they have the same percentile ranks on two tests.

More recently, Lord (1977,1980) has incorporated the notion of equity into a definition of equating tests. Equity is met when it is "a matter of indifference to applicants at every given ability level whether they are to take test x or test y." Several important requirements are implicit in this definition of test equating. First, equating makes sense only if two tests measure the same ability. Secondly, the equating should be the same regardless of which test is equated to the other. Third, the equating should be the same regardless of the population from which it is conducted. Finally, as shown in Theorem 13.3.1 (Lord, 1980, p.198), tests cannot be strictly equated unless the tests are equally reliable or perfectly parallel.

In practice, of course, scores are not perfectly reliable, and the above conditions are rarely met. A less rigorous definition has been used in connection with developing statistically equivalent tests. Under this definition, two tests are equated if examinees of equal ability would be expected to obtain the same score on each test. This has been referred to by Kolen (1981) as the definition of equating for non-parallel tests and by Whitely & Dawis (1974) as an equating of tau-equivalent measures.

Research into methods of equating tests has been an ongoing process for the better part of three decades. In the past decade, however, there has been an upsurge of interest due to the application of item response theory (IRT) methods to test equating. The focus of this inquiry has been on the development of new equating techniques and comparing their effectiveness with

that of traditional approaches. While representation in such journals as Journal of Educational Measurement and Applied Psychological Measurement has increased, even more papers have appeared at the meetings of many professional organizations, such as the American Educational Research Association, National Council on Measurement in Education, and the Minnesota Computerized Adaptive Testing Conferences.

This paper will look at the two major types of equating -- horizontal and vertical -- in terms of the kinds of questions test users are asking. In horizontal equating, the task is to provide test scores which are directly comparable across test forms designed to have similar psychometric properties. The major use for horizontal equating has usually been in developing alternate forms of standardized tests, such as the SAT, GRE, and ITBS. One may also ask about the feasibility of equating scores across different tests, for example, in obtaining CTBS equivalents from the ITBS.

For vertical equating, the problem is much more complex. The basic goal is to develop scores that will link on a single dimension several tests of intentionally different difficulties designed for groups of different abilities. It would be very convenient to have a way of comparing scores for examinees who took tests of unequal difficulty. A major application of this arises in out-of-level testing, where an examinee takes a level of a test that is appropriate to their ability level but which is different from the average for their group. A major focus of this review will be to see how far the present research has

progressed toward providing reasonable conclusions for those wishing to use IRT methods.

METHODS OF EQUATING SCORES

A complete understanding of test equating research requires familiarity with so-called traditional methods of equating and with several aspects of latent trait theory. These aspects include theoretical models, parameter estimation procedures, and equating techniques. It is beyond the scope of the this paper to provide a survey of all these topics. They have been covered in detail in other sources. In this section, some of these references will be provided. The authors will assume that the reader is familiar with at least some of these references.

There are two major types of traditional equating methods -- linear and equipercentile. Other types have been proposed, but the linear and equipercentile approaches have been the most commonly used, and recent research has focused almost exclusively on them. An extended treatment of these methods has been provided by Angoff (1971). That discussion summarizes theoretical distinctions, equating designs, methods for equally and unequally reliable tests, and standard errors.

When using item response theory to equate tests, one must first decide on the latent trait model that best fits the data. The most commonly used models are the Rasch, or one parameter logistic, model (Rasch, 1960) and the two and three parameter

models (Birnbaum, 1968). A summary of latent trait models can be found in Hambleton & Cook (1977) and Hambleton, Swaminathan, Cook, Eignor, & Gifford (1978).

Once the model is specified, item and person parameters need to be estimated. There is a great deal of literature concerning different estimation procedures. For test equating studies, methods based on maximum likelihood have been used almost exclusively. For the Rasch model, unconditional estimation procedures by Wright & Panchapakesan (1969) and the BICAL program of Wright & Mead (1976) have been used frequently. Conditional procedures have been developed (see Gustaffson, 1980), but they have been used only rarely in test equating research. For other latent trait models, the LOGIST program (Wood, Wingersky, & Lord, 1976) has been by far the most frequently utilized program.

When parameter estimates are obtained, test scores can then be placed on the same scale. This scale can be the ability scale, standard score scale, or raw score scale. Procedures for accomplishing this linking have been discussed by Lord (1977,1980), Marco (1977), and Wright (1977). These are based primarily on a linear transformation of the ability scale. Different approaches have been developed recently, and these will be mentioned later. All IRT equating is based on raw scores. Lord (1980) describes two methods for obtaining raw scores that can be used for equating -- estimated true scores and estimated observed scores. Almost all research has utilized the former method.

Because of the volume of recent research, it would be

impossible to review every paper. We have attempted here to include the most well-known papers as well as a fair sampling of lesser known studies. Any omissions are not intended to reflect on the quality of the papers not cited. Surely, by the end of this conference, a number of additional papers will have been presented, and this paper may to some extent become dated. A conclusion that will become apparent in this review is that the field of IRT equating is in midstream and has long way to go toward providing some definitive answers about these methods.

As an aid to summarizing the following studies, tables have been prepared. These list the papers and classify them according to the test used, models used, test length and type, sample size and type, method of assessment, equating design, and kinds of comparisons made. These are intended to aid the reader in referring quickly to some of the relevant dimensions of the studies.

THE RASCH MODEL

Not surprisingly, a majority of the research on test equating has focused on the Rasch or one parameter logistic model. The simplest of the latent trait models, it provides several advantages over other IRT models. Probably the most important of these is that the raw score is a sufficient statistic for estimating ability. Initial studies using the Rasch model investigated the so-called invariance properties of the model in one of two ways. First, two samples are administered the same set of items and the two sets of item

difficulty estimates compared. This is an example of person-free item calibration. Secondly, two sets of items are given to the same sample and the two sets of ability estimates compared. This latter situation is referred to as item-free person measurement and reflects a design that could be applied to test equating. The following studies are summarized in Table 1.

In his initial operationalization of the Rasch model, Wright (1968) illustrated both aspects of invariance with item response data from the Law School Admissions Test (LSAT). For sample-free item calibration, comparing high and low ability samples, Wright found that test calibration curves based on the two sets of ability estimates were very close together. With the same set of data, easy and difficult subtests were formed and ability estimates were obtained for each examinee, thus assessing item-free person measurement. Wright developed a "standardized difference score" for each individual and claimed that if only random error were present such differences would have a mean of zero and a standard deviation of one. In this study, obtained values were very close to zero and one, respctively. In this way, support for both types of invariance was found. Similar evidence was found by Anderson, Kearney, & Everett (1968) who, for two samples of nearly equal ability, obtained a correlation of .96 between the two sets of item difficulty estimates. These early studies therefore provided some evidence supporting the invariance claims of the Rasch model.

In a similar study, Tinsley & Dawis (1975) looked at four types of analogies tests and samples from four very different

populations. Ten comparisons were made between pairs of samples on the same test, and sets of difficulty and ahility estimates were correlated. The results did not follow a clear pattern. Correlations hetween sets of difficulty estimates ranged from -.08 to .98. Generally, lower correlations seemed to occur hetween the most dissimilar samples. Higher correlations tended to occur with larger sample sizes and longer tests. However, some notable exceptions made these generalizations very tentative. In particular, this study dealt with some very small samples and short tests. In this study, ability estimates between identical raw scores were correlated for pairs of samples. In all cases, this correlation was .999, even though item difficulties correlated as low as -.08. As noted by Whitely (1977) and Divgi (1981), test and item calibration are relatively independent concerns.

In hoth the Anderson and Tinsley papers, an attempt was made to assess the effect of removing items that did not fit the Rasch model. In the Anderson study, this resulted in an increase in the correlation between item estimates. In the Tinsley & Dawis study, changes in the correlation coefficients were inconsistent as misfitting items were removed. Some correlations decreased, and some increased and then decreased. In fact, one correlation decreased from .88 to .18 as misfitting items were removed Of course, this shortened the test considerably and probably made calibration less reliable. Clearly, an important variable in the equating outcome is the model upon which the test is constructed. Unfortunately, this topic has been largely neglected in the

equnting literature (see Cook & Eignor, 1981).

One of the first studies to compare different sets of items was conducted hy Whitely & Dawis (1974). These investigators divided n 60-item verbal nnalogies test into two 30-item suhtests in three different ways: odd and even items, easy and hard items, and random suhsets of items. They assessed their results in terms of the two ability estimates for each examinee, one from ench item set. Wright's standardized difference statistic was used to summarize the results. For the odd-even and random sets comparisons, the menns and standard deviations were very close to zero and one, respectively. However, in comparing easy and difficult items, the variance of the standardized difference scores was significantly greater than one. The authors suggested that poor fit to the Rasch model may have influenced this latter result. It turned out thnt 57% of the hard items and 23% of the easy items did not fit the model (40% for the entire test). Reasons for the misfit were not studied. Presumably, more guessing occurred on the harder test. At any rate, the authors concluded that the Rasch model was to some extent invariant for this set of data even though the items were fairly deviant from the Rasch model.

While the above studies investigated situntions relevant to test equating, they were not in fact equating studies since no sets of items were actually linked together. In 1974, the final report of the Anchor Test Study (ATS) was published by Loret, Seder, Bianchini, & Vale. This was a large-scale equating of several forms and levels of seven puhlished reading test

batteries. Linear and equipercentile approaches were used to link raw scores on these tests. While item response theory was not used in this study, it did provide the data used in several later studies.

The first of these involved an application of Rasch procedures to the ATS data by Rentz & Bashaw (1977). Rentz & Bashaw developed the National Reference Scale (NRS) for reading. This scale provided direct raw score comparisons for 28 test/level combinations. This, in effect, treated the 2,644 items on all tests as a calibrated item pool, any subset of which could produce a score on the NRS. Assessment of the adequacy of the equating process was done by looking at the variability of ability estimates across repeated administrations of the same test. This was accomplished because each test was administered to somewhere between 14 and 28 samples. The standard deviation for each raw score ability level was computed and averaged for all raw score groups to provide a single index for each test/level combination. These ranged from .008 to .041 logits. Relative to the ability scale itself and to the standard error of an individual's ability estimate, these values were quite small. Therefore, the authors concluded that there was sufficient invariance to justify the Rasch equating.

Up to this point, the research has shown some support for the validity of the Rasch model. With tests of similar difficulty and samples of comparable ability, Rasch horizontal equating seems to provide reasonable results. To be sure, this provides two major improvements over traditional methods because:

1) item subsets can be tailored to specific groups and 2) statistically equivalent forms can be developed despite unintended differences in difficulty. On the other hand, these studies do raise some questions about the limits of Rasch invariance. The Whitely & Dawis study suggests that ability estimates were not quite as invariant when subtests were deliberately different in difficulty. Tinsley & Dawis point to the potential problem of small samples and samples which are widely different from one another.

Some of the above difficulties have been investigated through vertical equating where tests of unequal difficulty and samples of unequal ability were employed. In an analysis of the Anchor Test Study data, Slinde & Linn (1977) found that vertical equating using the equipercentile approach resulted in large discrepancies in grade-equivalent and scaled scores for the same examinees on different levels of a published test. The authors suggest using latent trait models for vertical equating.

Slinde & Linn (1978) conducted a vertical equating investigation with the Rasch model that was both a replication and an extension on one part of the Whitely & Dawis and Wright (1968) studies. Whitely & Dawis noted that the variance of standardized difference scores on easy and difficult subtests was slightly larger than would be expected with random purely measurement error. Slinde & Linn replicated this situation. In addition, they investigated the stability of equating when item difficulty estimates were obtained from one sample and then applied to a different sample. In practical vertical equating

studies, these two samples may differ widely in ability.

In the Slinde & Linn study, a 36 item achievement test was divided into easy and difficult subtests. A sample of 1307 incoming college freshmen was divided into high, medium, and low ability groups based on their performance on the easy subtest. The high and low groups only were used for item and ability estimates. It turned out that the two subtests yielded similar results for a group (using Wright's standardized difference index) when that same group was used in the test calibration. This corroborated the results of Whitely & Dawis and Wright. However, this did not occur when a group other than the one for whom the results are applied was involved in the parameter estimation process. Substantially different ability estimates resulted from the two subtests, as much as 1.2 log ability units (logits). In other words, an examinee's equated score on different levels of a test varied depending on the ability level of the sample on which the equating was based. This was a clear violation of Rasch model invariance.

To be sure, the comparisons involved in this study were severe. The easy and difficult subtests differed by almost two raw score standard deviations. The high and low ability groups differed by about 1.8 logits. Nevertheless, some limits to Rasch invariance were demonstrated that cast doubts on its use in vertical equating.

Gustafsson (1979) criticized Slinde & Linn for dividing their sample into ability groups based on performance on the easy subtest of the same test used for equating. Gustafsson showed

through a simulation that a spurious lack of model fit could be introduced in such a situation due to a regression artifact.

Slinde & Linn (1979) conducted a reanalysis with a set of data drawn from the Anchor Test Study, using fifth graders. Three ability groups were formed based on data from the California Tests of Basic Skills Reading Comprehension subtest, but analysis was carried out on the SRA Reading Test. The procedures were the same as before except calibration was also done for the middle ability group, providing an additional comparison. The results generally supported the earlier study. Moreover, widely different ability estimates were obtained whenever the low group was used either for calibration or comparison. Results involving only the middle and high groups showed comparable ability estimates from the two subtests. This led the authors to conclude that guessing played a role in the poor results, a conclusion shared by Gustaffson (1979). The authors noted correlations of -.68 and -.38 between item difficulty and discrimination indices for the low and middle groups, respectively. Such a negative correlation is indicative of failure to estimate non-zero lower asymptotes. This would imply that a more fully-parameterized model would have provided better results. It also implies that Rasch vertical equating might give better results in situations where guessing is minimized.

Several researchers have examined the viability of Rasch vertical equating when differences between item sets and ability groups were less extreme. Loyd & Hoover (1980) used three

levels of Iowa Tests of Basic Skills (ITBS) Math Computations Test and three samples of pupils from the sixth through eighth grades. Equatings were conducted across adjacent and non-adjacent levels using the three samples as separate calibration groups Their results supported the Slinde & Linn studies in that the equating between any two levels was influenced by the group upon which the equating was based. There was no definite trend, except that perhaps, as in Slinde & Linn, an examinee would receive a higher ability estimate if he/she took the test at the level of the calibration group.

In looking for causes of the inadequate Rasch equatings, Loyd & Hoover were concerned that curriculum content across grade levels, particularly in mathematics, might not represent a unidimensional scale. To investigate this, they performed a principal axis factor analysis of the total item set. The analysis showed that more than one factor was present in the total item pool. Since unidimensionality is a basic assumption for IRT models, these results raise questions about the use of such models for certain tests. On the other hand, unidimensionality is implicit to test equating in general. For two tests to be equated in a meaningful way, they must measure the same trait. It may be that certain types of tests, such as curriculum-based tests, can not be equated because what they measure changes from level to level.

Finally, Cuskey (1981) recently conducted a Rasch vertical equating with the ITBS Reading Comprehension Test--levels 9-14. These tests and levels were also used by Rentz & Bashaw (1977).

Adjacent levels were equated, using only one calibration group at each level. Thus, issues raised by Slinde & Linn and Loyd & Hoover concerning cross-validation were not addressed here. However, Guskey compared the Rasch ability scale to the publisher's grade-equivalent scale for raw scores on each of the test levels. The two scales, not surprisingly, differed widely at extreme ability levels. Moreover, in one area near the middle of the ability scale, the three lower levels (9-11) of the test unexpectedly produced lower Rasch ability estimates than the three higher levels (12-14). Additional evidence suggested that the Rasch estimates were more indicative of the actual abilities of the examinees than grade-equivalent scores. While these results do not challenge sample invariance, they do suggest a clear improvement over publisher's grade-equivalent norms. On the other hand, Guskey minimized guessing by using only high ability examinees. The data therefore probably fit the Rasch model reasonably well.

THE THREE PARAMETER LOGISTIC AND OTHER MODELS

Following research on the Rasch model, several researchers have quite naturally investigated the application of other latent trait models, most notably the three parameter logistic model, to test equating. This research has focused on comparing different strategies using the same data set, the most frequent comparison being the three parameter versus the one parameter logistic

(Rasch) model. A large portion of this work has been conducted at Educational Testing Service under the direction of Dr. Frederic Lord. Perhaps because these papers tend to be more complex and larger in scope than thoe for the Rasch model alone, they are not as numerous. The studies that are discussed in this section have been summarized in Table 2.

One of the first comparative studies of this sort was done by Marco, Petersen, & Stewart (1979). This was a very large study, so it will discussed in some detail. Forty linear, two equipercentile, and the one and three parameter logistic models were examined under a variety of conditions including random and dissimilar samples, internal and external anchor tests, and different types of criterion scores and summary statistics. For any single comparison, only the best (least error) linear model was presented. Generally, there were two distinct studies in this project -- one, in which a test was equated to itself (horizontal equating) and two, in which tests of unequal difficulty were equated (vertical equating). In all situations, an anchor test design was used where one form of a total test was administered to one group of examinees, a second form given to a second group of examinees, and a common anchor test given to both groups. For evaluating the adequacy of the equating, two statistics were developed. Total error, or mean square error, was defined as a sum of squared differences (weighted) between equivalent scores. Secondly, squared bias was defined as the mean squared difference between equated scores. It is easily demonstrated that squared bias is a part of the total error.

TABLE 1 .

RASCH MODEL STUDIES

PAPER	TESTS	ITEMS (number=k)	SAMPLES (number=n)
Wright (1968)	LSAT	k=48	law students: n=976
Anderson et. al. (1968)	intelligence screening test	k=45	Australian armed forces: n_1=608 ; n_2=874
Whitely & Dawis (1974)	unpublished verbal analogies test	k=60 ; 30 items per subtest	college and high school students; n=949
Tinsley & Dawis (1975)	unpublished verbal analogies	4 tests: picture, word, symbol, numbers; k=25-60 per test	4 samples: high school & college students, Voc. Rehab. clients, civil service employees
Rentz & Bashaw (1977)	CAT, CTBS, ITBS, MAT, STEP II, SRA, SAT: see Anchor Test Study	2 forms of each test: k=59-121 per test; k=2,644 total	4th thru 6th graders n=1300-2000 per sample 42 total samples
Slinde & Linn (1978)	CEEB: Math Achievement Test Level I	n=36; 18 per subtest	incoming freshmen college n=1,307 total; 3 subsamples
Slinde & Linn (1979)	SRA -- Blue Level from Anchor Test Study	comprehension and vocabulary k=48,12 (respecitively)	5th graders; n=1,638 3 subsamples
Loyd & Hoover (1980)	ITBS: Math Comprehension Levels 12-14	k=45 per level; 30 item overlap between adj. levels	6th thru 8th graders; n=1,956
Guskey (1981)	ITBS; Reading Comprehension; Levels 9-14	k=60-80 perlevel; 38-58 item overlap between adj. levels	6th thru 8th graders of high ability; n=6,000; 1,000 per level

TABLE 1 (contd.)

PAPER	TYPE OF EQUATING	METHOD OF ASSESSMENT	COMPARISONS MADE
Wright (1968)	horizontal & vertical	standardized difference; comparison of test calibration curves	ability estimates from easy & hard subtests; difficulty estimates from "smart" & "dumb" samples
Anderson et.al. (1969)	samples of equal ability	correlations between sets of estimates	difficulty estimates from two samples
Whitely & Dawis (1974)	horizontal & vertical	standard difference scores	ability estimates based on easy vs. hard; odd vs. even random subsests of items
Tinsley & Dawis (1975)	samples of different ability	correlations between sets of ability & difficulty estimates	10 comparisons: pairs of samples responding to the same test
Rentz & Bashaw (1977)	horizontal & vertical	standard deviation of parameter estimates across items & persons	All test forms placed on the Rasch ability scale
Slinde & Linn (1978)	vertical	standardized differences	cross-validation of ability estimates
Slinde & Linn (1979)	vertical	standardized differences	cross-validation of ability estimates
Loyd & Hoover (1980)	vertical	graphic presentation	cross- validation of ability estimates
Guskey (1981)	vertical	correlations and mean differences between Rasch and G-E scores	comparison of Rasch ability scale with publisher's grade-equivalents

TABLE 2

THREE PARAMETER AND OTHER MODELS

PAPER	TESTS	MODELS	ITEMS (number=k)	SAMPLES (number=n)
Marco, Petersen, & Stewart (1979)	SAT -- Verbal	Rasch, three parameter, two equipercentile, forty linear methods	antonym, analogy, sentence completion, reading comprehension k=54,85 (total test) k=20,34-40 (anchor)	high school students (mostly) 2 samples/10subsamples n=1,577 per subsample random and dissmilar subsamples
Kolen (1981)	ITED -- 6th & 7th ed.	Linear, equipercentile, one, two, and three parameter models	vocabulary, quantitative k=40 (vocab.) k=36 (quant.) 7th ed. has 2 levels	9th thru 12th graders n=1,579-1,925 per grade/form combination
Petersen, Cook & Stocking (1981)	SAT -- Verbal & Math	3 linear models, equipercentile, three parameter model: partial pre-calibration & concurrent calibration	Verbal (see above) Math: math, data k=60	high school students (mostly) n=2,670 per sample 5 samples
Conk, Dunbar, & Eignor (1981)	SAT/PSAT-Nat. Merit	2 linear methods, equipercentile, three parameter model	Verbal and Math (see above) PSAT: k=65 (Verbal) k=50 (Math)	mostly high school students n=2,000 per sample 3 samples

TABLE 2 (contd.)

PAPER	TYPE OF EQUATING AND EQUATING DESIGN	METHODS OF ASSESSMENT
Marco, Petersen, & Stewart (1979)	A) Horizontal equating: equating a test to itself through anchor tests: a) internal anchor tests b) external anchor tests c) two internal anchors differing in difficulty from total test B) Vertical equating: total tests of different difficulty: a) through anchor of intermediate difficulty b) two total tests equated directly	A) Criterion score: test score itself B) Criterion scores calculated in two ways: a) IRT equipercentile b) direct equipercentile For both A and B two summary statistics calculated: a) mean square error (total error) b) squared bias (mean difference squared)
Kolen (1981)	A) Horizontal equating: 6th ed. & 7th ed. (Level II) B) Vertical equating: 6th ed. & 7th ed. (Level I) Test forms randomly assigned to examinees, therefore, randomly equivalent groups taking each test	Cross-validation statistic: total mean square error applied to an independent sample Original total test score used as criterion
Petersen, Cook, & Stocking (1981)	External anchor test -- scale drift: test equated to itself through five intervening forms, each with a separate anchor test	Summary statistics: Mean square error and squared bias (see above)
Cook, Dunbar, & Eignor (1981)	Internal anchor test: old SAT and new PSAT have items in common.	Mean square error and squared bias (see above) three parameter model scores used as criterion

TABLE 2 (contd.)

PAPER	TESTS	MODELS	ITEMS (number=k)	SAMPLES (number=n)
Covell (1981)	TOEFL -- 3 forms SLEP	modified and full three parameter, Rasch, and linear	Written Expression, Reading comprehension, Listening k=40,65,50 (resp.)	large and small samples n=2,069-3,172 (large) n=292-317 (small)
Kolen & Whitney (1981)	General Educational Development (GED) 12 forms	equipercentile, linear Rasch, three parameter	5 subtests: writing, science, social studies reading, math k=80,60,60,40,50(resp.)	adults -- high school equivalency n=205 per form n=20 in cross-validation
Patience (1981)	ITED -- Expression	Rasch, two, and three parameter models, equipercentile	Corrections, Spelling k=49,14 (resp.) k=6 (anchor) total test divided into easy, medium, and hard subtests of 25 items each	9th thru 12th graders n=1,000 per grade

TABLE 2 (contd.)

PAPER	TYPE OF EQUATING AND EQUATING DESIGN	METHODS OF ASSESSMENT
Cowell (1981)	Internal anchor test: equating alternate forms with common items Comparisons made between pairs of models and sample sizes	Summary statistics: mean squared difference, mean absoluted difference, squared bias, maximum absolute difference, variance of differences
Kolen & Whitney (1981)	Horizontal equating: each examinee given one test form and anchor test	squared bias and imprecision indices applied to cross-validation sample ANOVA: forms X methods anchor test score is used as the criterion score.
Patience (1981)	Vertical equating: calibration based on 12th grade sample. Easy, medium, and hard subtests are equated with 9th thru 11th grades, respectively	correlation between derived and obtained scores Score obtained from original data set used as criterion score (originally, all examinees responded to all items)

For the horizontal equating part of the study, Marco et. al. found that, when the anchor test was equal in difficulty to the two total tests, the linear and IRT methods performed well. With an internal anchor, the equipercentile approach also worked well. With an external anchor, the Rasch model did slightly better. With a parallel anchor test, the type of sample mattered very little. When the anchor test was easier or more difficult than the total tests, random samples showed very little error. On the other hand, the IRT models were vastly superior to the traditional methods with dissmilar samples (samples of unequal ahility). Neither IRT model was clearly superior to the other.

When tests of unequal difficulty were equated, the best linear method displayed large total errors, followed by the Rasch model. The three parameter model performed with the least amount of error when IRT-based criterion scores were used in the equating. The equipercentile method was the best method when an equipercentile criterion score was used. This indicates some degree of bias in the criterion (the authors also indicate that the horizontal equating criterion score may have favored the Rasch model). This would present a serious problem in interpreting the results.

The results of this large study clearly indicate the superiority of IRT methods in horizontal equating where samples are not randomly chosen. In practice, that is usually the case. For vertical equating, the Rasch model produced a large total error, a finding which is consistent with the Slinde & Linn (1978,1979) and Loyd & Hoover (1980) studies. Findings for the

equipercentile approaches were puzzling. This approach worked fairly well using equipercentile criterion scores. This conflicts with Slinde & Linn (1977) who showed poor results with that approach. It should be noted, however, that entirely different test batteries were used in the two studies. As will be discussed later, this is potentially a critical factor in comparing equating studies.

The three parameter model was far superior to the one parameter model for vertical equating in terms of total error, no matter which criterion was used. This is not surprising since the SAT Verbal items are known to be fairly difficult. The degree of guessing that could result would seem to suggest that estimating lower asymptotes of item characteristic curves will reduce total error. Interestingly, the one parameter model showed less squared bias than the three parameter model when equating was done between easy and medium difficulty tests and considerably more squared bias when the equating was done between medium and hard tests. Between easy and hard tests, the three parameter model showed less squared bias and total error. Perhaps, with the easier tests, less guessing occurred, and the one parameter model therefore provided closer fit to the data than it would with the more difficult tests. This conclusion would be supported by the Rasch model research reported earlier (Guskey, 1981) where the effect of guessing was minimized.

In another recent study, Kolen (1980) expanded on Marco et. al, by studying a number of IRT models as well as a linear and an equipercentile method. Kolen equated vocabulary and quantitative

thinking items separately in two editions of the Iowa Tests of Educational Development. One of the editions contained two difficulty levels, the more difficult of which was equivalent to the other edition. This provided both a horizontal and vertical equating comparison. With one, two, and three parameter models, two types of IRT equating were studied -- estimated true score equating (Lord, 1980; p. 199) and estimated observed score equating (Lord, 1980; p. 202). In addition, a modified Rasch model was used in which the common discrimination (slope) was allowed to vary between the two tests being equated. Another uniqueness to the Kolen study was the use of a cross-validation sample and statistic. The tests to be equated had no items in common, and each test was administered to an independent, random sample. Therefore, there was no repeated measurement anywhere in the design. But since the samples, including an independent cross- validation group, were randomly assigned to their tests, the expected ability distributions were the same. The statistic used for evaluation was a weighted mean square error of differences in equated scores for the cross-validation samples.

The results obtained with this statistic are somewhat confusing, suggesting a complex interaction between item content, difficulty level, and the model. For vertical equating, the linear and Rasch models performed poorly, supporting previous research. Also, the three parameter model and equipercentile models did very well. Of the two IRT equating methods, the estimated observed score method was slightly better than the estimated true score method. For horizontal equating, the

results were more inconsistent. There was a large difference between vocabulary and quantitative items and between supposedly alternate forms. In general, the estimated true score method for the three parameter model produced the best results, but the linear model did quite well. The estimated true score Rasch model did well for the quantitative items.

This study supports previous research on the inadequacy of using the Rasch model for vertical equating. The Marco and Kolen studies suggest that the three parameter model performs better in a variety of situations. Kolen attributes this to failure to account for guessing. For the three parameter model, he notes that the difficulty in obtaining true score equivalents below chance level may have been the reason for that method to perform relatively better at vertical equating. Also, there may be problems with the LOGIST program in assessing the lower asymptote parameter, which in this study had a differential impact on the two difficulty levels of the tests.

With regard to vertical equating, most of the research that has been discussed has demonstrated the superiority of the three-parameter model over the Rasch model. Failure to account for lower asymptotes has been cited as a primary reason for this. An interesting contradiction to this generalization can be found in a study by Patience (1981) using the Expression Test of the Iowa Tests of Educational Development (ITED) and samples of ninth through twelveth graders.

In Patience's study, scale scores were obtained from all examinees to all 63 items. These were the criterion scores used

for comparison against scores derived through equating. The total test was divided into high, medium, and low difficulty subtests. Item responses of eleventh graders to the hard test, tenth graders to the middle test, and ninth graders to the easy test, and an internal anchor test of six overlapping items between adjacent levels formed the basis for the equating. There were 1,000 examinees at each grade level. Correlations between equated ability estimates were used to compare the equipercentile , one, two, and three parameter models. In terms of these correlations, the three parameter model was outperformed by the other three models. Even with below chance level scores removed, the correlation for the three parameter model was lower than for the other methods. Patience offers several reasons for the results, short test length, small sample sizes, and lack of unidimensionality, the last issue of which would preclude IRT equating. In addition, the correlations produced here were based on ability estimates from the 25 items of moderate difficulty for each sample. These estimates were correlated with ability estimates based on 63 items of which the previous 25 were a subset. For ninth graders, the total test was more difficult than the subtest. For eleventh graders, the total test was easier. A spuriousness to the correlations could have been introduced by the overlap of items on which each ability estimate was based. There also could have been an effect due to grade level. It would be interesting to recalculate the correlations within each grade level separately and with overlapping items from the total test removed.

Several studies have been done that are concerned with comparing equating methods (Cook, Dunbar, & Eignor, 1981; Cowell, 1981; Kolen & Whitney, 1981; Petersen, Cook, & Stocking, 1981). Each has focused on different aspects of the equating problem.

Small sample size was suggested by Patience to be a contributing factor in the poor performance of the three parameter model because the estimation procedure had difficulty converging. Sample size as an independent variable was studied by Cowell (1981) with the Test of English as a Foreign Language (TOEFL). Cowell compared several IRT models with large samples (2,000-3,000) and small samples (about 300). In equating alternate forms of the TOEFL, differences between equating methods and samples sizes were quite small. The tests were probably very similar in difficulty and samples were probably equal in ability. In this study, stable three parameter estimates were produced by the small samples. Scores derived with the linear model were used as the criterion scores. the linear model. Discrepancies resulting from using small as opposed to large samples were less than discrepancies resulting from using the one as opposed to the three parameter model. This finding was clouded somewhat by using the linear model as the criterion, so that discrepancies represented agreement between the models rather than adequacy of the equating.

On the other hand, a study by Kolen & Whitney (1981) using the General Educational Development Tests (GED) found that with small samples (170-198), a number of extreme item parameter estimates were produced with the three parameter model. This

suggests problems with the estimation procedure that contributed significantly to equating error, a finding consistent with Patience's results. This study involved a horizontal equating of alternate GED test forms. It seems likely in these studies that the data fit the three parameter model to varying degrees. Patience,for example, decided initially to eliminate 12 of his original 75 items due to nonconvergence of item parameter estimates.

One source of error in the equating process for IRT methods is the translation of item difficulty (and ability) estimates for different sets of items to the same scale. Theoretically, this can be accomplished through a linear transformation involving the means and standard deviations of item difficulties. Petersen, Cook, & Stocking (1981) compared three translation procedures for the three parameter model with six editions of the SAT Verbal and Mathematics Tests. Three linear and one equipercentile method were also compared. The design of this study was fairly unique. An original SAT was equated through five intervening forms to itself. Each pairwise equating was done through an anchor test design of overlapping items. The initial test thus served as its own criterion. One of the translation procedures, called concurrent calibration, was a simultaneous estimation of all item responses from each pair of tests (including the anchor). For the other two methods, called partial pre-calibration, calibration was made separately for each test/ anchor combination. In a "fixed b's" procedure, item difficulties at one step were calibrated. For overlapping items, item

difficulties were fixed for the next calibration, thus forcing all tests to be placed on the same scale. For an "equated b's" procedure, each test/anchor combination was calibrated separately and then linked through a sequential linear transformation. Along with appropriate graphical presentations, Petersen et. al. used a weighted mean square difference (total error) and mean difference squared (squared bias) as summary indices (the same statistics were used by Marco et. al.,1979).

The results first of all showed that the Verbal and Math tests responded quite differently to the same equating methods. For the Verbal tests, the equated b's method was surprisingly close in predicting initial scale scores. The other procedures all overestimated initial scores, and the linear methods were all fairly close to one another. The equipercentile method had the largest total error. Moreover, the other methods were systematic in their overestimation, while the equipercentile approach showed a very erratic pattern of error. The total error for the traditional methods was at least three times that for any IRT method. These results were generally consistent with those of Marco et. al. (1979) for the SAT Verbal test. However, the models used in each study were not identical.

For the Math test, the equated b's method had a total error that was more than three times that for any of the other methods and more than double the scaled score standard deviation. Two of the linear methods -- Levine's Equally Reliable and Unequally Reliable methods (see Angoff, 1971) -- and the concurrent calibration method performed with the least amount of error. The

equipercentile approach again showed an erratic pattern of errors. All methods except equated b's overestimated scores. All IRT methods underestimated at the lower end of the ability scale and overestimated at the upper end.

Reasons for the inconsistency between Verbal and Math tests are not clear. There could have been a difference in the degree of parallelness between the test forms, or perhaps a difference in the degree of model fit. Item responses for reading comprehension items that are grouped around a passage are probably not locally independent. Another possibility is that the differences could be due to random fluctuations of the tests that happened to be chosen. However, the similarity of the results for the Verbal test to those of Marco et. al. tends to suggest that some sort of more systematic process is underlying test content in these cases.

In another comparison of verbal and quantitative items, Kolen (1981) also found substantial inconsistencies between the two types of tests with regard to the relative adequacy of the different equating models. Kolen & Whitney (1981) found smaller discrepancies between several different types of achievement tests. On the other hand, they noticed some differences in the factor structures of their tests.

The studies discussed so far have used a wide variety of tests. How the content of the test might have affected directly the equating results is not clear. However, underlying differences may be important. Studies that have factor analyzed their tests have shown that more than one substantial factor

typically exists, thus violating unidimensionality. If, for the above or other reasons, the content of the test affects equating outcome, as seems likely, then it becomes very difficult to compare results across studies where different tests are used. The recommendation for a practioner to use the method that gives the best results for a particular equating seems questionable because we do not know as yet how well such results will generalize to new samples.

MONTE CARLO METHODS

In the research discussed above, conclusions are derived from real test data. While the data has the desirable property of being representative of actual test results, it suffers from several major drawbacks. The sample sizes required for a test equating study almost necessitate the use of a data sets from large testing projects. Because of this, independent variables, such as samples sizes, test lengths, item factor structures, etc. can not be actively manipulated by the researcher. In every case cited above, it is difficult to interpret results because the superiority of a particular method could be due to sample size, poor data fit to a model, criterion bias, the content of the test, multidimensionality, and many other factors. In most real data cases, one cannot unconfound the influence of these factors.

Monte Carlo research offers the possibility of being able to manipulate independent variables in an experimental fashion. The

major drawback to such methods is of course simulating data that is realistic. In different terms, simulation can provide answers to many questions of internal validity with regard to equating research. At the same time, its major limitation is providing enough external validity, or generalizeability. Relative to empirical research, Monte Carlo studies have received little attention in the IRT literature. Most of the work has dealt with parameter estimation and robustness of models. To the authors' knowledge, no significant simulation of a test equating has been done. Still, some of the research has implications for equating.

Curry, Bashaw, & Rentz (1978) investigated the robustness of Rasch ability estimates when the equal discrimination condition was violated in a number of ways. They studied differences in the shape of the ability distribution, the difficulty of the test relative to the sample, the percentage of items fitting the Rasch model, and the degree of misfit for items not fitting the model. Misfit was described in terms of item discriminations unequal to the average discrimination. In comparing the estimated ability for each examinee with the true value, the error associated with a data set that perfectly fits the Rasch model was used as a yardstick.

The results suggested that estimated abilities were fairly close to their original value in most situations of misfit, relative to the calculated minimum standard error of measurement for an ability estimate. This would seem to indicate that the Rasch model is robust with regard to unequal discrimination. On the other hand, an ANOVA using absolute differences as a

dependent variable produced unexpected results. For tests of appropriate difficulty for the sample, the mean absolute difference between estimated and true ability increased as the percentage of fitting items increased The authors could offer no explanation of these findings, not did they attempt to assess the significance of this trend.

Several issues must be kept in mind when viewing these results. First, the authors attempted to make their simulation as realistic as possible by basing their choice of levels for each independent variable on that found in real test data (Rentz & Bashaw, 1977). This was commendable. On the other hand, the data was generated from the two parameter logistic model, thereby not including any effect due to guessing. This was further insured by a zero correlation between discrimination and difficulty parameters. In a recent study, Yen (1981) has shown through simulation that the relationship between sets of ability estimates from two latent trait models depends largely on the generating model. This suggests that perhaps different results would have been obtained had data been generated from a three parameter model in which lower asymptote values could be fixed.

In the Curry study, the average discrimination for all tests was held constant. In terms of realistic test equating, subsets of items rarely have equal average item discriminations. Thus, even when difficulty parameters are on the same scale, the ability estimates may not be. For example, if the test is appropriate in difficulty for the sample, ability estimates will increase more rapidly in the middle range for a more highly

discriminating test. Therefore, under such circumstances, unequal discrimination could affect Rasch ability estimates. Curry suggested this problem could be overcome by rescaling ability with the average item discrimination. However, Divgi (1981) provided evidence that a more highly discriminating test will show a higher ability estimate at the upper end of the distribution and a lower estimate at the lower end than will a test with a lower discrimination. Such a systematic error would not be remedied by a proportional constant across the entire scale. In his study, Kolen (1981) investigated a modified Rasch model in which the two tests being equated were allowed to have different average discriminations. However, this procedure did not consistently improve equating results.

Failure to account for guessing has been discussed as a reason for the inadequacy of Rasch vertical equating. Evidence for this is cited in a negative correlation between difficulty and discrimination parameter estimates. (To compute this correlation requires the estimation of discrimination parameters.) One simulation has been attempted by Gustaffson (1980) to assess the impact of difficulty-discrimination correlations on ability estimation. Bias was measured in this study in terms of ability estimates within the same group versus estimates derived from another group and applied to the first group. This represented a replication of the methods used by Slinde & Linn (1978,1979). The results showed that when difficulty and discrimination were uncorrelated, mean ability estimates from high and low ability groups were very similar.

However, for a positive or negative correlation, a considerable bias resulted. For a negative correlation, higher ability estimates resulted from estimation within the group whose ability matched the test's difficulty level. That is, for example, for easy items, a higher ability estimate would be obtained from parameters estimated by the low ability group. Such a bias is in the same direction as that reported by Slinde & Linn. These results suggest that guessing was a major factor in the poor Rasch vertical equating for those studies.

Several methodological improvements could be made for future work. For IRT research, a data matrix is usually generated by comparing probabilities of success for each item (these are based on the IRT model) with a uniformly distributed random number. It seems reasonable to use a set of parameters that reflect perfect fit to some model to obtain an estimate of the amount of random error in the simulation process. However, it might be possible that the expected distribution of errors for such a process could be derived. To date, the authors know of no one who has attempted to do this.

These few simulations have barely scratched the surface of what needs to be known about test equating. Monte Carlo studies are typically costly and difficult to run, and so there has probably been a purely economic reluctance to conduct such research. On the other hand, the active manipulation of independent variables is something that can not be accomplished with empirical studies, except on a post hoc basis. The scant research thus far has suggested that the Rasch model is sometimes

robust to ability differences in parameter estimation and sometimes not.

Some of the issues that could be dealt with through simulations include:

1) How does unequal average discrimination affect equating error?

2) How do various types of content or multidimensional fit affect equating error?

3) Can the effects due to shifts in population distributions of ability be separated from equating error?

4) What is the effect of differential reliability of the two tests on equating results?

One methodological consideration to be made by the researcher is to decide on the model from which the data will be generated. For the Rasch model, the two or three parameter logistic model could be used, but the researcher faces the issue of whether or not these models adequately represent test data in real life. What then becomes the focus of study is the extent of agreement between two models, one of which contains fewer parameters than the other. That can be useful, depending on how realistic the fuller model is. For research on the three parameter model, a major problem arises, namely, what should the generating model be? If a four parameter model is used, what should be the additional parameter? The same holds true for dimensionality studies, although perhaps some bootstrap approaches might be improvised (e.g. using factor scores). These methodological concerns have not yet been addressed.

Despite these problems, Monte Carlo methods hold a great

deal of promise for assessing some of the test equating problems that cannot be addressed through empirical studies. Moreover, the authors believe that until such work is completed, further work with existing data sets will not be very useful.

ISSUES RELEVANT TO TEST EQUATING

In reviewing test equating research, many more questions have been raised than have been answered. The results clearly indicate that no single method is superior to the others in all contexts. Because of this research needs to be broadened to include specific aspects of the equating problem. One thing that has become clear is that research which at the time seemed to support one model or another (in this case usually Rasch) can be challenged in light of what has been done since then.

In this final section, we shall discuss several issues that are pertinent to test equating but which have received little attention thus far. We shall also try to provide directions for future research and summarize the conclusions reached in this paper.

Assessing adequacy of equating

In the studies that have been discussed, a number of methods have been introduced for evaluating how well equating procedures performed. In many cases, the choice of method was guided by

limitations in the design of the study. Still, some comments are in order because the conclusions reached in a study are influenced by the manner in which the results are evaluated.

For example, Wright's standardized difference index has been used frequently. A mean of zero and standard deviation of one is a necessary but not sufficient condition for item-free person measurement. As a summary statistic, systematic differences in ability estimates can average out to a zero absolute mean difference. Divgi (1981) demonstrates how this can occur by using a residual plot across the raw score scale. In his example, using data from the Metropolitan Reading Test, he showed that a difficult subtest yielded higher ability estimates at the extremes and an easy subtest yielded higher ability estimates in the middle of the raw score distribution. Yet, the mean and standard deviation of standardized differences were -.024 and 1.21, respectively. How well this example generalizes remains to be seen. On the other hand, Whitely & Dawis (1974) reported values similar to those of Divgi's example. It would have been helpful in this and other studies to have seen standardized differences plotted as a function of the raw score or ability scale.

Much the same could be said of correlating sets of ability or difficulty estimates. As Tinsley & Dawis (1975) have shown, ability estimates can correlate almost perfectly even though difficulty estimates correlate negatively or not at all. Correlations can also be influenced by extreme cases, and they are fairly immune to differences between variances of two sets of

estimates. Because of the these problems, correlations are at best only a crude estimate of invariance. One does not know how high correlations should be. Most studies have subjectively appraised their values, but it could turn out that .95 is substantially lower than one would predict from a particular model.

Many studies, particularly those comparing several IRT models, have used a mean square error concept as a summary statistic. This is a traditional concept in assessing measurement error. With respect to test equating, comparisons between two sets of ability estimates or estimated true or observed scores imply that one of the sets is the criterion, or true, distribution. The total mean square can be broken down into bias and imprecision indices. The major problem with this approach is that both sets of estimates are based on fallible scores and neither one is truly a criterion measure. The problem of the criterion has never been solved. This is largely a theoretical issue. According to Lord (1980, Theorem 13.3.1), tests cannot be strictly equated unless they are perfectly reliable or strictly parallel. Because this is almost never true, tests can only be equated in a tau-equivalency sense (see Whitely & Dawis, 1974, p. 170), that is, in terms of expected scores on two tests. In the empirical studies reviewed here, error of equating is confounded with person measurement error. To the authors' knowledge, no study has yet been done to try to separate test unreliability from equating error.

A variance approach may also contain biases that can affect

a measure of how well certain models perform. Kolen & Whitney (1981), for example, noted that a relatively large variance of converted equating scores will result in a relatively large value of imprecision and hence of total error. The authors believe that this was a reason for a higher value for the three parameter model than for others. Conversely, the possibility exists for a model to look better than it really is, simply because of a small variance of equated scores.

Another aspect of this problem is bias in the criterion score itself. This score could be an estimated true score, ability score, scaled score, or a raw score. Marco, Petersen, & Stewart (1979) addressed criterion bias in their study. For horizontal equating, a test was equated to itself, with the test score serving as the criterion. The authors felt that this procedure may have favored the Rasch model because more ICC parameters were fixed at constant values than for other IRT models. In the vertical equating portion of their study, the criterion scores had to be calculated. It turned out that the mean square errors for the various models were considerably dependent on the method used to calculate the criterion scores. In other studies (Cowell, 1981; Cook, Dunbar, & Eignor, 1981), equated scores from one of the models was used as the criterion. In these cases, mean square error became a measure of agreement, and so it was impossible to tell if any method worked well at all.

What is to be done in light of such conflicting information? Obviously, some research needs to be done on the matter of

criterion bias. Also, we need to know more about equating from a distribution theory point of view. That is, what is the distribution of error from IRT equating methods? For the present, the authors recommend that conclusions based on a single summary statistic be considered very questionable. Multiple assessment procedures should be utilized. An especially important procedure is to examine differences in ability estimates at points across the entire ability scale so that any systematic errors may be spotted. Studies which have provided such graphical or scatterplot techniques have been inherently more useful.

Sources of equating error

The previous discussion has pointed out the need for more investigation into sources of equating error. The purpose of this is to see what systematic errors may result from the effects of different linking procedures, differential reliability, parameter estimation, and shrinkage.

In the Petersen, Cook, & Stocking (1981) study, vastly different results were obtained from the the three linking procedures even though all were based on the three parameter model. According to Lord (personal communication), perfect correlations between difficulty estimates are not found for two major reasons -- lack of model fit and sampling fluctuations. Lord points out that the latter probably predominates in real data sets. Most studies in the literature used a linea transformation based on the means and standard deviations of item

difficulty estimates. For the three parameter model, such a transformation ignores information from the discrimination and lower asymptote parameters. Recently, methods have been proposed which are based on minimizing the differences between item characteristic curves (Divgi, 1980; Haehara, 1980; Stocking & Lord, 1982). These methods are more complex, but initial results seem encouraging.

Another source of error in equating lies in the estimation of parameters. Difficulties in estimation were mentioned by Kolen (1981) and Patience (1981) as a problem with their results. While a detailed examination of procedures is beyond the scope of this paper, several comments can be made. One of the major criticisms made by Wright (1977) of the three parameter model was that discrimination and lower asymptote parameters could not be estimated without severe restrictions being placed on the estimates. In a simulation study Ree (1979) noted that the quality of parameter estimation with LOGIST and other programs depended partly on the characteristics of the ability distribution of the sample on which estimation was based. Work by Reckase (1979) investigated parameter estimation in conjunction with sample size, length of the anchor test, and the type of linking procedure. His results suggested that for larger sample sizes the length of the anchor test was not critical. Larger samples in this case meant 300 or more for the Rasch model and 1,000 or more for the three parameter model. The issue here is really a question of which is more costly in a practical sense -- poor parameter estimation or failure to include souces of

variation in the model. It's an interesting research topic and one that has rarely been explored directly.

Another source of equating error that has not to the authors' knowledge been investigated is the effect of unequal test reliabilities. As has been discussed previously, unequally reliable tests cannot meet Lord's equity requirement. However, some assessment needs to be made of this problem in order to estimate any systematic effects that may occur in the equating process. This seems particularly important for vertically equated tests, where examinees' scores on in-level and out-of-level tests are compared.

Finally, there is some concern over the generalizeability of equating results. In many of the early equating studies, parameter estimates used for equating were derived from the same group for whom the equating results were applied. However, in later studies (Kolen, 1981; Loyd & Hoover, 1980; Slinde & Linn, 1978,1979), calibration based on one group was applied to another. For the Rasch model, the results were not quite as encouraging. The situation is similar to that of shrinkage for multiple regression. The best recommendation to account for this, made by Kolen (1981) and Kolen & Whitney (1981), is to use a cross-validation sample whenever possible.

Multidimensionality

Violation of the unidimensionality assumption has been the least studied of several possible deviations from IRT models. And yet, this may ultimately be the most important source of

misfit. As has been previously pointed out, multidimensionality precludes the use of these IRT methods, but it also in a more general sense, precludes equating altogether. Consider,for instance, specifically changes in curriculum content across grade levels. Does it make sense to equate two test levels when their content differs? In some cases, it may not be meaningful at all to equate vertically.

Factor analysis is the most frequently used method of assessing unidimensionality. This is usually obtained with a non-rotated principal factors solution with estimated communalities in the diagonal of the interitem correlation matrix. Regardless of problems with the procedure, interpretation is problematic. Typically, one wishes to account for as much variance as possible with the first factor. Unfortunately, this first factor, frequently labeled a general ability factor, is not necessarily the trait that was supposed to be measured by the test, that is, reading comprehension, vocabulary, etc. If the first factor accounts for say 75% of the variance on a math test, this does not in any way indicate that the test is unidimensional and that dimension is math. The trait called math could only be extracted through rotating the solution, a procedure which tends to spread out variances across factors.

A major reason why multidimensionality has not been investigated is that it is an immensely more difficult issue to study. Multidimensional latent trait models have not yet been consistently theorized nor have parameter estimation procedures

been developed, although examples of two dimensional models can be found in Lumsden (1978) and Goldstein (1980).

An examination of empirical equating studies shows why this issue deserves more attention. Many of the studies (e.g. Kolen, 1981; Petersen, Cook, & Stocking, 1981) found considerable difference in the adequacy of various equating methods for different types of tests. As yet, we do not know what characteristics of these tests cause the differences. It certainly makes comparisons across different studies difficult if not impossible. The best that can be recommended to the practioner is to select the method that works best for their particular test. That is a weak recommendation because one does not know how consistent the results will be on cross-validation and because of the problems of defining the criterion score.

The lack of knowledge about multidimensionality is a major obstacle to interpreting equating results. This is not meant to be a criticism of previous research but an illustration of how new the field of IRT equating is and how far it needs to go before definitive answers can be attained.

Out-of-level Testing

A growing literature that has been relatively independent of the test equating studies is the research on out-of-level testing. Out-of-level testing concerns the testing of examinees at a level of a test battery other than the one that would usually be assigned to them according to grade or age. Several studies in recent years (Ayrer & McNamara, 1973; Long, Schaffran,

& Kellogg, 1977; Ozenne, 1979) have convincingly demonstrated that substantially different grade-equivalent scores result from in-level versus out-of-level testing of examinees. Most of this research has dealt with students functioning at a level below their grade level. Testing such examinees at a level appropriate for their skills allows for a much better assessment from a diagnostic and instructional point of view. However, no one is sure how to place scores on the scale of the in-level test.

Vertical equating offers a possible solution. The work of Slinde & Linn (1977) suggested that IRT methodology might provide a better solution than traditional approaches. However, test equating research has paid little attention so far to the literature on out-of-level testing. There is virtually no cross-referencing between the two literatures. A result of this is that test equating research has not addressed the problem of interpreting out-of-level test scores. Of the studies reviewed here, only one (Guskey, 1981) has compared the latent trait ability scale to the grade-equivalent scale. Another possible solution is out-of-level norms. To the authors' knowledge, no one has compared such norms to the latent ability scale.

School systems are facing today the problem of appropriate test levels. To date, the IRT literature has not focused on the problem directly, and a real needs exists for further study.

Practical implications of results

For test users, the results of IRT equating research thus far must seem quite confusing. The field has simply not

developed to the point that very many conclusions can be reached. In other words, many questions of critical concern have not been fully answered by the research thus far. These include:

1) Whether IRT methods provide better equating results that traditional methods.

2) Whether it is better to develop out-of-level testing scores through vertical equating or separate norms.

3) Whether latent trait ability estimates provide a more valid measurement of ability than grade-equivalent or other standard scores.

In terms of future research, further empirical studies with one test or another are not likely to be useful. Some work with Monte Carlo procedures would be very helpful in terms of examining potential influences on equating results. In addition, there will probably be a trend toward assessing specific sources of equating error such as parameter estimation, linking procedures, and test reliability. Finally, some theoretical work is needed with the distribution theory of equated scores.

Recommendations for practical applications of test equating based on our review of the literature can be summarized as follows:

For horizontal equating,

1) No single method is consistently superior to the others.

2) If the data are reasonably reliable, tests are nearly equal in difficulty, and samples are nearly equal in ability, probably any method will achieve satisfactory results.

3) The Rasch model should not be used where a substantial amount

of guessing has occurred.

4) The three parameter model should not be used with small sample sizes (less than 1,000).

5) Population changes should be investigated if the equatings take place over a long period of time.

For vertical equating,

1) The Rasch model should not be used at all unless test difficulties differences are small and guessing is minimized.

2) The three parameter model has not been proven to be superior to the Rasch model; very little work has been done with it in vertical equating.

3) In terms of content differences, it may not be meaningful at all to equate vertically.

4) When a vertical equating must be done, the safest procedure at this time would probably be to use an equipercentile approach.

In general,

1) Results should be cross-validated whenever possible.

2) Multiple procedures should be used to evaluate equating results. These include both summary statistics and graphical presentations.

REFERENCES

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

Angoff,W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (ed.). Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Ayrer, J.F. & McNamara, T.C. Survey testing on an out-of-level basis. Journal of Educational Measurement, 1973,10,79-84.

Birnbaum,A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Cook,L.L., Dunbar,S.B., & Eignor, D.R. IRT equating: A flexible alternative to conventional methods for solving practical testing problems. Paper presented at annual meeting of American Educational Research Association, Los Angeles, 1981.

Cook,L.L. & Eignor,D.R. Score equating and item response theory: Some practical considerations. Paper presented at annual meeting of National Council on Measurement in Education, Los Angeles, 1981.

Cowell,W. Applicability of a simplified three parameter logistic model for equating tests. Paper presented at annual meeting of American Educational Research Association, Los Angeles, 1981.

Curry,A.R., Bashaw,W.L., & Rentz, R.R. Invariance of Rasch model ability parameter estimates over different collections of items. Paper presented at annual meeting of American Educational Research Association, Toronto, 1978.

Divgi,D.R. Evaluation of scales for multi-level test batteries. Paper presented at annual meeting of American Educational Research Association, Boston, 1980.

Divgi,D.R. Does the Rasch model really work? Not if you look closely. Paper presented at annual meeting of American Educational Research Association, Los Angeles, 1981.

Goldstein,H. Dimensionality, bias, independence, and measurement scale problems in latent trait score models. British Journal of Mathematical and Statistical Psychology, 1980,33,234-246.

Guskey,T.R. Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. Applied Psychological Measurement, 1981,5,187-201.

Gustaffson,J.E. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. Journal of Educational Measurement, 1979,16,153-158.

Gustaffson,J.E. Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 1980,33,205-233.

Haebara,T. Equating logistic ability scales by a weighted least squares. Japanese Psychological Research, 1980,22,144-149.

Hambleton,R.K. & Cook,L.L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977,14,75-96.

Hambleton,R.K., Swaminathan,H., Cook,L.L., Eignor,D.R., & Gifford,J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978,48,467-510.

Kolen,M.J. Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 1981,18,1-11.

Kolen,M.J. & Whitney,D.R. Comparison of four procedures for equating the tests of General Educational Development. Paper presented at annual meeting of American Educational Research Association, Los Angeles, 1981.

Long,J.V., Schaffran,J.A., & Kellogg,T.M. Effects of out-of-level survey testing on reading achievement scores of Title I, ESEA students. Journal of Educational Measurement, 1977,14,203-213.

Lord,F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977,14,117-138.

Lord,F.M. Practical applications of item response theory. Hillsdale, N.J.: Lawrence Erlbaum, 1980.

Loret,P.G., Seder,A., Bianchini,J.C., & Vale,C. Anchor test study final report: Project report and volumes 1 through 30. Berkeley, Calif.: Educational Testing Service, 1974.

Loyd,B.H. & Hoover,H.D. Vertical equating using the Rasch model. Journal of Educational Measurement, 1980,17,179-193.

Lumsden,J. Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology, 1978,31,19-26.

Marco,G.L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977,14,139-160.

Marco,G.L., Petersen,N.S., & Cook,L.L. A test of the adequacy of curvilinear score equating models. Paper presented at the 1979 Computerized adaptive testing conference, Minneapolis, 1979.

Ozenne,D. You may not get what you think you are getting: What test scores mean in out-of-level testing. Paper presented at annual meeting of American Educational Research Association, San Francisco, 1979.

Patience,W. A comparison of latent trait and equipercentile methods of vertically equating tests. Paper presented at annual meeting of National Council on Measurement in Education, Los Angeles, 1981.

Petersen,N.S., Cook,L.L., & Stocking,M.L. IRT versus conventional equating methods: A comparative study of

scale drift. Paper presented at annual meeting of American Educational Research Association, Los Angeles, 1981.

Rasch,G. Probablistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Reckase,M.D. Item pool construction for use with latent trait models. Paper presented at annual meeting of American Educational Research Association, San Francisco, 1979.

Ree,M.T. Estimating item characteristic curves. Applied Psychological Measurement, 1979,3,371-385.

Rentz,R.R. Monitoring the quality of an item pool calibrated by the Rasch model. Paper presented at the annual meeting of National Council on Measurement in Education, Toronto, 1978.

Rentz,R.R. & Bashaw,W.L. The National Reference Scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977,14,161-179.

Slinde,J.A. & Linn,R.L. Verttically equated tests: Fact or phantom? Journal of Educational Measurement, 1977,14,23-32.

Slinde,J.A. & Linn,R.L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978,15,23-35.

Slinde,J.A. & Linn,R.L. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty, Journal of Educational Measurement, 1979,16,159-165.

Stocking,M.L. & Lord,F.M. Developing a common metric in tem response theory. Unpublished manuscript, 1982.

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

Whitely,S.E. Models, meanings, and misunderstandings: Some issues in applying Rasch's theory. Journal of Educational Measurement, 1977,14,227-236.

Whitely,S.E. & Davis,R.V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974,11,163-178.

Wood,R.L., Wingersky,M.S., & Lord,F.M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. RM 76-6. Princeton, N.J.: Educational Testing Service, 1976.

Wright,B.D. Sample-free test calibration and person measurement. In proceedings of the 1967 Computerized adaptive testing conference on testing problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright,B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977,14,97-116.

Wright,B.D. & Mead,R.J. BICAL: Calibrating rating scales with the Rasch model. Research Memorandum No. 23. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright,B.D. & Panchapakesan,N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969,29,23-48.

Yen,W.M. Using simulation results to choose a latent trait model. Applied Psychological Measurement, 1981,5,245-262.