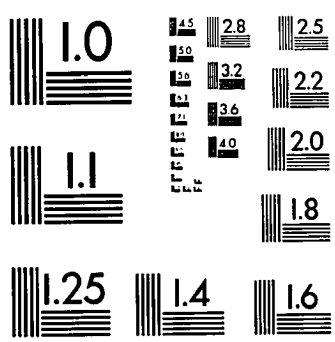

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

INFORMATION TO USERS

This reproduction was made from a copy of a manuscript sent to us for publication and microfilming. While the most advanced technology has been used to photograph and reproduce this manuscript, the quality of the reproduction is heavily dependent upon the quality of the material submitted. Pages in any manuscript may have indistinct print. In all cases the best available copy has been filmed.

The following explanation of techniques is provided to help clarify notations which may appear on this reproduction.

1. Manuscripts may not always be complete. When it is not possible to obtain missing pages, a note appears to indicate this.

2. When copyrighted materials are removed from the manuscript, a note appears to indicate this.

3. Oversize materials (maps, drawings, and charts) are photographed by sectioning the original, beginning at the upper left hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is also filmed as one exposure and is available, for an additional charge, as a standard 35mm slide or in black and white paper format.*

4. Most photographs reproduce acceptably on positive microfilm or microfiche but lack clarity on xerographic copies made from the microfilm. For an additional charge, all photographs are available in black and white standard 35mm slide format.*

***For more information about black and white slides or enlarged paper reproductions, please contact the Dissertations Customer Services Department.**

Doody, Evelyn Norine

EXAMINING THE EFFECTS OF MULTIDIMENSIONALITY ON VERTICAL EQUATING WITH THE THREE-PARAMETER LOGISTIC MODEL

The University of Oklahoma PH.D. 1985

THE UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

EXAMINING THE EFFECTS OF MULTIDIMENSIONALITY ON VERTICAL EQUATING WITH THE THREE-PARAMETER LOGISTIC MODEL

A DISSERTATION
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
degree of
DOCTOR OF PHILOSOPHY

BY
EVELYN NORINE DOODY
Norman, Oklahoma
1985

EXAMINING THE EFFECTS OF MULTIDIMENSIONALITY ON VERTICAL EQUATING WITH THE THREE-PARAMETER LOGISTIC MODEL

APPROVED BY

DISSERTATION COMMITTEE

ACKNOWLEDGEMENTS

I sincerely thank my dissertation committee for their suggestions and criticisms, and for the years of support and faith that I would eventually finish: Dr. W. Alan Nicewander for his guidance and intellectual stimulation, Dr. Larry Toothaker for his letters of encouragement, Dr. Michael Akiyama, Dr. Frank Durso, Dr. Richard Reardon, and Dr. Jim Price for accepting the task of reading my dissertation and for their constructive criticism. To all of you I send my deepest gratitude. May my career reflect the high quality of your teaching.

I am especially indebted to my dissertation committee co-chairmen, Drs. Nicewander and Toothaker, who gave me guidance and support across 3000 miles and many long-distance telephone calls. To Dr. Nicewander I extent special thanks for he is responsible in great part for my being able to continue graduate work at O.U. Thank you, Al, for taking me as your student, sight unseen, in my hour of greatest need. And thank you, Larry, for your special support at that time. I owe so much to both of you. Thank you also, Al, for hours of patient explanations, descriptions, and discussions in person, on the telephone, in letters, and especially for donating your time at the convention.

Thanks to O. U. Psychology Department secretaries, Glenda Smith and Janie Moonie. Thank you, Glenda, for invaluable advice and information, and for being a friend when I needed one most. Thank you, Janie, for your indispensable help, innumerable trips to handle papers, acquiring and typing of forms, reminders to my committee members, and in general, for being my proxy at O.U. while I was away. A special thanks to Janie and her husband, Robert, for being such wonderful hosts.

I thank all my friends and colleagues at CTB/McGraw-Hill for years of encouragement and support: Dr. Wendy Yen for advice, guidance, and patient explanations, for time and effort expended in sharing some of her depth of knowledge in a field about which I knew nothing when I began, without whose expertise I could not have completed the task,

for reading the first through the last drafts of my dissertation, offering invaluable suggestions with each reading. I thank Dr. D. Ross Green, Dr. George Burkett, and CTB/McGraw-Hill for allowing me unlimited access to CTB research facilities and computers; Dr. Beth Langhorst for teaching me how to use the microcomputer, printers, and software on which I typed this dissertation; Lowell T. Cooke for writing many of the computer programs necessary for this research and for helping me write and debug most of the rest; Phyllis O'Donovan for searching the library for papers, journals, and books I needed; and my other friends at CTB for encouragement and support.

I thank Dr. Susan Holmes for her belief in me, for her encouraging words and indispendable advice.

I am deeply grateful to Dr. Barbara A. Warsavage for being my friend, for support and unfailing encouragement, and for continuously telling me "Don't give up!"

I have saved my deepest love and warmest thanks for my two lovely daughters, Erika and Danika Sanders, for years of patience and for suffering in silence the many times I told them "I can't. I have to work on my dissertation." I extend special thanks to Danika for hours of patiently and accurately transcribing numbers into tables.

I give love and thanks to my brother, Bert Doody, and to my sister, Cara Minor, for years of faith and encouragement, and because over the years many of my successes were due to the fact that I could not betray their trust.

To my VERY, VERY special friend and husband, Bob Keeland, I extend my love and thanks for encouraging me through the final hours and the final draft and for helping me see the light at the end of the tunnel when I was too tired to write another word.

Last, but certainly not least, I am most indebted to and most deeply thank my mother, Margaret Catherine Doody, and my father, James Richard Doody. They are the ones who started me on my way, but who didn't get to see me finish. Thank you, Mom and Dad, I love you. Thank you for believing in me. I finally made it, as you always knew I would.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Ten multidimensional data configurations and one unidimensional data configuration for thirty item tests were simulated for 6000 examinees. Each condition consisted of specific item parameter configurations and a specific correlation between traits on two dimensions. Three differences in mean difficulty between the two tests to be equated were simulated.

Equatings were examined in terms of standardized root mean squared differences, standardized differences between means, the ratio of standard deviations, correlations, and tau equivalence. The strength of the correlation between the two true (generating) traits appeared to have little effect on the quality of the equating. In general, the horizontal equatings were better than the vertical equatings.

Results indicated that both the poorest item parameter estimation and the poorest equating occurred for the situation in which one test was unidimensional and one was multidimensional. When both tests were unidimensional or both were multidimensional the equatings were comparable to the unidimensional criterion equatings. There appeared to be a very slight advantage when both tests were unidimensional over equating tests that were both multidimensional when considering local equating bias (local equating consistency). However, the advantage appeared only at the extremes of the ability distribution, particularly at the lower end.

EXAMINING THE EFFECTS OF MULTIDIMENSIONALITY ON VERTICAL EQUATING WITH THE THREE-PARAMETER LOGISTIC MODEL

In many testing situations, it is desirable or even necessary to compare test scores of different examinees or the same examinee across multiple forms of a test or across levels of a test, or to compare two entirely different tests. These different forms, and especially different levels, cannot be expected to be of equal difficulty nor of equal reliability. Therefore, scores on different forms or levels must be transformed in some way so that comparisons between test scores will be meaningful.

Converting the scores of one test or test form to the system of units of the other test or test form so that the resulting scores are directly equivalent is called test equating (Angoff, 1984). Horizontal equating is equating test forms that measure the same attribute at the same difficulty level. Vertical equating equates tests that measure the same attribute at different difficulty levels.

The Need for Vertical Equating

The need for vertical equating arises when scores from tests of unequal difficulty must be compared. Examples of such situations include following an individual's progress through an academic year or across grades, comparing scores of different examinees taking different levels of the same test, or interpreting scores of students who are tested "out of level" because their ability levels are not the same as their grade placement.

Vertical equating is especially applicable to the case of a multilevel battery of tests whose levels are of increasing difficulty. An advantage of such a series of graded tests is that examinees ahead or behind their peers in development can take the most appropriate level, thereby increasing measurement accuracy. These out-of-level scores must then be scaled in some manner (vertically equated) if they are to serve a useful purpose.

The Unidimensionality Assumption

Despite widespread use of vertical equating, results must be examined with caution because many problems remain unsolved. One problem in many practical applications is the effect of violating the unidimensionality assumption. This assumption requires that the tests to be equated onto a common score scale must measure the same underlying (latent) trait or ability. In most equating models used to date unidimensionality is assumed. It is assumed that the test items measure a single latent trait or ability. This assumption frequently does not appear to be met; however, the need for vertical equating does exist. Therefore, the issue becomes that of the robustness of vertical equating to violation of the unidimensionality assumption.

Possible Causes of Violation of the Unidimensionality Assumption

Violation of the unidimensionality assumption could result from various factors. For example, the trait underlying mathematical computation may change from level to level with mathematics performance being differentially dependent upon school curriculum. Diffferent probabilities of success could be due to lack of emphasis on, or lack of introduction to, some aspect of mathematical computation such as fractions or decimals (Loyd & Hoover, 1980). An achievement test in science might require both reading and knowledge of science facts.

Any factor that influences an examinee's score on a test, other than the one latent trait (ability) assumed for the one-dimensional model, will violate the assumption of unidimensionality. The existence of two or more cognitive traits influencing an examinee's response to an item is one such possible factor. Other possible factors are bias or group membership, guessing, speededness, fatigue, cheating, deliberate random answering, or accidently overlooking or skipping an item. Some of these factors can be controlled, reduced, or eliminated, but the fact remains that vertical equating is needed, and the unidimensionality assumption will be violated in many practical situations.

Poor item parameter estimates can compound the problems found in equating. The accuracy of item parameter estimates can be affected by several things, including accuracy of the estimation program (McKinley & Reckase, 1980), and size of the calibration sample (Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978; Reckase, 1977). Violation of the unidimensionality assumption has also been suggested as a problem in estimation of item parameters (Loyd & Hoover, 1980; Cook & Eignor, 1981). Therefore, if multidimensional data have an adverse effect on vertical equating when accurate parameter estimates are used (and this problem is further compounded by the effect of multidimensionality on item parameter estimates), then the effect of multi-dimensionality on vertical equating could be quite extreme.

Equating Models

One distinction between equating models or methods is the dichotomy between the traditional techniques based on classical test theory versus the newer techniques based on item response theory (IRT) or latent trait theory. An extensive coverage of the traditional or classical test theory techniques can be found in Angoff (1984). Item response theory techniques are surveyed by Lord (1975b).

This research is concerned with vertical equating; therefore, IRT techniques were used. Use of traditional techniques is inappropriate and inadequate for vertical equating. Lord (1975b) states that IRT methods are the only existent methods that are appropriate for estimating the equipercentile line of relationship between raw scores when two tests to be equated are not parallel, when they are given to non-equivalent groups, and when every examinee takes an anchor test. Traditional equipercentile equating of two tests to an anchor test is inefficient, biased and inadequate for groups that differ in ability level.

There is ample research supporting the appropriateness and usefulness of the three-parameter model for vertical equating and the inappropriateness of traditional methods (Cook, Dunbar & Eignor, 1981; Cowell, 1981; Gialluca, Crichton, Vale & Ree, 1984; Kolen, 1981; Lord, 1975b, 1977; Marco, Petersen & Stewart, 1983; Petersen, Cook & Stocking, 1983; Slinde & Linn, 1977, 1978, 1979a, 1979b).

Objectives

The objectives of this research were to examine the robustness of vertical equating with the three-parameter logistic model to violation of the unidimensionality assumption and to examine the effects of specific multidimensional data configurations on vertical equating with the three-parameter logistic model.

Educational or Scientific Importance of the Study

Clearly, item pools in most practical situations do not strictly satisfy the assumption of unidimensionality (Lord, 1968). Because IRT models assume unidimensionality of the latent space, their applicability is questionable. Although it is acceptable to assume unidimensionality in the case of some general achievement tests, that assumption is unrealistic for most achievement tests, some criterion-referenced tests, and speeded tests (McKinley & Reckase,

1982; Reckase, 1979, 1981b; Sympson, 1978). Given that the assumption is not valid, it is informative to determine the effects of multidimensionality on vertical equating. The guidelines presented in this paper will be useful to educators and researchers in interpreting the effects of multidimensionality on vertical equating.

METHOD

The computer simulations used in this inquiry began with the generation of two-dimensional data sets from a multidimensional model using prespecified ('true' or generating) parameters. These 'true' parameters were then estimated, followed by an examination of the accuracy of the parameter estimation. Finally, equating was performed, followed by an examination of the accuracy of the equating and an investigation into the effects on vertical equating of various multidimensional conditions.

Test Analysis Model

The statistical model used in this study for analyzing item responses was the three-parameter logistic model of IRT (Birnbaum, 1968). This model assumes that an individual's performance on a test is influenced by only one important unobservable characteristic, θ, which is called a (latent) trait or ability. The three-parameter logistic model assumes that the probability of a correct response to item i by person j with ability level, θ, is:

$$P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + \exp(-1.7a_i(\theta_j - b_i))}, \qquad (1)$$

where a_i, b_i, and c_i are the discriminating power, difficulty and guessing parameters of item i, respectively.

Data Generation

The main question examined in this research was how robust vertical equating based on the three-parameter logistic model is to violation of the unidimensionality assumption underlying the equating. The unidimensionality assumption is violated when scores that are being equated are multidimensional in the sense that an examinee's score on a test is the result of two or more latent traits. The data can be the result of more than one latent trait, and can also vary in the degree of correlation that exists between these traits. Infinitely many multidimensional data sets fulfilling these requirements are possible; therefore, this research was necessarily limited to a few of the possibilities.

Number of latent dimensions. Examinee's scores in most academic testing situations probably result from two or more latent traits. The two-dimensional case was chosen for this research as possibly typical of some published tests and as a starting point for examining the robustness of vertical equating to violation of the unidimensionality assumption. One unidimensional data set was generated (m = 1 in the generating model) to use as a criterion against which the multidimensional data sets could be compared.

Degrees of correlation. The choice of correlations between dimensions was limited to that which seemed possible for a published test. A correlation of zero simulated data sets in which Traits 1 and 2 had no correlation. Correlations of .3 and .4 were chosen to simulate data sets in which the two traits had low correlation; .5, .6, and .75 were used for traits that were more highly correlated; and .9 was used to simulate two traits that were highly correlated. One unidimensional data set (equivalent to a situation in which the correlation between traits is 1.0) was also generated.

Multivariate model. Two-dimensional data sets were generated using a modified version of the multidimensional model described by Hattie (1982). This model, an extension of the three-parameter logistic latent-trait model (Doody-Bogan & Yen, 1983), allows one to vary the levels of difficulty, guessing, discrimination and the number of traits underlying the data. The model used was:

$$P_i(\theta_{jt}) = c_i + \frac{(1 - c_i)}{1 + \exp(-1.7\sum_{t=1}^{m} a_{it}(\theta_{jt} - b_{it}))}, \qquad (2)$$

where $P_i(\theta_{jt}) = P_i(\theta_{j1},\theta_{j2},\ldots,\theta_{jm})$, the probability of a correct response to item i by person j whose location in an m-dimensional latent space is described by abilities θ_{j1}, $\theta_{j2},\ldots,\theta_{jm}$; θ_{jt} is the ability of person j on trait t, a_{it} is the discrimination of item i with respect to trait t, b_{it} is the difficulty of item i with respect to trait t, and c_i is the guessing parameter for item i. Because this research was concerned with two-dimensional data sets, m = 2. Note that for m = 1, the model reduces to the univariate logistic three-parameter model of Birnbaum (1968).

Difficulty values. Base values of b_i were chosen to be reasonable for published tests (Lord,1968; Ross, 1966; Yen, 1982). The b_i's chosen were -2.00, -1.46, -1.25, -1.04, -.84, -.66, -.59, -.52, -.45, -.38, -.31, -.24, -.17, -.10, -.03, .04, .11, .18, .26, .32, .40, .49, .53, .62, .68, .74, .95, 1.02, 1.35 and 1.52 (Mean = -.03; SD = .82).

Three levels of differences in difficulty between the tests to be equated were simulated. For the first set, both tests to be equated had the same mean difficulty values. This simulated a horizontal equating situation. For the vertical equating situations, mean differences in difficulty were chosen to be 1 and 2. These differences in

difficulty represent adjacent levels and once-removed levels, respectively, and were represented as $\bar{b}_D - \bar{b}_E = 1.0$ and $\bar{b}_D - \bar{b}_E = 2.0$, where $\bar{b}_D$ is the mean difficulty of the harder test (Test D), and $\bar{b}_E$ is the mean difficulty of the easier test (Test E). Hereafter, the harder test will be called Test D, and the easier test will be called Test E. For $\bar{b}_D - \bar{b}_E = 1.0$, .5 was added to the base b_i values in order to simulate the b_i for the harder test (Test D), and .5 was subtracted from the base b_i values in order to simulate the b_i for the easier test (Test E). Similarly, 1.0 was added-to and subtracted-from the base b_i values in order to create $\bar{b}_D - \bar{b}_E = 2.0$.

Discrimination values. Discrimination values in IRT simulation studies are typically between 0.5 and 2.0 with a mean around 1.0 (Hattie, 1982; Lord, 1968; Ree, 1979; Ross, 1966; Warm, 1978). For this study, a_i values for the simulated items were chosen to be: .5, .6, .6, .7, .7, .8, .8, .8, .8, .9, .9, .9, .9, 1.0, 1.0, 1.0, 1.0, 1.0, 1.1, 1.1, 1.1, 1.1, 1.2, 1.2, 1.2, 1.3, 1.4, 1.6, 1.8, and 2.0 (Mean = 1.03; SD = .34).

Guessing values. For a multiple choice item with five alternatives, the expected value of c is .2 under the assumption of random guessing; however, values less than .2 are often found in actual test items because some incorrect alternatives are deliberately made more attractive than others. Values of the guessing parameter typically range from 0.0 to 0.4 (Hambleton & Traub, 1971; Hattie, 1982; Kingsbury & Weiss, 1979; Lord, 1968; Warm, 1978). In order to generate data similar to published tests for which the simulation results might be applicable, the following c_i values were chosen: .16 (seven items), .17 (three items), .18 (three items), .19 (three items), .20 (eight items), .21 (three items), and .22 (three items) (Mean = .19; SD = .02).

Item pool. The base item pool consisted of 30 items. The existence of two traits was assumed; therefore, the model required two discrimination parameters, two difficulty parameters and one guessing parameter per item. For each configuration, item parameters a_{i1}, b_{i1}, a_{i2}, b_{i2}, and c_i were randomly assigned to Traits 1 and 2 for Tests E and D. Randomizations were repeated until the desired correlations between latent traits were closely approximated.

Test configurations. Ten two-trait configurations were chosen as typical of possible achievement tests. As a starting point, CTBS/U and CTBS/V (CTB/McGraw-Hill, 1981) were examined from the point of view that the examinees' responses were the result of two traits rather than one. From this examination, specific configurations of correlations between traits and between item parameters were chosen to be simulated. Table 1 shows the desired trait and item parameter correlations for the simulated data sets. Table 2 shows the item parameters used for each data configuration.

For Configuration 1, both traits were assumed to be measured by both tests. Zero correlation between traits was assumed. All correlations for discrimination and difficulty were assumed to be zero. This configuration was chosen to represent a test, such as language mechanics, where Trait 1 is end punctuation (i.e., period, question mark), and Trait 2 is middle punctuation (i.e., comma, colon). Theoretically, a student could understand or master either trait with no knowledge of the other trait.

Configuration 2 was the same as Configuration 1 (both traits measured by both tests, a, b, c randomly assigned) except a moderate correlation (.6) was assumed between traits. This configuration might result from a science test where an examinee's item responses were the result of both reading ability and knowledge of science facts.

Table 1

Correlations for Simulated Data Sets

Configuration	Easy Test		Hard Test	
	Trait 1	Trait 2	Trait 1	Trait 2
1	$r(\theta_1,\theta_2) = 0$		same as	
	$r(a_1,b_1) = 0$	$r(a_2,b_2) = 0$	easy	
	$r(a_1,a_2) = 0$	$r(b_1,b_2) = 0$	easy	
2	$r(\theta_1,\theta_2) = .6$		same as	
	$r(a_1,b_1) = 0$	$r(a_2,b_2) = 0$	easy	
	$r(a_1,a_2) = 0$	$r(b_1,b_2) = 0$	test	
3	$r(\theta_1,\theta_2) = .75$		same as	
	$r(a_1,b_1) = .4$	$r(a_2,b_2) = .0$	easy	
	$r(a_1,a_2) = 0$	$r(b_1,b_2) = 0$	test	
4	$r(\theta_1,\theta_2) = .9$		same as	
	$r(a_1,b_1) = .0$	$r(a_2,b_2) = 0$	easy	
	$r(a_1,a_2) = 0$	$r(b_1,b_2) = 0$	test	
5	$r(\theta_1,\theta_2) = .9$		same as	
	$r(a_1,b_1) = 0$	$r(a_2,b_2) = 0$	easy	
	$r(a_1,a_2) = .8$	$r(b_1,b_2) = .8$	test	
6	$r(\theta_1,\theta_2) = .5$		same as	
	$r(a_1,b_1) = 0$	$r(a_2,b_2) = 0$	easy	
	$r(a_1,a_2) = -.8$	$r(b_1,b_2) = 0$	test	

(table continues)

Table 1 (continued)

Configuration	Easy Test		Hard Test	
	Trait 1	Trait 2	Trait 1	Trait 2
7	$r(\theta_1,\theta_2) = .4$ $r(a_1,b_1) = .2$ $r(a_2,b_2) = .5$ $r(a_1,a_2) = 0$ $r(b_1,b_2) = 0$		same as easy test	
8	$r(\theta_1,\theta_2) = .3$ $r(a_1,b_1) = -.3$ $r(a_2,b_2) = 0$ $r(a_1,a_2) = 0$ $r(b_1,b_2) = 0$		$r(\theta_1,\theta_2) = .3$ $r(a_1,b_1) = -.3$ $r(a_2,b_2) = 0$ $r(a_1,a_2) = .2$ $r(b_1,b_2) = 0$	
9	$r(\theta_1,\theta_2) = .5$ $r(a_1,b_1) = .4$ $r(a_2,b_2) = 0$ $r(a_1,a_2) = 0$ $r(b_1,b_2) = 0$		$r(\theta_1,\theta_2) = .5$ $r(a_1,b_1) = .4$ $r(a_2,b_2) = .7$ $r(a_1,a_2) = 0$ $r(b_1,b_2) = .5$	
10	$r(\theta_1,\theta_2) = .5$ $r(a_1,b_1) = -.3$ $r(a_2,b_2) = 0$ $r(a_1,a_2) = -.5$ $r(b_1,b_2) = 0$		same as easy test	
11	unidimensional		unidimensional	

Note: All item parameters are written without the item subscript, i.
All ability parameters are written without the person subscript, j.

Table 2

Base Item Parameters

	Easy Test					Hard Test				
Configuration	AL1	AL2	BL1	BL2	CL	AH1	AH2	BH1	BH2	CH
1	0.80	0.80	0.68	-1.25	0.18	0.90	1.00	0.62	-2.00	0.16
	2.00	0.80	0.18	1.52	0.20	0.90	1.00	-0.03	0.04	0.18
	1.00	2.00	0.62	-0.38	0.18	2.00	1.40	-0.31	-0.52	0.17
	1.10	1.00	-0.17	-2.00	0.20	1.00	1.10	1.35	0.49	0.20
	0.50	0.90	0.32	-0.03	0.16	0.60	1.80	1.02	0.18	0.19
	1.20	1.80	0.40	0.74	0.20	1.00	2.00	0.95	-1.25	0.20
	1.00	1.00	0.11	0.18	0.16	0.80	1.10	-0.84	0.53	0.16
	1.20	0.50	-0.66	0.62	0.16	1.00	0.50	-2.00	-0.10	0.18
	1.20	1.20	-0.31	0.11	0.18	1.10	0.90	0.26	-0.31	0.20
	0.90	0.80	-0.84	0.32	0.20	0.60	1.10	-0.10	0.32	0.22
	1.10	1.20	-2.00	-0.84	0.19	1.10	0.60	0.32	0.11	0.22
	0.60	0.70	-0.45	-1.46	0.19	1.00	1.00	-0.45	-0.84	0.20
	0.80	0.60	0.49	-0.17	0.16	0.50	1.00	0.49	-0.24	0.21
	1.00	1.10	-0.59	0.68	0.22	0.70	0.70	0.68	0.40	0.16
	0.90	1.10	-0.52	0.40	0.17	1.00	1.00	-1.46	1.02	0.17
	0.80	1.10	-1.25	-0.59	0.20	0.80	1.60	-0.59	-1.04	0.19
	0.90	1.00	-0.24	-1.04	0.20	1.20	0.80	-0.52	-1.46	0.16
	0.70	0.80	1.52	1.35	0.20	1.60	0.90	0.53	-0.17	0.20
	0.70	1.20	-0.10	1.02	0.16	0.90	0.80	-0.17	-0.66	0.19
	1.10	0.60	-0.38	0.26	0.16	1.80	0.90	0.04	-0.03	0.17
	0.60	1.40	-1.04	0.49	0.21	0.70	1.20	-1.04	0.62	0.16
	1.60	0.90	0.95	-0.24	0.21	1.30	0.80	-1.25	-0.45	0.18
	1.80	1.10	-1.46	0.95	0.20	0.80	0.90	-0.24	0.95	0.16
	1.00	1.30	1.02	0.53	0.22	1.20	1.20	0.18	1.35	0.20
	0.90	0.90	0.53	-0.66	0.22	0.90	1.20	0.74	0.26	0.21
	1.30	1.60	1.35	-0.52	0.21	1.10	0.80	-0.66	-0.59	0.16
	1.40	0.90	0.04	-0.10	0.16	0.80	0.60	-0.38	-0.38	0.20
	0.80	1.00	-0.03	0.04	0.17	1.10	0.70	0.11	0.74	0.21
	1.00	0.70	0.26	-0.31	0.17	1.40	1.30	0.40	1.52	0.20
	1.10	1.00	0.74	-0.45	0.19	1.20	1.10	1.52	0.68	0.22
2	same	as	Configuration		1	same	as	Configuration		1
3	0.80	0.80	0.53	0.49	0.20	0.90	0.60	1.52	0.40	0.16
	1.10	0.80	0.68	0.18	0.18	1.60	1.30	0.26	-0.66	0.20
	1.80	0.80	1.02	0.53	0.21	1.20	1.00	0.62	0.74	0.16
	1.10	1.00	-0.45	0.26	0.16	0.80	0.60	-0.03	-1.25	0.17
	1.20	1.10	0.95	-1.25	0.20	1.00	2.00	-0.66	1.35	0.16
	1.00	1.00	0.40	-0.03	0.20	1.40	0.90	1.02	0.95	0.20
	1.10	0.90	-0.10	0.04	0.17	1.10	0.90	-0.17	0.62	0.20
	1.00	2.00	-1.04	-1.46	0.16	1.00	1.10	0.32	-0.10	0.16
	1.00	1.10	0.26	1.52	0.16	0.50	1.20	-1.25	-2.00	0.20
	1.40	1.30	1.35	1.35	0.18	1.20	0.90	0.04	0.32	0.22
	0.90	1.00	-2.00	0.40	0.16	1.80	0.80	0.11	0.49	0.17
	0.80	0.50	0.49	0.11	0.17	1.00	1.10	0.68	0.18	0.21
	0.70	1.80	-1.46	0.32	0.20	1.00	0.70	-0.45	0.68	0.20
	1.10	0.80	-0.59	0.68	0.17	0.90	1.00	-0.31	-0.59	0.18
	0.70	1.20	0.74	1.02	0.21	0.90	1.20	1.35	0.11	0.16
	0.60	0.90	0.62	-0.59	0.20	0.90	0.50	0.74	-0.52	0.18
	1.30	1.00	0.18	0.74	0.20	1.30	1.60	-0.84	-1.04	0.19
	1.20	1.60	0.04	-0.17	0.18	2.00	0.90	0.95	-1.46	0.17
	1.60	1.20	-0.24	-0.45	0.16	1.20	1.10	0.40	0.26	0.20
	2.00	0.90	1.52	-0.31	0.16	1.10	1.80	0.18	-0.31	0.21
	0.80	1.00	-0.52	-0.10	0.20	0.70	0.70	-0.24	-0.45	0.16
	0.90	0.70	-0.66	-1.04	0.19	0.80	0.80	-2.00	0.53	0.16
	1.20	0.70	-0.84	-0.24	0.22	0.80	1.00	-0.38	1.02	0.22
	0.80	1.20	-1.25	0.95	0.19	1.10	1.00	-0.10	-0.24	0.18
	0.60	1.40	-0.17	-0.66	0.22	0.80	1.00	-0.59	-0.38	0.19
	0.50	0.60	0.11	-2.00	0.19	0.60	0.80	0.53	-0.84	0.22
	1.00	0.90	-0.38	-0.52	0.21	0.60	1.40	-1.46	-0.17	0.21
	0.90	1.10	-0.31	-0.38	0.22	1.00	1.20	-0.52	-0.03	0.19
	1.00	0.60	-0.03	0.62	0.16	0.70	0.80	-1.04	1.52	0.20
	0.90	1.10	0.32	-0.84	0.20	1.10	1.10	0.49	0.04	0.20

(table continues)

Table 2 (continued)

Configuration	Easy Test AL1	AL2	BL1	BL2	CL	Hard Test AH1	AH2	BH1	BH2	CH
4	same	as	Configuration		1	same	as	Configuration		1
5	0.50	0.80	1.52	1.35	0.21	0.50	0.80	-1.25	-2.00	0.17
	0.60	0.90	-2.00	-1.46	0.16	0.60	0.90	0.49	0.11	0.16
	0.60	0.70	-1.04	-1.04	0.16	0.60	0.70	0.04	0.49	0.18
	0.70	0.60	0.95	1.02	0.18	0.70	0.60	-0.24	-0.52	0.16
	0.70	0.90	-0.17	-0.31	0.19	0.70	0.80	-0.52	-0.24	0.20
	0.80	0.70	-0.66	-0.38	0.20	0.80	0.70	0.62	0.40	0.18
	0.80	1.00	-0.03	0.26	0.20	0.80	0.50	0.40	0.62	0.20
	0.80	0.50	0.26	-0.03	0.21	0.80	0.60	-1.46	-1.04	0.21
	0.80	0.90	-0.84	-0.59	0.21	0.80	1.00	0.11	0.04	0.17
	0.90	1.10	-0.45	-0.45	0.22	0.90	0.80	0.53	-0.31	0.20
	0.90	1.10	0.32	0.62	0.20	0.90	1.00	-1.04	-0.17	0.16
	0.90	0.60	0.68	0.49	0.20	0.90	1.20	-0.31	-1.25	0.21
	0.90	1.00	0.62	0.32	0.17	0.90	1.10	-0.17	0.74	0.19
	1.00	0.90	0.11	0.04	0.17	1.00	0.80	1.52	1.35	0.21
	1.00	0.80	0.04	0.11	0.20	1.00	1.20	-0.45	-0.66	0.16
	1.00	1.10	1.35	1.52	0.19	1.00	1.30	0.32	0.53	0.16
	1.00	0.80	0.40	-0.10	0.20	1.00	0.90	-0.03	-0.38	0.17
	1.00	1.20	-0.10	0.40	0.22	1.00	1.10	-0.38	-0.03	0.20
	1.10	1.00	-0.59	-0.84	0.18	1.10	0.90	-0.66	-0.45	0.19
	1.10	0.80	0.74	0.74	0.18	1.10	1.40	0.26	-0.59	0.22
	1.10	1.00	-1.46	-2.00	0.16	1.10	1.20	-0.59	0.26	0.22
	1.10	1.20	-0.31	-0.17	0.20	1.10	1.00	0.95	1.02	0.20
	1.20	1.10	-0.52	0.53	0.18	1.20	1.00	-0.84	-0.84	0.18
	1.20	1.00	0.53	-0.52	0.20	1.20	0.90	1.35	1.52	0.19
	1.20	1.30	0.49	0.68	0.22	1.20	1.10	0.18	-0.10	0.16
	1.30	1.60	-0.38	-1.25	0.16	1.30	1.00	1.02	0.95	0.22
	1.40	1.20	-0.24	-0.66	0.16	1.40	1.10	0.74	0.32	0.20
	1.60	1.40	-1.25	-0.24	0.17	1.60	1.60	-0.10	0.18	0.20
	1.80	2.00	1.02	0.95	0.16	1.80	2.00	0.68	0.68	0.16
	2.00	1.80	0.18	0.18	0.19	2.00	1.80	-2.00	-1.46	0.20
6	0.00	1.10	-0.66	0.04	0.17	0.00	1.20	-1.46	-0.66	0.18
	0.00	1.20	0.95	0.68	0.21	0.60	1.30	0.62	0.49	0.20
	0.60	1.80	-0.45	-0.31	0.17	0.60	0.00	0.40	-0.10	0.17
	0.70	0.00	-1.46	0.40	0.18	0.70	0.00	0.95	-0.45	0.20
	0.00	1.30	0.11	0.32	0.20	0.00	1.80	-2.00	1.02	0.16
	0.80	0.00	-0.31	1.02	0.16	0.80	0.00	1.52	1.52	0.20
	0.80	0.00	-0.24	1.52	0.19	0.80	0.00	0.04	-0.59	0.22
	0.80	1.10	0.40	-0.45	0.19	0.80	0.00	0.74	-1.25	0.21
	0.00	1.40	0.74	0.26	0.20	0.80	1.10	0.68	-0.31	0.16
	0.90	0.00	0.18	0.53	0.20	0.00	1.00	-0.24	-0.24	0.20
	0.90	0.90	-0.17	-0.84	0.20	0.90	0.00	0.18	0.68	0.18
	0.90	0.00	-2.00	0.11	0.19	0.00	1.10	1.35	1.35	0.20
	0.00	1.00	1.35	0.95	0.17	0.00	1.20	-1.04	-0.52	0.16
	1.00	0.00	0.49	-1.25	0.22	1.00	0.00	0.49	0.04	0.17
	0.00	1.20	0.32	-0.17	0.21	0.00	1.10	-0.66	0.18	0.16
	1.00	0.00	-0.84	-0.10	0.16	1.00	0.00	0.32	0.62	0.19
	0.00	1.00	-0.52	0.74	0.16	0.00	1.40	-0.84	0.32	0.20
	0.00	1.00	-0.10	-0.59	0.20	1.00	0.00	-0.45	0.11	0.19
	1.10	0.00	-0.59	0.18	0.20	1.10	0.90	0.26	-1.04	0.22
	1.10	0.00	-1.04	-0.66	0.16	1.10	0.00	-0.03	0.95	0.17
	0.00	0.80	0.04	-0.38	0.18	0.00	0.90	0.11	-0.03	0.16
	1.10	0.50	0.62	0.49	0.22	0.00	0.90	0.53	0.53	0.16
	0.00	0.80	1.02	-0.52	0.16	1.20	0.00	-0.59	0.26	0.19
	1.20	0.00	0.26	1.35	0.18	0.00	1.00	-0.31	-1.46	0.21
	1.20	0.00	1.52	-0.03	0.16	0.00	0.70	-0.52	0.40	0.20
	0.00	0.70	-1.25	-0.24	0.16	1.30	0.00	-0.10	0.74	0.18
	0.00	0.90	-0.03	-1.04	0.20	0.00	0.80	-0.17	-0.38	0.22
	1.60	0.00	0.53	-2.00	0.21	1.60	0.50	-0.38	-0.17	0.16
	0.00	0.60	-0.38	-1.46	0.20	0.00	0.60	1.02	-2.00	0.21
	2.00	0.00	0.68	0.62	0.22	2.00	0.00	-1.25	-0.84	0.20

(table continues)

Table 2 (continued)

Configuration	Easy Test					Hard Test				
	AL1	AL2	BL1	BL2	CL	AH1	AH2	BH1	BH2	CH
7	1.10	0.00	0.68	-0.03	0.18	0.80	0.90	0.53	-2.00	0.16
	2.00	1.60	0.11	1.02	0.19	0.90	1.10	-0.03	0.49	0.19
	1.20	0.00	0.62	-1.04	0.17	2.00	1.40	1.52	-0.45	0.16
	1.20	0.00	-0.17	0.68	0.19	1.40	1.00	1.35	-0.84	0.20
	0.50	0.00	0.26	0.62	0.16	0.60	1.60	0.68	0.95	0.18
	1.40	1.80	0.40	1.35	0.20	1.00	1.20	1.02	-1.04	0.22
	1.00	0.00	0.18	0.04	0.16	1.80	1.10	0.04	0.40	0.16
	1.00	0.00	-0.66	-1.46	0.16	1.00	0.80	-0.52	0.11	0.18
	1.10	0.00	-0.38	0.18	0.18	1.10	0.90	0.49	-0.10	0.20
	0.90	0.00	-0.84	0.11	0.20	0.60	1.10	0.74	0.32	0.20
	0.70	0.00	-2.00	-0.24	0.20	1.10	0.70	0.32	-0.38	0.20
	0.90	0.00	-0.45	-1.25	0.20	0.80	1.00	-0.45	0.04	0.21
	0.80	0.00	0.53	-0.17	0.16	0.50	1.00	0.26	0.74	0.21
	1.20	0.00	-0.59	-2.00	0.22	1.00	0.70	0.95	0.53	0.16
	0.60	0.00	-0.52	0.32	0.17	1.20	1.00	-1.04	1.02	0.17
	0.80	0.00	-1.04	0.53	0.20	1.00	0.80	-0.59	-1.25	0.19
	1.10	0.00	-0.31	-0.31	0.20	1.20	0.50	-2.00	-0.24	0.16
	1.10	0.00	1.52	-0.84	0.20	1.60	1.80	0.62	1.52	0.19
	0.80	0.00	-0.10	-0.10	0.16	0.80	0.80	-0.66	-0.66	0.20
	0.90	0.00	-0.24	0.95	0.16	0.90	0.90	-0.84	-0.03	0.18
	0.60	0.00	-1.25	0.49	0.22	0.70	1.20	-1.46	-0.31	0.16
	1.00	0.00	0.04	0.26	0.21	1.30	0.90	-1.25	0.18	0.17
	1.60	1.20	-1.46	-0.45	0.20	0.80	1.00	-0.24	-0.52	0.17
	1.00	0.00	1.35	-0.66	0.22	0.70	2.00	0.18	0.62	0.22
	0.90	0.00	0.49	-0.52	0.21	0.90	1.20	-0.10	0.26	0.21
	1.30	2.00	1.02	1.52	0.21	1.10	0.80	-0.17	-0.59	0.16
	1.80	1.40	0.95	0.40	0.16	1.20	0.60	-0.38	-1.46	0.20
	0.70	0.00	-0.03	0.74	0.18	1.10	0.60	0.11	-0.17	0.20
	1.00	0.00	0.32	-0.38	0.17	1.00	1.30	0.40	1.35	0.22
	0.80	0.00	0.74	-0.59	0.19	0.90	1.10	-0.31	0.68	0.20
8	0.80	0.00	1.02	-0.03	0.17	0.80	0.00	-0.45	-0.52	0.18
	0.80	0.00	0.40	-0.24	0.16	1.60	1.60	1.35	1.35	0.22
	1.10	0.00	-0.59	-0.59	0.21	0.70	0.00	1.02	0.26	0.16
	1.00	0.00	0.68	0.32	0.20	0.90	2.00	-1.25	-1.25	0.16
	1.20	0.00	0.11	-0.84	0.17	0.70	0.00	0.95	-0.45	0.20
	1.80	0.00	-2.00	-1.04	0.19	1.40	1.40	0.32	0.32	0.16
	1.40	0.00	-0.17	-0.66	0.22	1.00	0.00	0.11	0.18	0.18
	0.70	0.00	0.32	0.49	0.16	1.10	0.00	1.52	-0.84	0.20
	1.60	0.00	0.18	0.11	0.20	1.80	0.00	-2.00	-0.31	0.16
	1.00	0.00	-0.84	-1.46	0.20	1.00	0.00	-1.04	0.40	0.20
	1.10	0.00	-0.10	0.04	0.16	1.00	0.00	-0.24	0.62	0.19
	0.80	0.00	0.49	-2.00	0.20	1.20	0.00	-0.17	-1.46	0.21
	0.80	0.00	-0.38	-0.45	0.20	1.10	0.00	-0.38	-0.17	0.20
	1.00	0.00	-1.25	0.74	0.22	0.90	0.00	0.62	-0.03	0.20
	1.00	0.00	0.53	-0.38	0.18	1.20	0.00	0.04	-2.00	0.22
	1.10	0.00	0.95	0.95	0.16	1.10	0.00	-0.31	0.53	0.19
	0.50	0.00	0.04	-0.31	0.16	1.20	0.00	-1.46	1.52	0.17
	0.90	0.00	0.62	0.62	0.20	0.90	0.00	-0.03	-0.59	0.20
	1.20	0.00	-0.31	-1.25	0.20	0.60	0.00	0.40	-1.04	0.18
	0.60	0.00	0.26	0.18	0.19	1.00	1.80	0.49	0.49	0.21
	2.00	0.00	-0.66	0.53	0.18	1.30	1.30	-0.10	-0.10	0.16
	1.10	0.00	-1.04	1.02	0.21	0.80	0.00	-0.52	0.04	0.17
	0.90	0.00	1.35	-0.10	0.18	0.80	0.00	0.26	-0.38	0.16
	0.70	0.00	-0.03	1.35	0.20	1.10	0.00	-0.59	0.11	0.20
	0.90	0.00	-1.46	0.40	0.17	1.00	0.00	0.18	0.74	0.20
	1.20	0.00	1.52	0.26	0.16	0.90	0.00	0.68	1.02	0.19
	0.60	0.00	-0.45	0.68	0.22	0.80	0.00	0.53	0.95	0.17
	1.30	0.00	-0.52	-0.52	0.19	2.00	0.00	-0.66	-0.66	0.16
	1.00	0.00	-0.24	1.52	0.16	0.60	0.00	-0.84	0.68	0.21
	0.90	0.00	0.74	-0.17	0.21	0.50	0.00	0.74	-0.24	0.22

(table continues)

Table 2 (continued)

	Easy Test					Hard Test				
Configuration	AL1	AL2	BL1	BL2	CL	AH1	AH2	BH1	BH2	CH
9	0.50	0.00	-2.00	-0.38	0.18	0.90	0.00	-0.45	-0.10	0.21
	0.90	0.00	-0.24	-0.66	0.20	0.90	0.00	-1.04	-2.00	0.16
	0.60	0.00	0.40	-0.10	0.20	1.10	0.00	-0.66	-1.25	0.17
	0.70	0.00	-0.59	-0.31	0.20	0.60	0.00	-0.52	-0.17	0.22
	0.70	0.00	-1.46	1.35	0.21	0.70	0.00	0.04	-0.84	0.22
	0.80	0.00	-1.04	-0.52	0.19	1.00	0.00	-0.24	-0.66	0.20
	0.80	0.00	0.53	-0.84	0.20	0.80	0.00	-0.31	-0.59	0.17
	0.80	0.00	-0.84	-0.45	0.17	0.80	0.00	-1.46	-1.04	0.20
	1.20	0.00	-0.17	-0.03	0.22	2.00	0.00	0.40	-0.31	0.20
	0.90	0.00	-1.25	0.32	0.16	1.00	0.00	1.52	0.74	0.16
	0.90	0.00	0.68	0.04	0.20	1.20	0.00	0.68	-0.45	0.22
	0.60	0.00	-0.31	-0.24	0.16	1.10	0.00	0.18	-0.24	0.18
	0.90	0.00	0.26	-0.59	0.19	0.90	0.80	-0.84	1.02	0.16
	1.00	0.00	-0.52	-1.25	0.21	1.30	0.00	-0.10	-0.52	0.16
	1.00	0.00	0.62	0.11	0.22	1.40	0.00	0.74	0.04	0.20
	1.00	0.00	0.95	0.26	0.18	0.80	0.50	-0.59	0.62	0.16
	1.00	0.00	1.52	0.62	0.17	1.80	0.70	-0.17	0.95	0.17
	1.10	0.00	-0.03	0.53	0.16	1.20	0.00	1.02	0.18	0.16
	1.10	0.00	-0.45	0.49	0.16	0.70	0.00	-2.00	-1.46	0.18
	1.00	0.00	0.49	0.68	0.17	1.10	0.00	0.11	0.32	0.20
	1.10	0.00	-0.66	0.40	0.16	1.00	0.00	0.32	0.40	0.18
	1.10	0.00	0.11	-1.46	0.20	0.60	0.00	-0.38	-0.03	0.20
	0.80	0.00	0.74	-1.04	0.16	0.80	0.00	0.53	0.11	0.20
	1.20	0.00	1.02	-2.00	0.21	0.90	0.00	0.95	0.49	0.19
	1.20	0.00	-0.10	0.95	0.22	1.60	0.00	0.26	0.53	0.20
	1.30	0.00	0.04	1.52	0.18	0.50	0.00	-1.25	-0.38	0.21
	1.40	0.00	0.32	0.18	0.20	1.10	0.00	1.35	0.26	0.19
	1.60	0.00	0.18	-0.17	0.16	1.20	0.00	-0.03	0.68	0.16
	1.80	0.00	-0.38	0.74	0.20	1.00	1.00	0.49	1.52	0.19
	2.00	0.00	1.35	1.02	0.19	1.00	0.90	0.62	1.35	0.21
10	0.50	0.60	0.32	-0.10	0.16	0.50	1.60	0.32	0.53	0.16
	0.60	0.50	0.04	-0.45	0.21	0.60	1.30	-0.84	-0.84	0.16
	0.60	0.60	-0.10	0.11	0.19	0.60	1.40	-0.66	-1.04	0.20
	0.70	0.70	-0.59	0.32	0.20	0.70	2.00	-0.52	0.26	0.18
	0.70	0.70	0.74	0.62	0.19	0.70	1.80	0.11	0.62	0.16
	0.80	0.00	0.53	-2.00	0.21	0.80	0.00	-0.59	-0.66	0.17
	0.80	0.00	0.68	-1.46	0.17	0.80	0.00	0.62	-0.45	0.21
	0.90	0.00	-0.38	-0.03	0.16	0.90	0.00	0.26	1.02	0.20
	0.80	0.00	0.62	1.35	0.16	0.80	0.00	1.35	-0.17	0.17
	1.00	0.00	-0.66	-0.84	0.16	0.90	0.00	0.40	-0.52	0.16
	0.90	0.00	0.95	-0.59	0.20	0.90	0.00	1.52	-1.46	0.21
	0.90	0.00	1.02	0.40	0.18	0.80	0.00	0.04	0.68	0.16
	0.80	0.00	0.49	-1.04	0.17	0.90	0.00	0.18	0.74	0.20
	0.90	0.00	0.40	-1.25	0.20	1.00	0.00	-0.24	1.35	0.16
	1.00	0.00	-1.46	-0.66	0.20	1.00	0.00	-0.10	0.11	0.18
	1.00	0.00	-0.31	1.02	0.22	1.00	0.00	-0.38	-0.31	0.22
	1.00	0.00	-0.24	0.95	0.16	1.00	0.00	1.02	0.49	0.22
	1.20	0.00	-0.17	0.26	0.16	1.00	0.00	0.74	0.95	0.20
	1.10	0.00	1.35	0.74	0.17	1.10	0.00	0.95	0.32	0.22
	1.10	0.00	-2.00	0.68	0.16	1.10	0.00	-0.03	-1.25	0.17
	1.10	0.00	-0.45	-0.31	0.20	1.10	0.00	-0.45	-0.59	0.19
	1.10	0.00	-0.03	0.49	0.18	1.10	0.00	0.49	-0.38	0.18
	1.00	0.00	0.26	0.53	0.22	1.20	0.00	0.53	0.18	0.20
	1.20	0.00	1.52	0.04	0.22	1.80	0.00	-0.31	-0.03	0.20
	1.20	0.00	0.18	1.52	0.19	1.20	0.00	0.68	-0.10	0.21
	1.30	0.00	-0.84	0.18	0.21	1.30	0.00	-2.00	-0.24	0.16
	1.40	0.00	-1.04	-0.24	0.20	1.40	0.00	-1.25	0.04	0.20
	1.60	0.00	-1.25	-0.38	0.18	1.60	0.00	-0.17	-2.00	0.19
	1.80	0.00	0.11	-0.17	0.20	1.20	0.00	-1.46	1.52	0.19
	2.00	0.00	-0.52	-0.52	0.20	2.00	0.00	-1.04	0.40	0.20

(table continues)

Table 2 (continued)

Configuration	Easy Test					Hard Test				
	AL1	AL2	BL1	BL2	CL	AH1	AH2	BH1	BH2	CH
11	0.80		0.68		0.18	0.90		0.62		0.16
	2.00		0.18		0.20	0.90		-0.03		0.18
	1.00		0.62		0.18	2.00		-0.31		0.17
	1.10		-0.17		0.20	1.00		1.35		0.20
	0.50		0.32		0.16	0.60		1.02		0.19
	1.20		0.40		0.20	1.00		0.95		0.20
	1.00		0.11		0.16	0.80		-0.84		0.16
	1.20		-0.66		0.16	1.00		-2.00		0.18
	1.20		-0.31		0.18	1.10		0.26		0.20
	0.90		-0.84		0.20	0.60		-0.10		0.22
	1.10		-2.00		0.19	1.10		0.32		0.22
	0.60		-0.45		0.19	1.00		-0.45		0.20
	0.80		0.49		0.16	0.50		0.49		0.21
	1.00		-0.59		0.22	0.70		0.68		0.16
	0.90		-0.52		0.17	1.00		-1.46		0.17
	0.80		-1.25		0.20	0.80		-0.59		0.19
	0.90		-0.24		0.20	1.20		-0.52		0.16
	0.70		1.52		0.20	1.60		0.53		0.20
	0.70		-0.10		0.16	0.90		-0.17		0.19
	1.10		-0.38		0.16	1.80		0.04		0.17
	0.60		-1.04		0.21	0.70		-1.04		0.16
	1.60		0.95		0.21	1.30		-1.25		0.18
	1.80		-1.46		0.20	0.80		-0.24		0.16
	1.00		1.02		0.22	1.20		0.18		0.20
	0.90		0.53		0.22	0.90		0.74		0.21
	1.30		1.35		0.21	1.10		-0.66		0.16
	1.40		0.04		0.16	0.80		-0.38		0.20
	0.80		-0.03		0.17	1.10		0.11		0.21
	1.00		0.26		0.17	1.40		0.40		0.20
	1.10		0.74		0.19	1.20		1.52		0.22

Note: AL1 and BL1 are discrimination and difficulty for Trait 1 on the easy test.

AL2 and BL2 are discrimination and difficulty for Trait 2 on the easy test.

AH1 and BH1 are discrimination and difficulty for Trait 1 on the hard test.

AH2 and BH2 are discrimination and difficulty for Trait 2 on the hard test.

Base item parameters are before ± 0, ± 1, or ± 2 are added to the difficulty parameters.

For Configuration 3, it was assumed that both traits were measured by both tests, and a high (.75) correlation existed between the two traits. Discrimination and difficulty were correlated .4 in Trait 1. A test that involves both mathematics and reading might produce such a configuration.

Configuration 4 was the same as Configuration 1 except that a high correlation (.9) was assumed between traits. This configuration represents a test, such as language mechanics, where Trait 1 is reading ability and Trait 2 is punctuation ability.

A test such as reading comprehension might involve both reading and vocabulary as separate traits. A high correlation between traits would probably exist. Such a situation was assumed for Configuration 5, with difficulty and discrimination highly correlated across traits.

In a test such as social studies, where two traits might be reading ability and ability to understand graphs, some items usually pertain to graphs only, some require reading only, and a few may require both reading ability and an understanding of graphs. For Configuration 6, some items were assumed to measure only Trait 1, while having zero discrimination in Trait 2. Some items were assumed to measure Trait 2 only, with zero discrimination in Trait 1. A few items were assumed to measure both traits.

Configuration 7 involves a situation in which Trait 2 is measured in the easier test by a few items and in the harder test by most or all of the items. For the easier test, discrimination was high for the items in both traits, while all other items had a range of discrimination in Trait 1 and zero discrimination in Trait 2. The harder test had a range of discrimination in both traits. This might occur if knowledge of fractions was one trait measured by a few items, and all other items measured another trait.

Decimals are introduced after more basic concepts (number, addition, etc.) have been taught. Therefore, items measuring knowledge of decimals usually do not occur in the lowest or easiest levels of a series of tests designed to cover kindergarten through high school. Configuration 8 was chosen to represent a situation where the second trait is measured only by a few items in the harder test. For the harder test, these few items were assumed to have equal difficulty in both traits, while discrimination was assumed to be medium for Trait 1 and high for Trait 2. All other items include a range of discrimination values in Trait 1 and zero discrimination in Trait 2. A low correlation (.3) between the two traits was assumed.

In Configuration 9, Trait 2 was measured only by a few items in the harder test with moderate correlation between traits. Difficulty and discrimination were assumed to be moderately correlated in Trait 1 for both tests and highly correlated in Trait 2 for the harder test. Difficulty had medium correlation for the harder test. This was chosen to represent a test such as mathematics computation, where Trait 2 is ability with fractions (measured by a few items in the harder test), and Trait 1 is ability to solve nonfraction items (measured at both levels).

For Configuration 10 it was assumed that both traits were measured by both tests. The correlation between traits was .5. A few items were assumed to have low discrimination in Trait 1 for both tests, low discrimination in Trait 2 for the easier test and high discrimination in Trait 2 for the harder test. All other items were assumed to have high, medium, or low discrimination in Trait 1 and zero discrimination in Trait 2. This configuration was chosen as being similar to a test such as mathematics computation, where Trait 2 is ability in decimal problems, and Trait 1 is ability in computing all other problems.

Configuration 11 was the unidimensional situation. This configuration was simulated by setting m = 1 in the multidimensional model used to generate the data.

Data conditions. Combining the previously described configurations resulted in 33 data conditions: 11 (item parameter/correlation configurations) x 3 ($\bar{b}_D - \bar{b}_E$) values.

Configuration Groupings. Configurations 1 through 10 were generated by a multivariate model, and therefore are multidimensional. However, for the sake of clarity, the configurations were grouped into four groups based on the discrimination values. Items not measuring a trait were assigned zero discrimination values, because, if an item does not measure a trait it presumably will not discriminate on that trait. It was assumed that Trait 2 was not measured at all by some tests or was measured by only a few items (five of the 30 items). These tests, although not strictly unidimensional, may be considered to be essentially unidimensional. Tests that were assumed to measure both traits with all items were assigned nonzero discriminations and are clearly multidimensional.

In Configurations 1 through 5 and Configuration 7, Test 2, every item measured both dimensions, and over half the items (17) measured each trait in Configuration 6. Therefore, both tests of Configurations 1 through 6, and Test 2, Configuration 7 were multidimensional. Hence, Group M1, consisting of configurations having both tests clearly multidimensional, was composed of Configurations 1 through 5. For all items in both tests, this group had nonzero discriminations for both traits. Configuration 6 was the only configuration placed in Group M2 because it was the only configuration with about half of the items measuring each dimension. Group M2 was also considered to be a multidimensional group.

Another group (Group U) was considered as a group of configurations with both tests essentially unidimensional. Group U consisted of Configurations 8, 9, 10, and 11. For Test 1 of Configurations 8 and 9, Trait 2 was not measured at all; hence, it had zero discriminations for all items. Therefore, Test 1 for Configurations 8 and 9 was considered to be essentially unidimensional. Similarly, Test 1 of Configurations 7 and 10, and Test 2 of Configurations 8, 9, and 10 were all essentially unidimensional because 25 of 30 items did not involve the second trait.

Test 1, Configuration 7 was essentially unidimensional and Test 2 was multidimensional. Configuration 7 was placed into a group by itself, Group M3.

Simulated data. The simulated data sets included three groups of simulated examinees for each of the two traits: 2,000 of low ability, 2,000 of middle ability and 2,000 of high ability. The 2,000 ability or theta values for each of the three ability levels were generated using the IMSL multivariate normal random deviate generator, GGNSM (IMSL, 1979). In each case a normal distribution was assumed. (Mean = -.57; SD = 1.0 for the low ability group. Mean = 0.0; SD = 1.0 for the medium ability group. Mean = 0.57; SD = 1.0 for the high ability group.) These differences in mean ability were chosen to be similar to the differences between ability levels in published tests. A total group of all 6,000 observations pooled was also formed. For the 33 data conditions (11 configurations x 3 differences in mean difficulty between tests), response vectors were generated for the low, medium and high ability groups for tests of 30 items each.

Anchor test. Recall that for vertical equating, Test E was the easier test, and Test D was the harder test of each pair of tests to be equated. For horizontal equating, Tests E and D are equivalent. For each of the 33 data conditions, a 30 item anchor test was chosen using traditional difficulty or p-values. Each anchor test was composed of the 15 items from Test E with the lowest p-values and the 15 items from Test D with the highest p-values.

Response vectors. Using the prespecified parameters and the multidimensional model, the $P_i(\theta_{j1},\theta_{j2})$ were computed for each observation. Using these $P_i(\theta_{j1},\theta_{j2})$ values, binary responses (u_{ijk}) were generated for each item (i), examinee (j) and test (k) combination, where u_{ijk} is 1 (a correct response) if a random number is less than or equal to $P_i(\theta_{j1},\theta_{j2})$ or 0 (a wrong response), otherwise. Random numbers were generated from a uniform random number generator distribution provided by the IMSL subroutine, GGUBS (IMSL, 1979).

Responses were generated for all simulated examinees for both tests (E and D). For item parameter estimation, only the responses to Test E were used from the low ability group, only the responses to Test D were used from the high ability group, and only the responses to the anchor test were used from the medium ability group. Responses for all simulated examinees to both tests were used to examine the accuracy of the equating.

Data verification

Means and standard deviations of number-correct scores, item difficulties (p-values), and KR-20 test reliability coefficients from the simulated tests were examined in order to verify that the simulations were veridical.

Parameter Estimation

Before equating can be done using an IRT model, the estimated item parameters ($\hat{a}$, $\hat{b}$, $\hat{c}$), and the ability

estimates ($\hat{\theta}$), must be derived from the test response vectors. These estimates are usually obtained using computer programs, such as LOGIST (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976; Wingersky, Barton, & Lord, 1982). With the LOGIST program, the estimates are computed by a modified maximum likelihood procedure.

For each of the 33 data conditions, item parameters were estimated in two separate runs with LOGIST (Wingersky, Barton, & Lord, 1982). LOGIST has a not-reached option to exclude from the parameter estimation those items which are beyond the examinee's last marked response on the test. These items are not considered incorrect, but are ignored as if they were not part of the test that the examinee took. Table 3 shows the LOGIST runs used for parameter estimation.

Table 3

LOGIST Runs for Parameter Estimation

LOGIST Run	Sample	N	Items	
			Test E (30 items)	Test D (30 items)
1	low ability	2,000	responses[1]	NR[2]
	medium ability	2,000	15 NR 15 responses	15 NR 15 responses
2	medium ability	2,000	15 NR 15 responses	15 NR 15 responses
	high ability	2,000	NR	responses

Note: [1]"Responses" means the item responses were used.
[2]"NR" means "not reached" was used in place of item responses.

Item parameters for Test E and the anchor were estimated using the responses from the low and medium ability groups. The 30 items from Test D were considered not-reached (NR) items for the low ability group. The 15 Test D and the 15 Test E items not in the anchor were considered NR items for the medium ability group. Similarly, item parameters for Test D and the anchor were estimated using the responses from the medium and high ability groups combined. Items from Test E were considered NR for the high ability group. The 30 Test D and Test E items not in the anchor were NR for the medium ability group.

This allowed combined samples of 4,000 examinees; a sample size found to be adequate for obtaining very stable item parameter estimates (Yen, 1983). Because separate pairs of LOGIST runs were made for each of the 33 data conditions, the result was two sets of estimated item parameters and estimated abilities per condition. For each of these 33 conditions, the accuracy of the parameter estimates was examined; equating was performed, and the accuracy of the equating was examined.

Accuracy of the Parameter Estimation

Although in real life situations the true parameters are not known, an advantage of using simulated data is that the "true" item parameters are known; the "true" item parameters are those used to generate the data. Hence comparisons between true and estimated parameters can be made, and such comparisons can be used to examine the accuracy of the estimation procedure within levels.

Within level comparisons. The accuracy of the estimation procedure was examined within levels by comparing the estimated item parameters from Tests E and D to both sets of true item parameters for Traits 1 and 2 using correlations between the appropriate parameters and by comparing true thetas to estimated thetas. Estimated b_i

(a_i) values were compared to each of the two sets of true b_i (a_i) values, b_{i1} (a_{i1}) and b_{i2} (a_{i2}) for both traits. The estimated c_i values were compared to the one set of true c_i values, and the estimated thetas ($\hat{\theta}$) were compared with the true thetas (θ_1 and θ_2) for the two traits.

Item Response Theory Equating

In IRT equating, scores from two tests are considered equivalent if they correspond to the same level of estimated ability, $\hat{\theta}$. In other words, two tests are considered equated if the same ability estimate, $\hat{\theta}$, can be expected for any examinee on both tests after equating. The basic principle behind equating ability estimates is that if the data fit the IRT model, then an estimate of an examinee's ability can be obtained that is independent of the set of items to which the examinee responds. The ability estimates for each examinee obtained from two tests will be identical except for sampling error.

Equating using an IRT model involves first estimating item and ability parameters, followed by equating. Equating consists of placing the item parameter estimates for the tests to be equated onto the same scale, and then placing the ability estimates, $\hat{\theta}$, onto this scale. A transformation is chosen such that scores from the two tests are equated if they correspond to the same estimated level of the latent trait ability underlying both tests.

Using the estimated item parameters from the parameter estimation runs, the means and standard deviations of the b_i values for the anchor items were equated as follows:

a. $M_{\hat{b}1}$, $SD_{\hat{b}1}$, $M_{\hat{b}2}$, and $SD_{\hat{b}2}$ were calculated where $M_{\hat{b}1}$ and $SD_{\hat{b}1}$ are the mean and standard deviation of the item difficulty parameter, b_i, for the anchor items as estimated in Run 1, and $M_{\hat{b}2}$ and $SD_{\hat{b}2}$ are the mean and standard deviation of the b_i as estimated in Run 2.

b. Let $K_1 = SD_{\hat{b}1}/SD_{\hat{b}2}$, and $K_2 = M_{\hat{b}1} - K_1 M_{\hat{b}2}$.

c. The estimated parameters for a, b, and c for Run 2 were then adjusted as follows: $\hat{c}_{2^*} = \hat{c}_2$, $\hat{a}_{2^*} = \hat{a}_2/K_1$, and $\hat{b}_{2^*} = \hat{b}_2K_1 + K_2$. The estimated item parameters for Run 2 were then on the same scale as those for Run 1. A second set of LOGIST runs (using fixed equated item parameters from the parameter estimation) was made. Because fixed equated item parameters were used, the result of these runs was that the $\hat{\theta}$ of Test D were equated to the $\hat{\theta}$ of Test E.

Accuracy of the Equating

The accuracy of the equating was examined in light of three effects: the effect of each multidimensional data configuration, the size of the difference between mean difficulty on the two tests that were equated, and the effect of the accuracy (or inaccuracy) of the item parameter estimates.

Between level comparisons. For each of the 33 data conditions, thetas (θ) were estimated in two separate runs with item parameters ($\hat{a}$, $\hat{b}$, $\hat{c}$) fixed (and equated). In Run 1, thetas were estimated for Test E using item responses for all three ability levels on Test E and not reached (NR) for Test D. Item parameters were fixed at the values estimated in the parameter estimation LOGIST runs. In Run 2, thetas were estimated for Test D, using item responses on Test D, NR on Test E, and fixed item parameters. Because equated fixed item parameters were used, estimated thetas for Tests E and D were equated after these LOGIST runs. The accuracy of the equating was then examined.

The accuracy of the equating was examined by comparing the estimated thetas for all simulated examinees on the easier test with the estimated thetas for all simulated examinees on the harder test. Standardized root mean squared differences (SRMSD), standardized differences between means (SDM), the ratio of the standard deviations, correlations, and degree of tau equivalence were used.

The RMSD between pairs of estimated thetas, scaled by the variances of the two equated trait estimates being compared, was computed as follows:

$$SRMSD = \sqrt{\frac{\frac{1}{n}\sum_{j=1}^{n}(\hat{\theta}_{Ej} - \hat{\theta}_{Dj})^2}{(S_E^2 + S_D^2)/2}}, \qquad (3)$$

where $\hat{\theta}_{Ej}$, $\hat{\theta}_{Dj}$, S_E^2, and S_D^2 are the ability estimates for person j and the variances of the theta estimates for Tests E and D, respectively.

The standardized difference between means (SDM), an estimate of overall bias between the estimated thetas, is the difference in mean thetas for Test E and Test D divided by a pooled estimate of the standard deviations:

$$SDM = \frac{\bar{\hat{\theta}}_E - \bar{\hat{\theta}}_D}{(S_E^2 + S_D^2)^{1/2}}, \qquad (4)$$

where $\bar{\hat{\theta}}_E$, $\bar{\hat{\theta}}_D$, S_E^2, and S_D^2 are means of the estimated thetas and variances of Tests E and D, respectively.

A third method used to check the accuracy of equating was the examination of the degree of tau-equivalence of the equated, estimated thetas (Yen, 1983). Two scores, X_1 and X_2 are tau-equivalent if $E(X_1|T) = E(X_2|T)$, where T is the true score. Local equating bias of $\hat{\theta}_E$ and $\hat{\theta}_D$ is defined as $B(\hat{\theta}_E,\hat{\theta}_D|T) = E(\hat{\theta}_E|T) - E(\hat{\theta}_D|T)$. To estimate local bias, examinees were rank ordered on the basis of the average of the two trait values and grouped into five cells with 20

percent of the observations in each cell. The estimate of local equating accuracy that was used is

$$B_m(\hat{\theta}_E,\hat{\theta}_D|T) = \frac{\bar{\hat{\theta}}_{Em} - \bar{\hat{\theta}}_{Dm}}{\sqrt{(S_E^2 + S_D^2)/2}} \qquad (5)$$

for cell m, where $\bar{\hat{\theta}}_{Em}$ and $\bar{\hat{\theta}}_{Dm}$ are the means of $\hat{\theta}_E$ and $\hat{\theta}_D$ for cell m. The standardized difference between means was found for each of the five cells. The average of the absolute values of the standardized mean differences for the five cells was used as a summary measure of local bias. Note that this is not really bias in the conventional, statistical sense, but actually is local consistency of the scaled, estimated thetas.

RESULTS

True Item Parameters

All attained correlations among the item parameters used for generating the two-trait data were within $\pm$.10 of the desired correlations. Hence, the attained correlations appeared to be acceptable.

Simulated Thetas

The attained correlations between generated thetas on the two traits were within .03 of the desired correlations. For all three ability levels mean thetas were within .06 of the desired mean ability, and the standard deviation of the true thetas were within .04 of the desired standard deviations. The simulated thetas appeared to be acceptable.

Simulated Item Responses

Table 4 contains means and standard deviations of number-correct scores for the pairs of simulated tests. The simulated tests appeared to be realistic. As was expected, Tests E and D differed a great deal in difficulty for the $\bar{b}_D - \bar{b}_E = 2$ conditions, and had very similar (number-correct) difficulties when $\bar{b}_D - \bar{b}_E = 0$. Also as expected, Test E appeared easiest and Test D hardest when $\bar{b}_D - \bar{b}_E = 2$. (Note that the easiest test of Configurations 1 through 5 had a loss of more than 1000 examinees due to perfect scores. Thetas for examinees with perfect or zero scores cannot be estimated by maximum likelihood methods; hence LOGIST automatically terminates such runs because results from a run with a large percentage of deleted examinees would be inaccurate.)

Table 5 contains KR-20 values for the pairs of tests. The KR-20 values ranged from .74 to .94. The overall trend appeared to be that the hardest test for each configuration ($\bar{b}_D - \bar{b}_E = 2$, Test D) had the smallest KR-20. Most of the tests appeared to be acceptably reliable.

Table 4

Means and Standard Deviations of Number-Correct Scores

Configuration	$\bar{b}_p - \bar{b}_g$	Test	Ability Level: Low + Medium Mean	Low + Medium S.D.	Medium + High Mean	Medium + High S.D.
1	2	E	24.03	6.05	*	*
	1		19.87	7.13	24.47	5.79
	0		15.53	7.10	20.53	7.06
	0	D	15.78	6.92	20.56	6.70
	1		11.76	6.04	16.26	6.93
	2		9.07	4.76	12.56	6.30
2	2	E	23.20	7.08	*	*
	1		19.71	8.03	23.87	6.80
	0		15.79	8.00	20.33	7.89
	0	D	15.99	7.76	20.42	7.48
	1		12.40	6.94	16.44	7.75
	2		9.63	5.69	12.93	7.15
3	2	E	23.14	6.92	*	*
	1		19.42	7.81	23.49	6.73
	0		15.75	7.84	20.02	7.77
	0	D	15.50	8.01	19.88	8.11
	1		11.91	6.90	15.93	8.04
	2		9.42	5.64	12.49	7.17
4	2	E	23.25	7.28	*	*
	1		19.58	8.24	23.66	7.15
	0		16.03	8.20	20.30	8.12
	0	D	16.15	8.03	20.33	7.93
	1		12.45	7.23	16.53	8.06
	2		9.84	5.88	13.19	7.46
5	2	E	22.85	6.74	*	*
	1		19.55	7.45	23.22	6.59
	0		15.93	7.47	19.98	7.43
	0	D	15.88	7.40	19.93	7.44
	1		12.70	6.71	16.49	7.52
	2		10.16	5.54	13.29	6.82

(table continues)

Table 4 (continued)

Configuration	$\bar{b}_D - \bar{b}_E$	Test	Ability Level: Low + Medium Mean	Low + Medium S.D.	Medium + High Mean	Medium + High S.D.
6	2	E	22.69	5.53	25.46	4.40
	1		19.70	6.05	23.01	5.46
	0		16.84	5.98	20.27	5.94
	0	D	15.93	6.02	19.41	6.01
	1		12.99	5.54	16.32	6.20
	2		10.62	4.83	13.35	5.61
7	2	E	21.53	5.94	24.40	5.01
	1		18.68	6.38	21.98	6.01
	0		15.56	6.14	18.97	6.25
	0	D	15.41	7.42	19.99	7.48
	1		12.08	6.58	16.39	7.54
	2		9.32	5.11	12.39	6.58
8	2	E	21.95	5.71	24.51	4.94
	1		19.22	6.30	22.32	5.64
	0		16.52	6.39	19.53	6.22
	0	D	16.51	6.15	19.66	6.05
	1		13.66	5.79	16.88	6.17
	2		11.22	5.09	14.06	5.92
9	2	E	21.94	6.04	24.69	5.15
	1		19.04	6.49	22.17	6.04
	0		16.15	6.52	19.40	6.44
	0	D	15.33	6.43	18.72	6.62
	1		12.69	5.66	15.77	6.51
	2		10.55	4.85	12.98	5.82
10	2	E	22.10	5.94	24.94	4.87
	1		19.32	6.45	22.54	5.72
	0		16.47	6.49	19.97	6.36
	0	D	16.25	6.36	19.76	6.36
	1		13.30	5.84	16.54	6.42
	2		10.60	5.02	13.58	5.99
11	2	E	21.90	5.99	24.60	5.05
	1		19.04	6.54	22.16	5.89
	0		16.31	6.33	19.59	6.35
	0	D	16.45	6.37	19.74	6.32
	1		13.66	6.00	16.65	6.38
	2		11.37	5.21	14.03	6.04

* Incomplete LOGIST run due to too many perfect scores.

Table 5

KR-20 Values of Number-Correct Scores

Configuration	$\bar{b}_p - \bar{b}_E$	Test	Ability Level: Low + Medium	Ability Level: Medium + High
1	2	E	.91	*
	1		.91	.91
	0		.90	.91
	0	D	.89	.91
	1		.85	.89
	2		.76	.86
2	2	E	.94	*
	1		.94	.93
	0		.92	.94
	0	D	.92	.93
	1		.89	.92
	2		.83	.90
3	2	E	.93	*
	1		.93	.93
	0		.92	.93
	0	D	.92	.94
	1		.89	.93
	2		.84	.90
4	2	E	.94	*
	1		.94	.94
	0		.93	.94
	0	D	.92	.94
	1		.90	.93
	2		.85	.91
5	2	E	.93	*
	1		.93	.93
	0		.91	.93
	0	D	.91	.93
	1		.89	.92
	2		.83	.89

(table continues)

Table 5 (continued)

KR-20 Values of Number-Correct Scores

Configuration	$\bar{b}_p - \bar{b}_E$	Test	Ability Level: Low + Medium	Ability Level: Medium + High
6	2	E	.86	.84
	1		.86	.87
	0		.84	.86
	0	D	.84	.86
	1		.80	.85
	2		.75	.81
7	2	E	.88	.86
	1		.88	.88
	0		.85	.87
	0	D	.91	.93
	1		.88	.91
	2		.79	.88
8	2	E	.87	.86
	1		.87	.87
	0		.87	.87
	0	D	.85	.87
	1		.83	.86
	2		.78	.84
9	2	E	.88	.88
	1		.88	.88
	0		.87	.88
	0	D	.86	.88
	1		.82	.87
	2		.76	.83
10	2	E	.88	.87
	1		.88	.88
	0		.87	.88
	0	D	.87	.88
	1		.83	.87
	2		.77	.84
11	2	E	.88	.87
	1		.88	.88
	0		.86	.88
	0	D	.86	.87
	1		.84	.86
	2		.78	.84

*Incomplete LOGIST run due to too many perfect scores.

Parameter Estimation. Tables 6 and 7 contain comparisons of the true and estimated parameters. Recall that both tests (E and D) of Group U and Test E of Group M3 were unidimensional tests according to the values of the (generating) item discrimination parameters; both tests of Group M1 and Test D of Group M3 were multidimensional (all items measuring both traits); and both tests of Group M2 were multidimensional (approximately half of the items measuring each trait). Table 6 lists the correlations between the estimated and true item parameters. For Configuration 5 and the unidimensional tests, the correlations between difficulty on Trait 1 (b_1) and estimated difficulty ($\hat{b}$) were .90 and above. For most of the multidimensional tests the correlations were under .80 (mostly between .60 and .78). The only correlations under .59 were in M2, Test D (.26, .36 and .34).

For the correlations between estimated difficulty ($\hat{b}$) and difficulty on Trait 2 (b_2), the best correlations (.87 to .95) were for Configuration 5, the multidimensional configuration with the highest correlation between traits. All other correlations were .77 or less. For most of the unidimensional tests the correlations were below .20. For most of the multidimensional tests the correlations were between .50 and .78.

Note the large differences (.78 to .96) between $r(\hat{b},b_2)$ and $r(\hat{b},b_1)$ for the unidimensional tests. Note also the differences between $r(\hat{b},b_1)$ and $r(\hat{b},b_2)$ for Group M2. For Trait 1, Test E had the higher correlations (.77 and .78) and Test D had the lower (.26 to .36). The reverse was true for Trait 2. Test D had the higher correlations (.67 to .77), and Test E had the lower (.24 to .30).

The correlations between estimated discrimination ($\hat{a}$) and true discrimination on Trait 1 (a_1) followed some of the same patterns as the correlations between estimated and

true difficulty. The highest correlations between $\hat{a}$ and a_1 were for the unidimensional tests as was true for the correlations between $\hat{b}$ and b_1. The correlations for most of the multidimensional tests were less than .72. Group M2, again an exception, had the lowest overall correlations on Trait 1 (.11 to .41).

For the correlations between estimated discrimination ($\hat{a}$) with discrimination on Trait 2 (a_2), the unidimensional test of Group M3 had the largest correlations (.80 and .83). All other correlations were .73 or less. Most of the Group M1 correlations were between .50 and .73. Group M2 correlations were mostly in the .30s, and correlations for Group U were less than .31.

There appeared to be no pattern for the correlations between $\hat{c}$ and c. These correlations fell between -.14 and .80. Some of the poorest correlations were for Group U.

Table 7 lists the correlations between estimated and true trait values. The correlations for Trait 1 ranged from .44 to .92. The unidimensional tests had correlations mostly in the .90s, and the correlations for the multidimensional tests were mostly in the .70s and .80s. Configuration 1 yielded the smallest correlations -- from .44 to .68.

For the multidimensional tests the correlations between $\hat{\theta}$ and θ_2 were approximately equal to the correlations between $\hat{\theta}$ and θ_1 for corresponding conditions. For the unidimensional tests the Trait 2 correlations were all .28 to .56 less than the corresponding Trait 1 correlations.

Table 6

Correlations Between True and Estimated Item Parameters

Conf.	Test	$\bar{b}_D - \bar{b}_R$	$r(\hat{b}, b_1)$	$r(\hat{b}, b_2)$	$r(\hat{a}, a_1)$	$r(\hat{a}, a_2)$	$r(\hat{c}, c)$
1	E	2	.74	.64	.46	.73	.40
		1	.74	.65	.51	.60	.35
		0	.74	.63	.41	.63	.65
	D	0	.69	.71	.56	.58	.43
		1	.68	.74	.46	.54	.60
		2	.68	.73	.54	.52	.64
2	E	2	.74	.62	.63	.44	.40
		1	.74	.62	.64	.50	.45
		0	.59	.51	.56	.44	.69
	D	0	.68	.71	.61	.50	.42
		1	.70	.72	.56	.55	.79
		2	.69	.71	.31	.53	.54
3	E	2	.65	.76	.59	.30	.41
		1	.66	.74	.58	.31	.38
		0	.64	.71	.41	.17	.55
	D	0	.62	.67	.45	.52	.18
		1	.66	.66	.61	.55	.28
		2	.69	.66	.44	.56	.60
4	E	2	.74	.62	.60	.42	.51
		1	.73	.62	.56	.51	.63
		0	.74	.59	.50	.50	.34
	D	0	.71	.67	.41	.69	.56
		1	.70	.70	.48	.66	.72
		2	.69	.71	.41	.50	.57
5	E	2	.93	.95	.62	.64	.22
		1	.91	.93	.71	.69	.39
		0	.90	.93	.51	.53	.59
	D	0	.93	.91	.69	.68	.67
		1	.93	.87	.56	.58	.40
		2	.94	.94	.44	.64	.26
6	E	2	.77	.24	.38	.32	.43
		1	.78	.30	.34	.31	.44
		0	.77	.27	.11	.51	.22
	D	0	.26	.77	.37	.36	-.05
		1	.36	.71	.37	.37	.09
		2	.34	.67	.41	.33	.66
7	E	2	.98	.12	.87	.83	.14
		1	.98	.15	.90	.80	.47
		0	.96	.18	.92	.68	.19
	D	0	.76	.61	.52	.55	.53
		1	.76	.62	.60	.52	.59
		2	.74	.66	.61	.25	.70
8	E	2	.98	.09	.88	*	.49
		1	.97	.13	.85	*	.39
		0	.96	.14	.74	*	.50
	D	0	.95	.12	.84	-.01	.24
		1	.97	.13	.73	-.11	.11
		2	.97	.13	.82	-.02	.51
9	E	2	.99	-.10	.92	*	.35
		1	.99	-.05	.95	*	.28
		0	1.00	-.05	.95	*	.55
	D	0	.96	.66	.91	.23	.21
		1	.94	.68	.90	.30	.42
		2	.93	.72	.89	.29	.55
10	E	2	.96	-.02	.80	-.13	.15
		1	.97	-.01	.78	-.20	.32
		0	.96	-.03	.84	-.31	.44
	D	0	.96	-.01	.77	-.22	.39
		1	.97	-.02	.80	-.18	.49
		2	.97	-.01	.80	-.25	.44
11	E	2	.99	-	.95	-	.06
		1	.99	-	.90	-	.26
		0	.98	-	.94	-	.40
	D	0	.98	-	.90	-	-.13
		1	.99	-	.92	-	.37
		2	.99	-	.90	-	.36

Note: *Discriminations of Trait 2 are all zero.

Table 7

Correlations Between True and Estimated Traits

Conf.	Test	$\bar{b}_p - \bar{b}_g$	N*	$r(\hat{\theta},\theta_1)$	$r(\hat{\theta},\theta_2)$
1	E	2	4850	.44	.46
		1	5277	.67	.65
		0	5711	.67	.68
	D	0	5703	.68	.67
		1	5921	.66	.61
		2	5983	.62	.55
2	E	2	4588	.52	.53
		1	4996	.83	.82
		0	5529	.81	.81
	D	0	5538	.81	.81
		1	5833	.76	.74
		2	5945	.71	.67
3	E	2	4664	.55	.55
		1	5052	.86	.86
		0	5607	.81	.83
	D	0	5473	.82	.85
		1	5800	.80	.83
		2	5932	.71	.75
4	E	2	4558	.56	.56
		1	4915	.90	.90
		0	5448	.88	.88
	D	0	5409	.88	.88
		1	5755	.82	.81
		2	5915	.76	.75
5	E	2	4921	.59	.59
		1	5261	.91	.91
		0	5680	.88	.88
	D	0	5646	.88	.89
		1	5869	.88	.88
		2	5958	.83	.83
6	E	2	5306	.78	.78
		1	5680	.80	.81
		0	5881	.78	.79
	D	0	5932	.80	.79
		1	5976	.75	.77
		2	5993	.67	.70
7	E	2	5469	.90	.58
		1	5720	.91	.57
		0	5929	.90	.54
	D	0	5573	.80	.76
		1	5847	.76	.71
		2	5965	.71	.62
8	E	2	5491	.91	.35
		1	5784	.92	.39
		0	5920	.90	.36
	D	0	5922	.90	.46
		1	5981	.86	.46
		2	5995	.82	.42
9	E	2	5313	.92	.49
		1	5698	.92	.51
		0	5887	.90	.52
	D	0	5882	.90	.55
		1	5949	.85	.53
		2	5988	.79	.47
10	E	2	5416	.91	.54
		1	5681	.91	.54
		0	5900	.90	.56
	D	0	5862	.90	.62
		1	5955	.85	.56
		2	5991	.81	.53
11	E	2	5431	.92	-
		1	5751	.92	-
		0	5885	.91	-
	D	0	5897	.89	-
		1	5981	.84	-
		2	5992	.80	-

Note: N* is total N minus examinees with zero or perfect scores.

Equating. Table 8 shows means, standard deviations and correlations between the estimated thetas for Test E versus Test D. Correlations were .83 or greater for the horizontal equating conditions; .80 to .85 when $\bar{b}_D - \bar{b}_E = 1$; and the poorest equating was when $\bar{b}_D - \bar{b}_E = 2$. The correlations for the multidimensional data sets were as good as for the unidimensional data sets except for Group M3 (equating a multidimensional test to a unidimensional test). Group M3 had the worst equating (correlations .64 to .78).

The standardized root mean squared differences (SRMSD) between thetas from Tests E and D, standardized differences between the means (SDM), ratios of standard deviations and summary measures of local bias are presented in Table 9. Except for Group M3, the SRMSD were mostly in the .50s for the horizontal equating conditions, in the .60s for the $\bar{b}_D - \bar{b}_E = 1$ conditions, and .68 to .79 for the most extreme vertical equating situations. Group M3 had the worst (largest) SRMSD. The ratios between standard deviations were mostly near 1.00. Note the small ratios for Group M3 and the large ratios for Configurations 2 and 4.

The SDM values were small for Configuration 5, Group U and for the horizontal equating condition for most multidimensional configurations. Configurations 3 and 7 were exceptions (SDM values for the $\bar{b}_D - \bar{b}_E = 0$ condition were largest). The SDM values were largest in each configuration for the most extreme vertical equating situation. As for the correlations between estimated thetas and SRMSD, Group M3 had the worst SDM. Test D means were usually higher than Test E means.

The local bias measure tended to increase as the mean difference in difficulty between the tests increased; however, the largest local bias for Group M3 was in the $\bar{b}_D - \bar{b}_E = 0$ condition. Group M3 had the worst overall local bias.

Table 8

Means, Standard Deviations and Correlations of Equated Theta Estimates for Test E versus Test D

Configuration	$\bar{b}_D - \bar{b}_E$	M_E	SD_E	M_D	SD_D	$r(\hat{\theta}_E, \hat{\theta}_D)$
1	0	.28	1.03	.29	.96	.88
	1	.33	.92	.43	.89	.84
	2	*	*	.85	.85	*
2	0	.23	1.02	.17	.91	.87
	1	.27	.96	.37	.77	.85
	2	*	*	.86	.81	*
3	0	.24	.84	.12	1.04	.85
	1	.22	.91	.24	.99	.83
	2	*	*	.51	1.23	*
4	0	.27	.99	.26	.97	.91
	1	.29	.95	.39	.81	.85
	2	*	*	.92	.83	*
5	0	.23	.96	.23	.91	.87
	1	.28	.94	.25	1.03	.83
	2	*	*	.29	1.38	*
6	0	.28	1.07	.20	1.11	.83
	1	.26	.98	.33	.96	.82
	2	.18	.88	.30	.88	.71
7	0	.27	.90	.12	1.17	.77
	1	.36	.94	.44	.96	.78
	2	.39	.87	.56	1.04	.65
8	0	.22	1.00	.20	1.08	.84
	1	.30	.95	.26	1.02	.81
	2	.22	.94	.31	.97	.78
9	0	.20	1.02	.22	1.09	.84
	1	.21	1.00	.22	1.08	.81
	2	.18	.90	.26	1.04	.70
10	0	.26	.98	.26	1.03	.86
	1	.27	.93	.32	.88	.83
	2	.23	.88	.31	.94	.76
11	0	.25	1.01	.25	.98	.85
	1	.26	.94	.31	.86	.83
	2	.18	.92	.23	.92	.75

*Incomplete LOGIST runs due to too many perfect scores.

Note: Test E is the easier test and Test D is the harder test.

Table 9

Standardized Root Mean Squared Differences, Standardized Differences Between Means, Ratios of Standard Deviations and Local Bias[a] for Test E versus Test D

Configuration	$\bar{b}_D - \bar{b}_E$	SRMSD	SD_E/SD_D	SDM	Local Bias
1	0	.49	1.07	−.01	.05
	1	.58	1.04	−.11	.11
	2	b	b	b	b
2	0	.52	1.12	.07	.07
	1	.60	1.24	−.11	.16
	2	b	b	b	b
3	0	.60	.81	.12	.12
	1	.59	.93	−.03	.09
	2	b	b	b	b
4	0	.43	1.02	.01	.03
	1	.58	1.16	−.11	.14
	2	b	b	b	b
5	0	.52	1.06	.00	.05
	1	.60	.92	.03	.08
	2	b	b	b	b
6	0	.59	.96	.06	.07
	1	.60	1.03	−.07	.07
	2	.77	1.00	−.13	.13
7	0	.74	.77	.14	.25
	1	.67	.98	−.08	.08
	2	.87	.83	−.18	.18
8	0	.58	.93	.02	.04
	1	.62	.93	.04	.05
	2	.68	.98	−.09	.09
9	0	.56	.93	−.02	.05
	1	.62	.92	−.01	.08
	2	.79	.87	−.08	.17
10	0	.53	.96	−.01	.04
	1	.59	1.06	−.05	.05
	2	.71	.94	−.09	.09
11	0	.54	1.03	−.01	.04
	1	.60	1.09	−.05	.09
	2	.71	1.01	−.05	.05

[a]Local bias is not really bias in the conventional, statistical sense, but actually is local consistency of scaled estimated thetas.

[b]Incomplete LOGIST run due to too many perfect scores on Test E.

Note: A positive SDM indicates that the Test E mean is higher than the Test D mean.

Local equating bias by ability quintile is shown in Table 10. Recall that this measure is not really bias in the conventional, statistical sense, but actually is local consistency of scaled estimated thetas.

The conditions with large SDM and large overall amounts of local bias in Table 9 also tended to have large amounts of local equating bias in two or more quintiles, or extremely large bias in one quintile. In the $\bar{b}_D - \bar{b}_E = 2$ conditions (extreme vertical equating), a moderate to high amount of local equating bias in at least two quintiles or a very high amount in one quintile appeared for all of the multidimensional conditions, except the configurations for which there was no data. This was consistent with the correlations, which indicated that the poorest equating occurred in the most extreme vertical equating situations. Configurations 3 and 7 were again exceptions. Configuration 7 showed the largest equating bias for the $\bar{b}_D - \bar{b}_E = 0$ condition and the smallest for the $\bar{b}_D - \bar{b}_E = 1$ condition, which was consistent with the SRMSD, SDM and Local Bias data.

Note that the unidimensional configurations had a small amount of bias in all quintiles for all conditions. Configuration 1 and Group M2 followed a similar pattern of low bias in all quintiles for all conditions, although a few quintiles had moderate bias.

Table 10

Local Equating Bias[a] for Estimated Thetas for Test E versus Test D

Configuration	$\bar{b}_E - \bar{b}_S$	N[b]	Quintile 1	2	3	4	5
1	0	5198	-.09	.02	-.02	-.04	.08
	1	4642	-.22	-.06	-.08	-.07	-.14
	2	c	c	c	c	c	c
2	0	4757	.01	.07	.01	.02	.24
	1	4196	-.50	-.09	-.05	-.04	.12
	2	c	c	c	c	c	c
3	0	4748	.57	.01	.03	.00	.00
	1	4348	.16	-.05	-.08	-.04	-.13
	2	c	c	c	c	c	c
4	0	4552	-.05	.04	.00	.01	.03
	1	3979	-.38	-.05	-.06	-.12	.08
	2	c	c	c	c	c	c
5	0	5064	-.13	-.00	.01	.04	.08
	1	4602	.26	-.08	-.02	.01	.01
	2	c	c	c	c	c	c
6	0	5679	.14	.12	.07	.01	-.00
	1	5374	-.05	-.14	-.10	-.07	.01
	2	4840	-.08	-.20	-.19	-.10	-.08
7	0	5139	.59	.27	.14	-.06	-.21
	1	4824	-.09	-.06	-.05	-.06	-.17
	2	4131	-.16	-.23	-.12	-.15	-.25
8	0	5593	.14	.00	-.02	.01	-.05
	1	5277	.19	-.00	-.00	-.01	.03
	2	5073	-.04	-.11	-.14	-.07	-.07
9	0	5634	.07	.00	-.02	-.03	-.10
	1	5442	.15	-.07	-.06	-.01	-.09
	2	4900	.23	-.16	-.19	-.11	-.17
10	0	5455	.07	.02	-.02	-.04	-.06
	1	5177	-.16	-.02	-.05	-.03	-.01
	2	4864	-.04	-.10	-.08	-.06	-.17
11	0	5493	-.10	.03	.03	.01	.00
	1	5243	-.15	-.12	-.07	-.01	.09
	2	5033	-.10	-.10	-.04	-.03	-.00

[a]Local equating bias is not really bias in the conventional statistical sense, but actually is local consistency of scaled estimated thetas.

[b]Number of simulated examinees for whom thetas were estimated on both tests.

[c]Incomplete data: too many perfect scores in Test 1.

Note: Quintiles range from 1 for the lowest ability to 5 for the highest. A positive difference indicates that the Test 1 mean is higher than the Test 2 mean.

DISCUSSION

The purpose of this research was to examine the effects of various multidimensional data configurations on vertical equating using a three-parameter logistic model. Eleven configurations were chosen, ten two-trait and one unidimensional. For each of these eleven configurations, one horizontal and two vertical equating conditions were simulated. Data were generated using a multidimensional model with the correlation between traits ranging from 0.00 to 1.00.

Simulations. The simulated item parameters and thetas were well within acceptable limits of the desired values. Means, standard deviations and KR-20 values of number-correct scores indicated that all the conditions simulated realistic test configurations even though the most extreme vertical equating condition produced tests that had quite different difficulties.

Parameter estimation. How well the item parameters were estimated appeared to depend to some extent on whether the tests were unidimensional or multidimensional. For Group U (both tests essentially unidimensional) difficulty for Trait 1 (b_1) was estimated very well, and difficulty for Trait 2 (b_2) was estimated extremely poorly. For most of Group M1 (both tests multidimensional), b_1 was not estimated as well as for Group U. Most Group M1 b_2 estimates, although mediocre, were better than Group U estimates. An exception in Group M1 was Configuration 5, which had excellent estimation of both b_1 and b_2.

Note in Table 1 that the correlation between b_1 and b_2 was zero for most configurations. Because the correlation between b_1 and b_2 was high (.80) for Configuration 5, the excellent estimation of both could be expected. (Recall that the same set of base difficulties was used for both traits; therefore, a high correlation between b_1 and b_2 would indicate that most b_2s were the same as the b_1s.) Similarily, the medium correlation between b_1 and $\hat{b}$ for Configuration 9, Test D could be expected because the correlation between b_1 and b_2 was .5. However, for Group U, with zero correlation between b_1 and b_2, and b_1 well estimated, b_2 could not be estimated very well.)

Especially notable is the extremely poor estimation of b_1 for Test D, Group M2 (about half the items measured each trait), and the poor estimation of b_2 for Test E, Group M2 and for Group U. (Note, that when $a_2 = 0$, then $a_2(\theta - b_2) = 0$, and b_2 has no influence on an examinee's response; therefore, discussing the estimates of b_2 has no meaning. Nevertheless, because many of the Group U tests examined here have a few items with nonzero discrimination, the quality of the difficulty estimates was discussed).

For Group M2, good b_1 estimates and poor b_2 estimates for Test E and good b_2 estimates and poor b_1 estimates for Test D followed from the fact that nearly half the items (13 of 30) did not measure Trait 1, 13 others did not measure Trait 2, and only four measured both traits. Therefore, approximately half the items discriminated on Trait 1 and half on Trait 2 which allowed one trait to be well estimated while the other was estimated poorly.

In general, when $r(b_1,b_2)$ was high, the estimation of both b_1 and b_2 was good. When $r(b_1,b_2)$ was medium, b_1 was estimated well, and the estimation of b_2 was mediocre. When $r(b_1,b_2)$ was zero, and Trait 2 was

measured by few or no items, then b was estimated well for the measured trait and poorly for the trait not thoroughly measured.

It appears that how well the difficulty parameter is estimated depends on an interaction between whether or not the test is unidimensional according to the discrimination criterion, and how closely correlated the difficulties for the two traits are. If the items clearly measure one trait and not the other (i.e., the test is unidimensional), the difficulty parameter on the trait measured is estimated well. However, if the test measures both traits (i.e., the test is multidimensional), then mediocre estimation of the difficulty parameter for both traits can be expected unless the correlation between difficulty on both traits is high.

The estimation of the discrimination parameter followed some of the grouping patterns established. For Group U, a_1 was estimated well. For Group M1, the estimation of a_1 was mostly mediocre, and Group M2 had the poorest estimations overall. For Group M3, both a_1 and a_2 were estimated well on Test E (unidimensional) and mediocre on Test D (multidimensional).

The estimation of Trait 2 discrimination (a_2) did not seem to follow a discernable pattern. Most $r(\hat{a},a_2)$ were mediocre to poor. In general, the estimation of a_2 for Group M1 was better than that for Group U. The size of the correlation between a_1 and a_2 appeared to have no effect on the estimation of either a_1 or a_2; however, it did appear to be dependent on whether the tests were unidimensional or not. If the test was unidimensional, then a_1 was well estimated, and a_2 was poorly estimated. If the test was multidimensional, then the estimation of both a_1 and a_2 was mediocre to poor.

The estimation of the c parameter was poor to mediocre for all conditions of all configurations. No consistent

pattern occurred. Lord (1975a) and others have shown that LOGIST parameter estimates for the three-parameter model are adequate if $N \geq 1000$ and $n \geq 50$. If there are no or few examinees at the lower end of the range of abilities, there is no information from which to estimate c; hence, LOGIST estimates a fixed value of c for all such items (Wingersky, 1983). There may have been too few examinees with low test scores for these data. This appears implausible because the number of examinees used here for parameter estimation (4000) is well over the 1000 recommended by Lord. As the difficulty of the items increased there should have been more examinees with low test scores; hence, c should have been better estimated for the harder tests. This also does not appear to be the case here. Possibly the 30-item tests were too small for good estimation of the c parameter.

Correlations between true and estimated trait values followed the dimensional grouping. Group U had the highest correlations between estimated ability and ability on Trait 1. Groups M1 and M2 had correlations about .1 lower. Configuration 1 had the lowest. Notice for Groups M1 and M2 that, in general, the correlation between $\hat{\theta}$ and θ_1 increased as the correlation between θ_1 and θ_2 increased.

When both tests were unidimensional (Group U), θ_1 was well estimated but θ_2 was poorly estimated. Trait 2 estimation improved as the correlation between traits increased. When one or both tests were multidimensional (Groups M1, M2, and M3), then the estimation of both traits was approximately the same. When one or both tests were multidimensional, as the correlation between traits increased, the correlation between $\hat{\theta}$ and θ increased until it was as good as or better than the correlations for the corresponding conditions of the unidimensional criterion (except for the easiest test where the correlation remained less than .60).

For most tests, the correlation between $\hat{\theta}$ and θ decreased as the test got harder. Examination of Table 7 reveals that Group M1 configurations had a correlation under .60 on the easiest test for both traits (.19 to .34 less than the next harder test). All other configurations had virtually equal correlations on the easiest and next to easiest tests. For these five configurations (Group M1), the loss of examinees (1150, 1412, 1336, 1442, 1079, respectively) was large. It appeared that a loss of over 1000 examinees lowered the correlations between estimated thetas and true thetas on both dimensions. All multidimensional tests, (except Configuration 6 tests), lost more examinees to zero or perfect scores than the undimensional tests. This could have serious implications for item and ability parameter estimation.

How well both traits were estimated appeared to depend on 1) how strong the correlation between true ability on the two traits was, 2) whether the two tests were unidimensional or multidimensional, and 3) how many examinees were lost due to zero or perfect scores. The higher the correlation between true ability on the two traits, the better the estimation of both traits was. If one or both tests were multidimensional, then both traits were estimated fairly well; however, if both tests were unidimensional, then one trait was well estimated, and one was estimated poorly. If over 1000 examinees were lost, the traits were poorly estimated, despite a larger N than the criterion set by previous researchers. These results support several studies where the same conclusions were reached (Christoffersson, 1975; Drasgow & Parsons, 1983; McKinley, 1983; Reckase, 1977, 1979; Yen, 1984d).

Equating. In general, the greater the difference in difficulty between the tests to be equated, the worse the equating. There was little to discriminate between the unidimensional criterion and the other configurations (except Group M3) in terms of means, standard deviations, SRMSD, or correlations between θ as estimated by Test E and θ as estimated by Test D. The worst equating was for Group M3, where one test was unidimensional, and one was multidimensional. If both tests were unidimensional (Group U) or both were multidimensional (Groups M1 and M2), equating results were equally good.

The strength of the correlation between true traits, θ_1 and θ_2, also appeared to have no effect, and the ratio of standard deviations showed little pattern. When one test was unidimensional and one multidimensional, the equating was poor, as seen by the low ratios (SD_E/SD_D) in Group M3. The standardized difference between means (SDM) and local bias were nearly all small. As above, the poorest equating appeared for the most extreme vertical equating situations. In contrast, Group M3 (again standing out as an exception) had the poorest SDM and local bias in the horizontal equating condition.

Upon examining local equating bias (local consistency of scaled θs), clearly Group M3 had the poorest equating for the horizontal condition and for the most extreme vertical equating condition. The unidimensional criterion had the smallest overall bias in all quintiles. In general, the multidimensional configurations showed as little bias as the unidimensional criterion except in the lowest quintiles. (The multidimensional equatings were as good as the unidimensional ones except in the lowest quintile for some conditions.) It was not clear why Configurations 2, 3, and 4 had a large amount of bias in the first quintile.

This research supports the conclusion reached by many researchers using real data that the results of equating multidimensional data sets are as acceptable as the results from equating unidimensional data sets. (Cook, Dorans, Eignor & Petersen, 1983; Cook & Eignor, 1983; Cook, Eignor & Petersen, 1982; Cook, Eignor & Taft, 1984; Holmes & Doody-Bogan, 1983; Yen 1980, 1984c).

Cook et al. (1982) found acceptable three-parameter IRT equating results when equating achievement tests in Biology and American History & Social Studies, despite the instability of item parameter estimates.

Cook et al. (1984) indicated that three-parameter IRT equating methods were quite robust to violation of the unidimensionality assumption if the multidimensionality can be represented in a similar factor structure in the two tests to be equated. However, if the multidimensionality represented different factor structures (i.e. Group M3 in this paper), then the IRT equating results were poor. Both of these results are supported by the data presented here.

Cook et al. (1983) found that, despite the multidimensionality found in SAT verbal and Mathematics tests, IRT true-formula-score equating results were reasonable. Holmes and Doody-Bogan (1983) found acceptable results when using three different IRT equating methods to equate two content areas and two grade combinations.

Yen (1984b) shows results indicating that, while equating one-dimensional and two-dimensional tests is problematic, it can result in acceptable equating. The results obtained in this research support this conclusion. Group M3 was a case of equating an essentially one-dimensional test to a two-dimensional test. The results are comparable to Yen's.

Summary and conclusions. It is accepted knowledge that many standardized tests, such as most achievement tests and many aptitude tests, do not satisfy the undimensionality assumption of the three-parameter logistic model (Bejar, 1983; Kingston & Dorans, 1982; Reckase, 1977, 1979). Therefore, the question is not whether the assumption is satisfied but whether a specific use of the model is robust to violations of the assumption (Hambleton & Cook, 1977; Reckase, 1981). Hambleton, Swaminathan, Cook, Eignor and Gifford (1978) and Yen (1984a, 1984b) presented evidence that the models are robust to some departures. The results of this research present more evidence of this robustness.

Although the strength of the correlation between true (generating) traits appeared to have little effect on the quality of equating, there was evidence that the dimensionality of the tests to be equated (as determined by discrimination values on the two dimensions) did have an effect on both item parameter estimation and equating. The poorest parameter estimation and the poorest equating occurred in a situation in which one test was unidimensional and one was multidimensional. This situation resulted in worse equating than if both tests were unidimensional or both were multidimensional. Whether poor equating was due to poor item parameter estimation or test dimensionality was not clear.

In general, the greater the difference between b_1 and b_2, the worse the equating. Horizontal equating was better than vertical, and in most cases, the worst equating was in the most extreme vertical equating condition.

When both tests to be equated were unidimensional (both with large discrimination values on the same dimension), then a_1 and b_1 were well estimated, and a_2 and b_2 were poorly estimated. When both tests to be equated were multidimensional, then b_1 was estimated fairly well, b_2 was poorly estimated, and a_1 and a_2 were mostly poorly

estimated. When both tests were multidimensional -- the easier test having large discrimination values on one dimension, the harder test on the other -- then b_1, b_2, a_1, and a_2 were all poorly estimated. When the easier test was unidimensional and the harder multidimensional, then a_1 and b_1 were well estimated, and a_2 and b_2 were poorly estimated for the easier test, while, for the harder test, b_1 was estimated fairly well, and b_2, a_1, and a_2 were poorly estimated. The estimation of the c parameter was poor for all conditions of all configurations.

When considering local equating bias, there appears to be a very slight advantage when both tests are essentially unidimensional over equating tests that are both multidimensional. This advantage appeared mostly at the extremes of the ability distribution, particularly the lower end. This could be due to poor estimation of the discrimination parameter for both traits, however, the very poor estimates of a_1, b_1, a_2, and b_2 for Group M2 seemed to have little or no effect on the equating, as did the poor estimation of b_2 and a_2 for Configurations 8, 9, and 10.

In conclusion, for the conditions in this research, the results indicated that the three-parameter logistic model is as good, in most instances, for vertical equating of multidimensional data as it is for unidimensional data. Clearly, caution should be exercised when vertically equating a unidimensional test to a multidimensional test. However, both the unidimensional and the multidimensional equatings could be judged as extremely poor instead of good because equating was as good for the multidimensional case (for which the model is not appropriate) as for the unidimensional case (for which the model is appropriate). Due to the absence of an appropriate criterion for judging the adequacy of equating, a determination of whether any equating is good or bad is not possible.

REFERENCES

Angoff, W. H. (1984). Scales, norms and equivalent scores. Princeton, NJ: Educational Testing Service.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton, (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institite of British Columbia.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. Novick, Statistical theories of mental test scores. (pp. 395-549) Reading, Mass.: Addison-Wesley.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Comprehensive Tests of Basic Skills, Forms U & V (1981) Monterey, CA: CTB/McGraw-Hill.

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1983, April). An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

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

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

Cook, L. L., & Eignor, D. R. (1983, April). An investigation of the feasibility of applying item response theory to equate achievement tests. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Cook, L. L., Eignor, D. R., & Petersen, N. S. (1982, March). A study of the temporal stability of IRT item parameter estimates. Paper presented at the annual meeting of the American Educational Research Association, New York.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1984, April). A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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

Doody-Bogan, E., & Yen, W. M. (1983, April). Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement 7, 189-199.

Gialluca, K. A., Crichton, L. I., Vale, C. D. & Ree, M. J. (1984, November). Methods for equating mental tests. (AFHRL-TR-84-35). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory.

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

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

Hambleton, R. K., & Traub, R. E. (1971). Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 24, 273-281.

Hattie, J. (1982). Decision criteria for determining unidimensionality: An empirical study. Armidale, N.S.W., Australia: Center for Behavioral Studies.

Holmes, S. E., & Doody-Bogan, E. N. (1983, April). An empirical study of vertical equating methods using the three-parameter logistic model. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

IMSL Library (7th Ed.). (1979). Houston TX: International Mathematical and Statistical Libraries.

Kingston, N. M. & Dorans, N.J. (1982). Equating of GRE aptitude tests using item response theory: Effects of multidimensionality. Paper presented at the annual meeting of the American Psychological Association.

Kingsbury, G.G., & Weiss, D.J. (1979, April). Relationships among achievement level estimates from three item characteristics curve scoring methods. (Research Report 79-3). Minneapolis: University of Minnesota, Department of Phsychology, Psychometric Methods Program.

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

Lord, F.M. (1968) Analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.

Lord, F. M. (1975a). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (Research Bulletin 75-33). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1975b). A survey of equating methods based on item characteristic curve theory (ETS RB 75-13). Princeton, NJ: Educational Testing Service.

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

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

Marco, G. L., Petersen, N. S. & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. J. Weiss, (Ed.), New Horizons in testing. New York: Academic Press.

McKinley, R. L. (1983, April). A multidimensional extension of the two-parameter logistic latent-trait model. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

McKinley, R.L. & Reckase, M.D. (1980, December). A comparison of the ANCILLES and LOGIST parameter estimation procedures for the three-parameter logistic model using goodness of fit as a criterion. (Research Report 80-2). Columbia: University of Missouri, Department of Educational Psychology.

McKinley, R. L., & Reckase, M. D. (1982, March). Multidimensional latent trait models. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Petersen, N. S., Cook, L. L. & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 137-156.

Reckase, M.D. (1977). Ability estimation and item calibration using the one and three parameter logistic models: A comparative study (Research Report 77-1). Columbia: University of Missouri, Department of Educational Psychology.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Reckase, M. D. (1981a, April). The validity of latent trait models through the analysis of fit and invariance. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Reckase, M. D. (1981b, August). The formation of homogeneous item sets when guessing is a factor in item responses (Research Report 81-5). Columbia, MO: University of Missouri, Educational Psychology Department, Tailored Testing Laboratory.

Ree, M.J. (1979, March). Measurements and latent-trait theory. Paper presented at the 1979 AFHRL Conference on Human Assessment.

Ross, J. (1966). An empirical study of a logistic mental test model. Psychometrika, 31, 325-340.

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

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 15, 23-25.

Slinde, J. A., & Linn, R. L. (1979a). 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, 16, 159-165.

Slinde, J. A., & Linn, R. L. (1979b). The Rasch model, objective measurement, equating and robustness. Applied Psychological Measurement, 3, 437-452.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota.

Warm, T.A. (1978). A primer of item response theory. Oklahoma City, OK: U.S. Coast Guard Institute.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. VAncouver, BC: Educational Research Institute of British Columbia.

Wingersky, M.A., Barton, M.S., & Lord, F.M. (1982) LOGIST user's guide. Princeton: Educational Testing Service.

Wood, R. L., & Lord, F. M. (1976). A User's guide to LOGIST (Research Memorandum RM-76-4). Princeton, NJ: Educational Testing Service.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). LOGIST: A computer program for estimating examinee's ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, NJ: Educational Testing Service.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Statistics, 17, 297-311.

Yen, W. M. (1982, March). Use of three-parameter item response theory in the development of CTBS, Form U, and TCS. Paper presented at the National Council on Measurement in Education, New York.

Yen, W.M. (1983). Tau equivalence and equipercentile equating. Psychometrika, 48, 353-369.

Yen, W. M. (1984a). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W. M. (1984b, June). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. Paper presented at the Psychometric Society, Santa Barbara, CA.

Yen, W. M. (1984c). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. Journal of Educational Statistics, 21, 93-111.

Yen, W. M. (1984d, April). The choice of scale for educational measurement: An IRT perspective. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

APPENDIX

REVIEW OF RELEVANT VERTICAL EQUATING RESEARCH

Review of Relevant Vertical Equating Research

In many testing situations, it is desirable or even necessary to compare test scores of different examinees or the same examinee across multiple forms or across levels of a test, or to compare two entirely different tests. These different forms and levels cannot be expected to be of equal difficulty nor to be equally reliable. Therefore, scores on different forms or levels must be transformed so the desired comparisons between test scores will be meaningful.

Converting the system of units of one test or form to the system of units of the other test or form so the resulting scores are directly equivalent is called test equating (Angoff, 1984). Horizontal equating is equating test forms that measure the same attribute at the same difficulty level. Vertical equating equates tests that measure the same attribute at different difficulty levels. Lord's (1977) definition of equating is that two tests are considered equated if it is a matter of indifference to each examinee which test is taken. Under this definition, tests measuring different traits and raw scores on unequally reliable tests cannot be equated. Hence most equatings must be considered approximate. In particular, vertical equating of raw scores is not possible.

Lord (1980) states three important requirements for equating unidimensional tests measuring the same ability: equity, invariance across groups, and symmetry. Equity implies that Lord's definition for equating is satisfied. (I.e., It would be a matter of indifference to each examinee which test is taken.). The invariance requirement means that the transformed scores on Y must be the same no matter what population is used to determine the transformation. The symmetry requirement is that the equating must be the same no matter which test is labeled X and which is labeled Y.

Lord shows that the equity requirement will not hold if either X or Y is a fallible (not perfectly reliable) measure unless X and Y are strictly parallel tests, in which case the need for equating is obviated. Therefore, the equity requirement implies that scores x and y on two tests cannot be equated unless either both scores x and y are perfectly reliable (infallible) or the two tests are strictly parallel. Since in practice scores are not perfectly reliable, equating of raw scores is not possible. If the tests differ in difficulty, the easy test is more reliable at low abilities, and the difficult test is more reliable at high abilities. Therefore, although the (curvilinear) relationship between true-scores can be determined, no single 'equating' of observed scores can be equitable within all subgroups of examinees (Lord, 1977). However, equating of true scores can be done in such a way as to satisfy all three of the above requirements.

Horizontal equating is appropriate whenever tests are constructed to be interchangeable, with the same degree of difficulty and content. Alternate forms of the same test is the major use for horizontal equating. The equating is needed in order to compare the scores of different examinees taking different forms of the same test. Other situations requiring horizontal equating are comparing performance on tests of different publishers that measure the same attribute at the same difficulty level, and pre-equating (deriving the relationship between test editions before they are operationally administered). (Cowell, 1982; Marco, 1977).

The need for vertical equating arises when scores from tests of unequal difficulty must be compared, such as following an individual's progress through one academic year, following an individual's progress across grades, comparing scores of different examinees taking different levels of the same test, interpreting scores of students who are tested "out of level" because their ability levels are not the same as their grade placement, or comparing performance on tests published by different publishers. Vertical equating is especially applicable to the case of a multilevel battery of tests with levels of increasing difficulty (Yen, 1983). An advantage of such a series of graded tests is that examinees ahead or behind their peers in their development can take the most appropriate level, thereby increasing measurement accuracy. These out-of-level scores must then be scaled in some manner (vertically equated) if they are to serve a useful purpose. Other practical applications include evaluating a multilevel mathematics aptitude test, and pre-equating a test (Marco, 1977).

EQUATING

Any equating procedure, whether horizontal or vertical, requires both a data collection design and a statistical model for analyzing the data.

Data Collection Design

The three most common data collection designs are the single-group design, the equivalent-groups design and the anchor-test design (Lord, 1975).

Single-group design. In the single-group design, the tests to be equated are administered to all examinees. If one entire test is administered before the other, a bias exists due to practice and fatigue. Therefore, some form of counterbalancing is necessary.

Equivalent-groups design. For the equivalent-groups design one test is administered to both groups, where the groups are equivalent random samples of examinees. In this case, an unknown amount of bias exists due to the difference in ability between the groups. If the groups are randomly selected, this bias is approximately inversely proportional to the square root of the number of cases (Lord, 1975). If the groups are not equivalent or if the groups are small, even though they are randomly chosen, the anchor test design must be used. (This design is sometimes used to equate alternate forms of a test. The two randomly selected groups each take a different form of the test. This method of equating is valid only if the two groups are truly equivalent, which is rare in practice.)

Anchor-test design. The anchor-test design is a combination of the first two designs. It involves both the administration to different groups of the different tests that are to be equated, plus the administration of a common set of items to all examinees. This common set of items, the anchor test, measures the difference between the groups and is used to reduce the bias due to differences between groups. Therefore, the different groups need not be equivalent nor random. The anchor test may be either a subset of each of the tests to be equated (an internal anchor), or it may be administered separately for the test to be equated (an external anchor). How effectively the anchor test eliminates bias depends on its correlation with the tests being equated (Lord, 1975).

In each of these three data collection designs there are either common items or common examinees, or both. The single-groups design consists of two different tests taken by a common group of examinees. The equivalent-groups design consists of a common set of items taken by two different groups of examinees. The anchor test design contains a common subset of items taken by two different groups. See Table 1.

Table 1. Data Collection Designs

DATA COLLECTION DESIGN	COMMON ELEMENTS	RELATIONSHIPS OF TESTS TO GROUPS
single-group	All examinees are common. No items are common.	One group takes both Test 1 and Test 2.
equivalent-groups	No examinees are common. All items are common.	Group 1 and Group 2 take Test 1.
anchor-test	No examinees are common.	Group 1 takes Test 1 = ANCHOR + items not in Test 2.
	Some items are common. (The ANCHOR)	Group 2 takes Test 2 = ANCHOR + items not in Test 1.

EQUATING MODELS

The wide variety of test equating techniques that are found in the psychometric literature can be separated various ways. One such distinction is that of the traditional techniques based on classical test theory and the newer techniques based on item response theory (IRT) or latent trait theory. An extensive coverage of the traditional or classical test theory techniques can be found in Angoff (1984). See Table 2. Item response theory techniques are surveyed by Lord (1975). An equating method is a procedure for determining a transformation for converting the scores on one test to the system of units in another test in such a way that it does not matter to the examinee which form she/he takes. There are three general techniques for determining the transformation. The first two are traditional techniques. A description of these techniques as iterated by Angoff (1984) and Lord (1975) follow.

Linear equating involves choosing a transformation that results in scores from two tests being equated if they correspond to the same number of standard deviations from the mean in some group of examinees. Equipercentile equating involves choosing a transformation resulting in scores from the two tests being equated if they correspond to the same percentile rank in a group of examinees. Item response theory equating involves choosing a transformation that results in scores from two tests being equated if they correspond to the same estimated level of ability underlying the two tests.

Table 2. Angoff Equating Designs

I. Random groups - a different test administered to each group (half take Form X, half Form Y).
   A. Equally reliable tests
      1. Linear procedure
      2. Curvilinear analog
   B. Unequally reliable tests

II. Random groups - both tests administered to each group, counterbalanced, half take Form X, then Form Y; half take Form Y, then Form X).
   A. Equally reliable tests
      1. Linear procedure
      2. Curvilinear analog
   B. Unequally reliable tests

III. Random groups - a different test administered to each group, common equating test administered to both groups.
   A. Equally reliable tests
   B. Unequally reliable tests

IV. Nonrandom groups - a different test to each group, common equating test to both groups.
   A. Basic linear method for groups not widely different in ability
   B. Curvilinear method for groups not widely different in ability
   C. Linear methods for samples of different ability
      1. Equally reliable tests
      2. Unequally reliable tests

V. Other methods involving score data.
   A. Forms X and Y equated to a common test
      1. Linear procedure
      2. Curvilinear analog
   B. Forms X and Y predicted by a common test
      1. Linear procedure
      2. Curvilinear analog
   C. Forms X and Y predicting a common test
      1. Linear procedure
      2. Curvilinear analog

VI. Methods of score equating based on item data.
   A. Thurstone's absolute scaling method
   B. Swineford - Fan method of equating

Traditional Equating

Traditional equating techniques are commonly of two types: linear and equipercentile. Before discussing linear and equipercentile equating, it is necessary to define some terms.

Two sets of scores are tau-equivalent if their expected means are equal. Also, their true scores will be equal for every examinee and their true score variances will be equal. Two sets of scores are parallel if their expected means, variances, and corelations with other variables are equal. Two sets of scores are strictly parallel or equivalent if their conditional distributions (of observed scores for a fixed true score) are equal. They are expected to have identical distributions, identical moments, and identical relationships with other variables. Therefore, although tests with unequal difficulty or unequal reliability cannot be equated (will not be equivalent) it is possible to calibrate them (have tau-equivalency).

Linear equating. One definition of equivalent scores is that scores on two test may be considered equivalent if their frequency distributions for a particular group of examinees is identical (Cook, 1982). For horizontal equating, where the difficulty levels of two tests are similar, the shape of the raw score distributions will differ only in the means and standard deviations (first and second moments) when administered to the same group of examinees. The shape of the raw score distributions can be made identical by transforming only the first and second moments of one of the distributions. Such a transformation is linear since the resulting relationship between raw scores on the two tests is linear.

Assuming that the raw scores of two tests are linearly related, a linear transformation is chosen such that scores from the tests are equated if they are the same number of standard deviations from the mean in some group of examinees. The two tests are equated by a linear transformation of the mean and standard deviation of the scores on one test. The assumption that the raw scores are linearly related usually holds approximately for forms that are very carefully constructed to be strictly parallel. Hence the score distributions have similar shapes and the standard scores are equivalent. If the assumption does not hold, the standard scores will not be equivalent and a table or graph may be used to show score transformations (Allen & Yen, 1979). An additional assumption for linear equating is that the tests have equal reliabilities.

Linear equating is used in designs I.A.1 and II.A.1 in Angoff (1984). The linear equating technique may be easily adapted for equating tests of unequal reliability in which case this method becomes design I.B. or II.B in Angoff. If an anchor test is used, this linear equating becomes one of designs III.A, III.B, IV.A, or IV.C.

Equipercentile equating. If the two tests differ in level of difficulty (i.e. a vertical equating situation), the shape of the raw score distributions for the two tests as administered to the same group of examinees will be considerably different. The two distributions can be forced into the same shape by setting equal all scores that have the same percentile ranks in the two distributions. This stretches and/or compresses the raw score scale on one of the tests (i.e., all moments of the distribution are transformed), and the resulting relationship between the raw scores on the two tests is curvilinear. For a non-linear (curvilinear) equating of raw scores, the equipercentile method is the traditional method usually used. Equipercentile equating requires that the tests to be equated be administered to randomly selected equivalent groups. If the groups are equivalent, then a raw score on one test is equivalent to a raw score on a second test when the scores are at the same percentile ranks in their respective distributions. Since equipercentile equating involves setting equal all scores in the two distributions that have the same percentile rank, it will transform all moments of one distribution so that its shape will be identical to the shape of the other distribution. The tests are equated by first obtaining the frequency distributions and percentile ranks of raw scores for both tests. Scores with the same percentile rank are then considered equivalent. Interpolation is used to estimate the percentile ranks of raw scores that do not occur in the observed distributions.

Designs I.A and II.A.2 in Angoff (1984) are equipercentile methods. The equipercentile method produces a linear equating of the scores when the distributions of scores differ only in their means and standard deviations. If the distributions differ in more than means and standard deviations, the result is a non-linear equating. If an anchor test is used, Angoff's design IV.B applies.

Results of curvilinear and linear methods will be identical if the distributions of raw scores differ only in their first and second moments. Linear methods should be viewed as approximations to curvilinear methods (Cook, 1982).

If anchor test data is used, equipercentile and linear methods can be equated either directly or indirectly by frequency estimation (Marco, Petersen & Stewart, 1979). For direct equating, scores on the tests to be equated are equated separately to the anchor test. Scores on the two tests to be equated are assumed to be equivalent if they correspond to the same score on the anchor test.

In indirect equipercentile equating, the frequencies of the tests to be equated are estimated from the combined distribution of scores from all examinees on the anchor test. The resulting frequencies represent the estimated distributions on the tests to be equated for all examinees as if they were a single-group design. The usual equipercentile equating is then applied to the resulting marginal distributions on the tests to be equated. For indirect linear equating, the tests to be equated are equated by applying the usual linear equating technique to the resulting estimated means and standard deviations. This is Angoff's design IV.B.

Item Response Theory

Item response theory (IRT) assumes that an individual's performance on a test is influenced by only one important unobservable characteristic, θ, which is called a (latent) trait or ability. The relationship between performance and ability is assumed to be a mathematical function that relates the probability of a correct response on an item to an examinee's ability (Lord & Novick, 1968).

The IRT three-parameter logistic model assumes that the probability of a correct response to item i by a person with ability level, θ is:

$$P_i(\theta) = c_i + (1 - c_i)/\{1 + \exp[-1.7a_i(\theta - b_i)]\}, \quad (1)$$

where a_i, b_i, and c_i are the discriminating power, difficulty, and lower asymptote or guessing parameters of item i, respectively.

IRT Equating

Lord (1975) states that he knows of no equating method other than IRT methods that can estimate the equipercentile line of relationship between raw scores when the two tests to be equated are not parallel, are given to non-equivalent groups, and every examinee takes an anchor test. The traditional equipercentile equating of two tests to an anchor test is inefficient, biased, and inadequate for groups that differ in ability level.

In IRT equating, two scores from two tests are considered equivalent if they correspond to the same estimated level of ability, $\hat{\theta}$. In other words, two tests are considered equated if the same ability estimate, $\hat{\theta}$, can be expected for any examinee on both tests after equating.

Several techniques exist for equating tests both horizontally or vertically, using an IRT model (Lord, 1975, 1980). In general, IRT equating involves the choice of some transformation that results in the equating of estimated true scores $\hat{\tau}$, estimated true-formula scores $\hat{\xi}$, or ability estimates $\hat{\theta}$. In practice, ability estimates, $\hat{\theta}$, are most commonly equated. Four methods of transforming ability estimates onto a common scale will be discussed here. The first three are linear transformations chosen to transform the mean and standard deviation of one distribution of abilities or difficulties to equal the mean and standard deviation of the other distribution of abilities or difficulties. The fourth, a method developed by Stocking and Lord (1983), is a linear transformation that minimizes the average squared difference between test characteristic curves or minimizes the sum of squared deviations between test characteristic curves.

The basic principle behind equating ability estimates, $\hat{\theta}$, is that if the data fit the IRT model, then an estimate ($\hat{\theta}$) of an examinee's ability can be obtained that is independent of the set of items to which the examinee responds. The ability estimates for each examinee obtained from the two tests will be identical except for sampling error.

Estimating ability and item parameters. Before equating can be done using an IRT model, the estimated item parameters, $\hat{a}$, $\hat{b}$, and $\hat{c}$, and the ability estimates, $\hat{\theta}$, must be calculated from the test response vectors. These estimates are usually obtained using computer programs such as LOGIST (Wood & Lord, 1976; Wood, Wingersky & Lord, 1976). With the LOGIST program, the estimates are computed by a modified maximum likelihood procedure.

After being estimated by the LOGIST program, the item parameters and the ability estimates for each test separately (and in some cases for both tests together) are on a common scale. (i.e. For each test all the item discrimination estimates, $\hat{a}$, are on a common discrimination scale; all the item difficulty estimates, $\hat{b}$, are on a common difficulty scale; all the item guessing estimates, $\hat{c}$, are on a common guessing scale; and all the ability estimates, $\hat{\theta}$, are on a common ability scale.)

Placing item parameters for the two tests on a common scale. The next step in IRT equating is to place the item parameter estimates from the two tests on the same scale. If either the single-group design, or the anchor-test design of data collection is used and the parameters are estimated simultaneously in a single LOGIST run, then the item parameters for both tests will already be on a common scale. Also, if the equivalent-groups design is used and the parameters are estimated in a single LOGIST run, then the item parameters will be on a common scale.

For the single-group design, the two sets of item parameters from the two tests may be estimated in two separate LOGIST runs. For the equivalent-groups design and the anchor-test design, the item parameters are usually estimated in two separate LOGIST runs.

If the item parameters for the two tests to be equated are estimated in two separate LOGIST runs then the item parameters, $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$, for one test and the item parameters, $\hat{a}_j$, $\hat{b}_j$, and $\hat{c}_j$, for the other test must be placed on the same scale (a_i on the same scale with a_j, etc.).

Theoretically, the item characteristic curves are independent of the groups used to derive them. Therefore, except for sampling error, the two sets of item parameter estimates should be identical. Unless item parameters or ability parameters are known in advance, LOGIST estimates the item and ability parameters for the three-parameter logistic model in the same run. An arbitrarily chosen mean of zero and standard deviation of one are chosen for the scale on which the ability estimates are placed. The difficulty and discrimination estimates are adjusted accordingly. Whenever the two groups differ in ability level, the item parameter estimates will not be the same but the item difficulty estimates of the two groups (and $\hat{\theta}$s, which are on the same metric as the difficulties) will be linearly related. This linear relationship is then used to place all parameter estimates on the same scale. These common item parameters are then used to equate the two tests.

Formulas for placing item parameters on the same scale. This method is necessary only if the parameter estimations are not performed in a single LOGIST run. If they are performed in a single LOGIST run, the parameters are automatically on the same scale except for sampling error.

Consider the two tests to be equated to be Test 1 and Test 2. The first step is to estimate the parameters in two LOGIST runs. If the data were collected by the single-group design then both Test 1 and Test 2 would have been administered to the same group. The first LOGIST run will be to estimate the parameters for Test 1 using responses from the group. Run 2 estimates the parameters for Test 2 using responses from the same group. The parameters are then placed on the same scale by Method 1 as follows:

a. Calculate $M_{\hat{\theta}_i}$, $SD_{\hat{\theta}_i}$, $M_{\hat{\theta}_j}$, $SD_{\hat{\theta}_j}$,

where the i and j refer to Test 1 and Test 2 respectively, and/or Run 1 and Run 2. M and SD refer to the mean and standard deviation of ability.

b. If the assumptions of the model hold, the $\hat{\theta}$'s will have the following linear relationship:

$$\hat{\theta}_i = A\hat{\theta}_j + B$$

where $A = SD_{\hat{\theta}_i} / SD_{\hat{\theta}_j}$ and $B = M_{\hat{\theta}_i} - AM_{\hat{\theta}_j}$

c. The estimated parameters for $\hat{a}$, $\hat{b}$, and $\hat{c}$ are adjusted as follows:

$$\hat{b}_j* = A\hat{b}_j + B$$

$$\hat{c}_j* = \hat{c}_j$$

$$\hat{a}_j* = \hat{a}_j/A$$

As a result of these transformations, Test 2 is equated to Test 1.

If the data were collected by the equivalent-groups design, then Groups 1 and 2 both took one test. In the first LOGIST run the parameters for this test are estimated using the responses from Group 1. In Run 2 the parameters for the same test are estimated using the responses from Group 2. The parameters are then placed on the same scale by Method 2 as follows:

a. Calculate $M_{\hat{b}_{i1}}$, $SD_{\hat{b}_{i1}}$, $M_{\hat{b}_{i1}}$, $SD_{\hat{b}_{i1}}$

where $\hat{b}_{i1}$ is the estimate of difficulty of item i in Test 1 as estimated from Group 1 responses; $\hat{b}_{i2}$ is the estimate of item difficulty for the same item in Test 1, as estimated from Group 2 responses; M and SD refer to the mean and standard deviation of the item difficulty parameter.

b. If the assumptions of the model hold, the b's will have the following linear relationship:

$$\hat{b}_{i1} = A\hat{b}_{i2} + B$$

where $A = SD_{\hat{b}_{i1}} / SD_{\hat{b}_{i2}}$ and $B = M_{\hat{b}_{i1}} - AM_{\hat{b}_{i2}}$

c. The estimated parameters, $\hat{a}$, $\hat{c}$ and $\hat{\theta}$ are adjusted as follows:

$$\hat{c}*_{i2} = \hat{c}_{i2}$$

$$\hat{a}_{i2}* = \hat{a}_{i2}/A$$

$$\hat{\theta}_{i2}* = A\hat{\theta}_{i2} + B$$

As a result, the scores for Group 2 on the common test are equated to the scores for Group 1 on the common test.

If the anchor-test design is used for data collection, then Test 1 taken by Group 1 and Test 2 taken by Group 2 contain a common set of items, the anchor. In Run 1, the parameters for Test 1 are estimated using the responses from Group 1. In Run 2, the parameters for Test 2 are estimated using the responses from Group 2. Using Method 2 and the anchor item responses (which are collected

according to the equivalent-groups design), the linear relation-ship between these anchor items is calculated, and this relationship is used to adjust all item parameters and ability estimates to a common scale.

Since the LOGIST program can accept parameter estimates from a previous run, an alternate method for estimating the parameters is available for the anchor-test method. In Run 1, the parameters for Test 1, (including the ANCHOR) are estimated using the responses from group 1. For Run 2, the parameters for Test 2 are estimated using the responses from Group 2 with the exception that the b values for the anchor items are fixed at the values obtained from Group 1. All item parameters and ability estimates will then be on a common scale (the Test 1 scale).

Note that, since parameter estimates from the separate IRT parameter estimate runs differ only in origin and unit of measurement, then the transformation required to place two sets of parameter estimates on the same scale is a simple linear transformation. Typically, the A and B parameters for this linear transformation are chosen such that the mean and standard deviation of the transformed distribution of estimated ability (Method 1) or estimated item difficulty (Method 2) obtained from the second test are equal to the mean and standard deviation of the estimated ability or estimated item difficulty for the first test.

A fourth method of equating ability estimates is the Stocking and Lord (1983) characteristic curve method. The A and B in Method 2 are chosen through an iterative process such that the average squared difference between test characteristic curves is minimized.

RELATED RESEARCH

The Anchor Test Study (Loret,Seder,Bianchini & Vale, 1974) was the first equating study spanning a number of tests and grades. Eight standardized reading tests were equated at grades 4, 5 and 6. Difficulty levels and content varied widely on the tests. Various patterns of common and different levels were equated at all three grades. These diverse tests produced by different publishers would intuitively appear to be nonparallel, in which case, equipercentile equating methods should be inadequate. In another study involving verbal tests, Petersen,Marco, and Stewart (1982) found the linear model to be better than equipercentile when both are appropriate. They found equipercentile equating superior to linear when the samples were dissimilar and the tests differed in difficulty.

Linn (1975) found equipercentile equatings "quite satisfactory for most practical purposes." This opinion was based on an estimated error that was "generally less than one raw score point." However, major exceptions were found for test scores in the "chance" range. Also, when the data were analyzed in subgroups of IQ and race, large differences in equating tables were found in regions where the data were scarce. In particular, the largest difference for the minority groups was found where the sample size was extremely small. Slinde and Linn (1977) reanalyzed the Anchor Test Data using equipercentile equating across grade levels (vertical equating). They found that vertical equating with an anchor test is inadequate although the equipercentile method appeared adequate for tests of approximately equal difficulty. Kolen (1983) compared linear equating versus several equipercentile equating methods at sample sizes of 500 or more per form for equating four forms of each of the four tests of the ACT Assessment Program. The linear method was mostly less adequate than the least adequate equipercentile method. He found the equipercentile smoothing procedure using cubic smoothing splines equating as good as those produced by other smoothing methods.

Studies dealing with applications of the Rasch model to horizontal equating generally report favorable overall results (Beard & Pettie, 1979; Renz & Bashaw, 1977). In each case, the Rasch model equating results were found to agree closely with those from more conventional equating methods.

Beard and Pettie (1979) found Rasch equating with an anchor test to be comparable to linear equating in the analysis of longitudinal trends in student achievement of basic skills. They also found that linear equating and Rasch equating were different near the ends of the test score distribution. This is probably to be expected since scores near the distribution ends would usually be adjusted with linear equating.

Rentz and Bashaw (1977) used Rasch equating on the Anchor Test Data. They found close similarity between the Rasch equatings and the equipercentile equatings. However, in view of the Slinde and Linn (1977) findings that equipercentile methods were inadequate for equating tests of different difficulties, it would appear that if Rasch equatings yield similar results, then Rasch equating for tests of differing difficulties would also be inadequate.

In these cases, where the equating was horizontal instead of vertical, the largest discrepancies were found in the lower tails where there was little data. It would seem that, while equipercentile and Rasch horizontal equating may be "adequate" overall, inadequate equating results are occurring in regions where there is a scarcity of data.

On the other hand, almost all equipercentile and Rasch vertical equating results are questionable. Applications of the Rasch model to vertical equating has been less satisfactory than horizontal equating with the Rasch model. Although Guskey (1981), Guskey and Katims (1978), Gustafsson (1979a, 1979b), Slinde and Linn (1979b), Whitely and Dawis (1974) and Wright (1968) suggest that the Rasch model is acceptable for vertical equating, Divgi (1981), Holmes (1982), Kolen (1979), Loyd and Hoover (1980) and Slinde and Linn (1977, 1978, 1979a) found the Rasch model inadequate for equating nonparallel tests.

Guskey (1981),in an investigation of the utility of the Rasch model for vertical equating, found the Rasch ability estimates as accurate as, and in some cases more accurate than, grade-equivalent estimates throughout the range of test scores. He suggested that the Rasch model is a feasible model for cross-validation of score-equating procedures. Guskey and Katims (1978) found highly consistent and stable results in their comparison of traditional and Rasch model vertical equating results.

Gustafsson (1979a) found a negative correlation between item difficulty and discrimination due to examinee guessing. Gustafson (1979b) also found that the Rasch model was adequate for vertical equating, as long as guessing is not present and there does not exist a correlation between item difficulty and discrimination. He found that when there exists a positive or negative correlation between item difficulty and discrimination, a bias dependent on the sign of the correlation will exist. Also, the Rasch model does not account for guessing. When guessing is present, a negative correlation results between item discrimination and guessing, which contributes to poor equating.

Slinde and Linn (1979b), examined groups not widely separated in ability, and concluded that the Rasch model provided a "reasonably satisfactory" means of vertical equating. Wright (1968) reported highly consistent ability estimates for easy and hard item sets in an investigation of the model item-free person measurement claim. Whitely and Dawis (1974) reported similar results.

In a study of vertical equating using the Rasch model, Divgi (1981) found systematic differences between individual ability estimates based on difficult and easy tests. Sometimes these differences were quite large. He concluded that the Rasch model is unsuitable for vertical equating of multiple-choice tests.

Holmes (1982), found that when applying vertical equating with the Rasch model, there were large differences in ability as estimated for examinees in the lower ability range and even larger differences when the equating results were applied to a new sample (a sample on which the equating etimates were not based). Even when both of the tests that were vertically equated were composed of items specifically selected to fit the Rasch model, these discrepancies were found.

Using the equipercentile method as a criterion, Kolen (1979) looked at the mean absolute difference between the equipercentile and eight other equating methods, including one-, two-, and three-parameter methods. The only general tendency found was that the mean absolute differences (from the equipercentile) were larger for lower scores. Kolen concluded that it was "presumably because few examinees in the sample had very low scores." Marked differences were found between Rasch equating and equipercentile results. In an attempt to vertically equate with the Rasch model, Slinde and Linn (1978) found that while easy and difficult tests are "roughly equivalent statistically, they differ substantially in their precision for some levels of ability." Slinde and Linn (1979b) also found ability differences that were "considerably larger at the extremes of the score distribution."

Loyd and Hoover (1980) reported that ability had a pronounced effect on vertical equating results. They found that vertical equating of the mathematics computation test of the Iowa Tests of Basic Skills is not independent of the ability group used. These equatings were highly dependent on the ability of the group, which supports the Slinde and Linn (1978) findings. Examinees taking an easier level of the test, whose scores were equated to a more difficult level, had higher resulting scores when the equating was based on the higher ability group. Examinees taking a more difficult level of the test, whose scores were equated to an easier level, had higher resulting scores when the equating was based on the lower ability group. In addition, large discrepancies between equipercentile and Rasch model equatings were observed.

While these research studies indicate the inadequacy of the Rasch model for vertical equating in general, they also indicate that the largest discrepancies are usually found in the tails of the ability distribution, particularly in the lower range. These studies further suggest that a possible cause is the scarcity of scores in the regions of greatest discrepancy.

The results of several studies (Dinero & Haertel, 1977; Hambleton & Traub, 1971; Kolen, 1981; Marco, Petersen & Stewart, 1979; and Slinde & Linn, 1978, 1979a, 1979b) imply that the Rasch model may not suffice, especially for low-ability examinees and that a two- or three-parameter model may be required. Slinde and Linn, in their series of studies (1977, 1978, 1979a) found linear,

equipercentile and Rasch equating inadequate for vertical equating, especially when differences between test difficulty and between ability levels of equating samples were greatest. They suggested that the three-parameter model might be most useful for vertical equating.

Although as mentioned above, Slinde and Linn (1979b) found the Rasch model possibly useful for vertical equating, they state as in their earlier studies that two- or three-parameter models would be expected to do even better. Furthermore, they conclude that the results of other studies (Dinero & Haertel, 1977, Hambleton & Traub, 1971; Slinde & Linn, 1978, 1979a) imply that the Rasch model may not suffice for low-ability examinees and that the two- or three-parameter model may be required.

Slinde and Linn (1979a) state that a possible explanation for the poor showing of the Rasch model for vertical equating is lack of data fit due to the effects of guessing. Patience (1981) concluded that it may be a problem with violation of the unidimensionality assumption, while Gustafsson (1979b) also proposed the effects of guessing.

Other factors have been proposed that may account for the unfavorable vertical equating results found with the Rasch model. The groups used in the equating usually differed greatly with respect to ability. The tests being studied differed greatly with respect to difficulty. Model-data fit was not assessed. The lack of unidimensionality was frequently offered as an explanation for poor equating results.

On the other hand, vertical equating results using the three-parameter model have been much more satisfactory.

Comparisons of traditional and IRT three-parameter equatings were made by many researchers. Differing results were found for the different equating schemes. When equipercentile assumptions were greatly violated, equipercentile and IRT equatings differed widely. Lord(1975,1977) found the three-parameter model intuitively appealing and also felt it showed promising empirical outcomes, while the equipercentile technique was inadequate for vertical equating due to difficulties with tests varying widely in difficulty and score distributions of groups with different levels of ability.

Woods and Wiley (1977, 1978) found that for nonparallel tests, the equatings from a modified three-parameter logistic model (fixed c) and "the scaling test" equipercentile method (Hieronymus & Lindquist, 1974) were nearly equivalent. Cowell (1979) found that results from IRT equating with the three-parameter logistic model and traditional methods were nearly identical over most of the score range on the Tests of English as a Foreign Language.

Cook, Dunbar, and Eignor (1981), found that when equating nonparallel tests given to nonrandomly equivalent groups, traditional linear, equipercentile, frequency estimation equipercentile and IRT equating using the three-parameter model all yielded comparable results for most of the score reporting range. Where the methods did not coincide (at the upper end of the distribution), the three-parameter method was assumed to produce the most appropriate equating. Cook et al. also found that it is feasible to equate using the three-parameter method on nonparallel tests given to nonrandomly equivalent groups in the absence of an anchor test.

Cowell (1981), in a comparison of a linear model, a one-parameter model and two three-parameter models, found all four models yielded similar results. Since the three-parameter model was the best model for equating with small samples and the cost of item calibration was found to be reduced more by reduction in sample size than by reduction in the number of parameters to be estimated, Cowell concluded that using the three-parameter model on reduced sample sizes would be preferable to using his simplified model at either level of sample size (large or small).

Marco, Petersen and Stewart (1979) compared several combinations of traditional and IRT methods on equating the verbal portion of the Scholastic Aptitude Test. They examined five equating models: direct linear, indirect (frequency estimation) linear, direct equipercentile, indirect (frequency estimation) equipercentile and IRT (one-and three-parameter) models; and two types of samples: random samples, and samples dissimilar in ability. The IRT methods were clearly superior for equating tests of differing difficulty. Marco et al. found, as did Slinde and Linn (1978) that the one-parameter IRT model yielded poorer results when total tests differed considerably in difficulty (i.e. vertical equating). In this situation the linear models and the frequency estimation methods were also inadequate. However, Marco et al. noted that their conclusions are tenative since the criterion used for judging the superiority of methods may have been biased against certain methods when used to equate tests of different difficulty. Nevertheless, in spite of this possible criterion bias, they suggest that the three-parameter logistic model would result in the best equatings under extreme or unusual conditions.

Kolen (1981), used a cross-validation statistic to compare the stability of equating results obtained from seven IRT models and traditional linear and equipercentile methods. The results lend support to the Slinde and Linn findings, and Kolen reached similar conclusions: When guessing was present, Rasch vertical equating yielded inadequate results, while three-parameter equating was the most stable of the nine equating methods used.

Kolen and Whitney (1982) found that linear, equipercentile, and one-parameter model equatings produced similar cross-validation results when equating twelve forms of each of the five (achievement) tests of General Education Development with sample sizes of approximately 200 and when both forms of similar difficulty had been administered to the same examinees. In this situation, the three-parameter equating method produced unstable results. Perhaps due to equating sample sizes being too small and possible score scale condensing (from non-existence of estimated true-scores below the "pseudo-chance" level of the test).

Patience (1981) evaluated four equating methods in terms of their ability to obtain an equated score scale which matched their overall score scale obtained from examinees taking all items. The equipercentile method resulted in the highest correlations of equated scales to overall scores. The one- and two-parameter IRT equating methods were comparable to the equipercentile method. The three-parameter equating resulted in the lowest correlation. Patience states that the poorer IRT results may have been due to several circumstances: There were only six overlapping items; they were not chosen with respect to quality nor model-fit; the item calibration sample size may have been too small; the tests may not have been sufficiently unidimensional; ability estimates were based on subgroups of only 25 items.

Gialluca, Crichton, Vale, and Ree (1984) compared equipercentile, linear, IRT and strong true score theory (STST) using single-group, equivalent-groups and anchor test designs. They found that linear and equipercentile methods were best for horizontal equating. Only IRT and STST methods were suitable for vertical equating. Sample sizes of 1000 were found to be sufficient. Although data collection design seemed to have little effect for samples of equivalent ability levels, the anchor-test design was the only design found appropriate for equating using nonequivalent groups.

Using scale drift as a criterion, Petersen, Cook and Stocking (1983) evaluated the relative accuracy of four traditional methods (three linear and one equipercentile) versus IRT equating with the three-parameter model on the Scholastic Aptitude Test (SAT). Using a chain so that each form was linked to the preceding and following forms by an anchor test, they investigated the effect of equating a test to itself through five intervening forms. It was concluded that the three IRT linking methods used (concurrent calibration, fixed b's and equated b's) were superior to the traditional methods on the verbal portion of the test. On the mathematics portion neither the results from the IRT fixed b's nor the equated b's was as accurate as the equating from two of the linear methods. However, the IRT three-parameter concurrent method was comparable. The equipercentile equating was unsatisfactory for both verbal and mathematics tests. Since in this study, equating was done for tests with similar levels of difficulty and administered to groups of examinees with similar levels of ability, the traditional methods would be expected to yield accurate results. Hence, they conclude that equipercentile equating is less satisfactory than linear equating. Their final conclusion was that of the seven models used in this study, the IRT concurrent method was the best procedure for reducing scale drift over time on the SAT.

In a study assessing the feasibility of item response theory for pre-equating the Test of Standard Written English, Bejar and Wingersky (1981), found the three-parameter model adequate and useful. Lord (1982) equated the verbal score on a 90-item form of the SAT to an 85-item form. In a comparison of the standard errors of IRT, linear, and equipercentile equating he found the following: IRT standard errors were greater than linear standard errors with the exception of the upper ability levels; and IRT standard errors were less than equipercentile with the exception of very low ability levels. Research on vertical equating, then, suggests that the three-parameter model may be the best model available for vertical equating. Although research in different situations has resulted in different conclusions, the usefulness of the three-parameter model for vertical equating has been amply demonstrated.

Despite the proven usefulness of IRT for vertical equating and the agreement by many researchers that the three-parameter model may be the best IRT model for vertical equating, vertical equating with the three-parameter model is not free of problems, as noted in some of the studies above.

Vertical equating with the three-parameter model has also shown large discrepancies in the tails of the distribution as did the Rasch model vertical equating. Also, the largest discrepancies were usually in the lower ability range for three-parameter model vertical equating as was true for Rasch vertical equating (Cook, Dunbar, & Eignor, 1981; Cook and Eignor, 1981, Cowell, 1979, 1981, Modu, 1982; Marco, 1977).

Cook, Dunbar, and Eignor (1981) found when equating nonparallel tests given to nonrandomly equivalent groups, that traditional linear, equipercentile, frequency estimation equipercentile, and IRT equating using the three-parameter model all yielded comparable results for most of the score reporting range. The three-parameter method was assumed to produce the most appropriate equating since a curvilinear relationship would be expected between tests with different difficulties. This curvilinear relationship theoretically negates the possibility of raw score equating (such as traditional linear and equipercentile) and requires the use of IRT equating (Lord, 1980). Cook et al. submit that the discrepancies at the extremes of the distribution between IRT and both curvilinear and linear traditional methods are probably due to instability in the traditional methods from scarcity of data at the extremes. They further state that the IRT true formula score methods that they used are not affected by lack of data at the upper end of the distributon but are affected by lack of data for below chance score levels.

However, as discussed earlier, inadequate data in either extreme of the distribution can result in poor estimation of any or all of the discrimination, difficulty, and pseudo-guessing parameters. This in turn affects estimation of the ability parameters which then can be expected to affect the quality of the equating.

Cook and Eignor (1981) also found larger discrepancies at the extremes and especially in the lower end of the score range in a vertical equating study using the three parameter model. The most general pattern they found was that the IRT equating produced slightly lower scaled scores than either traditional linear or equipercentile methods at portions of the score range which contained the majority of observed scores (i.e. the middle portion). At the upper and lower ends of the score range, the IRT equating produced higher scaled scores, and the discrepancy was usually larger than that for the middle portion of the score range. They submit that the IRT equatings are affected by poorly fitting items, especially at lower ability levels, related to inadequate estimation of the pseudoguessing parameter and by problems in estimating the pseudo-guessing parameter due to insufficient data at the lower ability levels. They present as a possible cause inadequate estimation of discrimination and difficulty for very easy or very difficult items due to scarcity of data.

Cowell (1979, 1981) also found larger differences between methods in the lower extreme. Cowell (1979) found that results for IRT equating with the three-parameter logistic model and traditional methods were nearly identical over most of the score range on the Tests of English as a Foreign Language. The largest differences were found in the "chance score range". Cowell (1981), in a comparison of a linear model, a one-parameter model and two three-parameter models, found all four models yielded similar results. Again, the largest differences were found in the lower half of the range. On the average, the differences were less than one point on a converted score scale while differences greater than two points were found at the extremes, mostly below chance level.

Modu (1982), while investigating the extent to which IRT equating results parallel those of traditional equating methods under conditions that violate the unidimensionality assumption, found agreement between all methods of one point or less, except at the two extremes of the scale score. Scarcity of data was the reason offered for the discrepancies at the extremes.

Marco (1977) found that when equating the scores on 45 items of the Advanced Placement Program to the scores on 50 items of the College-Level Examination Program, linear equating and the three-parameter logistic model equating were in close agreement except at the ends of the scales. In a preequating study, Marco found similar results for equipercentile, linear and three-parameter model equating. The largest discrepancies were again found at the ends of the scale. The maximum discrepancies between scaled scores for the traditional models and the three-parameter model ranged from -15.1 for the center of the scaled score range to -23.8 for the extremes of the range. The difference between the three-parameter equating and the pre-equating discrepancies was even greater. The maximum discrepancy in the center of the range was 7.0 while the maximum discrepancy at the extremes was 21.3. The maximum difference in scaled scores between these two IRT equating methods was more than three times as large in the extremes of the range as in the center.

As was noted earlier, if guessing is present, the Rasch model is entirely inadequate for vertical equating, while the three-parameter model takes guessing into account. However, problems do exist for vertical equating with the three-parameter model when the c or pseudo-guessing parameter is poorly estimated.

In the research mentioned previously, Kolen (1981) suggests that poor estimation of the lower asymptote (pseudo-guessing) parameter may have caused some instability in the observed results. Verbal and quantitative thinking tests were administered at two levels: Level I -- grades 9 and 10, and Level II -- grades 11 and 12. Levels I and II were equated to tests of vocabulary and thinking that were administered as one level to grades 9 through 12. The most stable cross-validation results were produced by the three-parameter estimated observed-score method at Level I with only moderate accuracy at Level II. The three-parameter estimated true-score equating method was the

most stable method at Level II with only moderate stability at Level I. Kolen hypothesized that these results could be partly due to difficulty in estimating the "c" parameter.

In a study of the effects of local item dependence on equating with the three-parameter model, Yen (1984) found unsystematic errors of equating when equating multidimensional tests consisting of collections of different dimensions. Systematic errors of equating were a problem when the two tests measured quite different dimensions.

Doody-Bogan and Yen (1983) found that both horizontal and vertical equatings were as good for multidimensional data as for unidimensional when examining correlations and standardized root mean squared differences. The multidimensional tests usually had worse equatings than the undimensional tests, especially for vertical equatings, when means were examined using the standardized difference between means and local bias statistics.

On the other hand, Kingston and Dorans (1982) found little difference in mean scores for four methods of IRT equating and linear and equipercentile equating, when equating GRE aptitude tests. However, a major exception was poor results for the linear equating model at the upper end of the score scale. No conclusive results were found in relation to the effect of multidimensionality.

Research in vertical equating with the three-parameter model, then, points at possible problems in the equating due to multidimensionality in the data.

A slightly different attack on the problem of vertical equating has been examined by Yen (1982). If test scores are not perfectly reliable they cannot be strictly equated unless they are strictly parallel (Lord, 1980). Since vertical equating by definition implies that the tests do not have the same difficulty, then strictly parallel or parallel results cannot be achieved from vertical equating. Hence, the scores are not interchangeable. However, it is possible that tests with unequal difficulties or standard errors can have tau-equivalent scores.

Yen (1982) demonstrates theoretically, with simulated data, that performing an equipercentile equating of observed scores in an attempt to equate or calibrate tests of unequal difficulty or reliability may not be reasonable since not only score equivalence is lost but tau-equivalence may also be lost. For this reason, equipercentile equating results should not be used as the criterion for judging other procedures.

IRT tau-equivalent "equating" (calibration) is possible and the quality of tau-equivalent calibrations may be examined by verifying whether the calibrated scores have the same mean for various groups of examinees although not necessarily the same variances or the same distributions. Such tau-equivalent "equating" is useful since, on the average, examinees could expect to receive the same score from two tau-equivalent tests and groups means can be compared since their means are expected to be the same for both tests.

ISSUES IN EQUATING

It is quite obvious from the preceding review that solutions to some of the problems involved in equating have been found. However, many issues are not yet resolved or research into their solutions is inconclusive. Some of these problems are how to judge the adequacy of equating, issues involving the samples behind the scores x and y, issues involving the total tests X and Y that are being equated, the anchor test taken alone and/or in relation to the total tests X and Y, issues related to IRT equating only, and miscellaneous issues involving both IRT equating only, and miscellaneous issues involving both IRT equating and traditional equating.

One major problem that remains in the area of equating is how to judge the adequacy of an equating. Appropriate methods for measuring the adequacy of an equating and the choice of an appropriate criterion have not been established.

How can the stability or reliability and the precision or validity of an equating be measured? Several methods have been used and/or proposed. Kolen and Whitney (1982) proposed both an equating bias index and an equating imprecision index. Kolen et al. also used Kolen's (1981) percentile comparison index, which is a measure of the disimilarity between distributions of anchor form scores and converted scores on an equating form.

Divgi (1981) defines the amount of equating bias as the mean difference between expected values on two tests and hence equating bias depends on ability. A problem with this method is that it requires a very large sample in order to result in a reasonably small standard error of the mean and a large number of groups in order that the variation of ability within groups may be as small as possible.

A discrepancy index that was presented by Marco, Petersen and Stewart (1979) was used by Cook, Dunbar and Eignor (1981), and by Petersen, Cook and Stocking (1983). The discrepancy index is a weighted mean square difference between an estimated score and a criterion score, which gives the greatest weight to the values that are most likely to occur.

Angoff (1984) discusses the standard error of equating that was described by Lord (1950) as the standard deviation of converted scores. Slinde and Linn (1977) compared grade equivalent scores in terms of the estimated errors of equating. Slinde and Linn (1978, 1979a, 1979b) used Wright's (1968) "standardized difference scores" comparisons between "easy" and "hard" tests. Standardized difference scores were also used by Holmes (1982) and by Whitely and Dawis (1974). Average standard errors were used by Woods and Wiley (1977, 1978). Doody-Bogan and Yen (1983) used standardized root mean squared differences, standardized differences between means, the ratio of standard deviations, correlations, and tau-equivalence. Gialluca, Crichton, Vale, and Ree (1984) used bias and root mean squared error.

Cowell (1981) used standard deviations of converted scores for pairs of equating methods to compare equating results. Summary statistics of five differences were analyzed. The first one, the maximum absolute difference, shows the largest absolute difference between score conversion tables. Their second summary statistic is the mean absolute difference (weighted). Their third summary statistic, the mean squared difference (weighted) is further subdivided into fourth and fifth summary statistics. These two components are the mean difference squared (weighted) which is a measure of the bias in the differences and the variance of the differences (weighted).

Many researchers have examined the value of equatings by simply looking at and/or comparing equating tables from two or more equating methods along with plots or graphs of the equating values. Beard and Pettie (1979) compared Rasch equating to linear equating by looking at the differences between equated values obtained by the two methods. Cook, Dunbar and Eignor (1981) used graphical comparisons to give an overview of the relative agreement of traditional methods of equating with IRT methods. Cowell (1979) used graphs of scaled scores and score conversion table comparisons. Cowell (1981) used graphs along with the standard deviations of converted scores.

Guskey (1981) used plots to compare grade equivalency estimates of raw scores versus Rasch ability estimates. Kolen and Whitney (1982) examined equating curves, item parameter estimates for extremes and compared equating sample means and variances. Lord (1975, 1977, 1980) used plots to evaluate equating methods. Loyd and Hoover (1980) used plots, evaluation of the means and standard deviations of the item difficulty estimates, maximum difference between two equatings and mean absolute difference between raw scores or between equatings. Marco (1977) looked at plots and the means of groups used for equating. Patience (1981) used plots and correlations between total tests and subsets of tests. Slinde and Linn (1978,1979a) used plots of estimated log abilities on high versus low groups along with standard difference scores and mean differences in log ability estimates.

Other methods used to evaluate equating results include examination of the differences between estimates of the mean of ability resulting from different equating methods (Gustafsson, 1979b), comparison of which equating method consistently produces a closer correspondence between the common scale score distribution for two forms taken by two different groups (Kolen, 1979), comparing various traditional and IRT equating methods by examining scale drift in a chain of six equatings leading back to the original test, (Petersen, Cook & Stocking, 1983), and a set of five evaluative methods proposed by Jaeger (1981) to be used to allow a choice between linear and equipercentile equating methods: similarity of the cumulative score distributions, shape of the raw-score to scaled-score transformations, consistency of the equating results, similarity of item difficulty distributions and similarity of the item discrimination distributions.

It appears then that the number of methods for judging the adequacy of an equating is almost as large and diverse as the number of researchers doing research on equating. None of the methods however stand out as being intuitively superior for the task.

Another problem related to choice of an equating method is how to choose an appropriate criterion. Arbitrary choice of an equating method as criterion seems to be the rule. Marco, Petersen, and Stewart (1979) judged the results of equating a test to itself by using the test itself as a criterion. Cook, Dunbar and Eignor (1981) used an IRT criterion. When a test was equated to a different test they used two "criterion equatings" in which they equated the scores using all their examinees in a single sample, a situation to which they referred as being an "ideal equating situation." The two "criterion equatings" were established by two methods of equating: an equipercentile equating of observed scores and an equipercentile equating of estimated true scores derived from the three-parameter logistic model. They state that the IRT criterion could be biased in favor of the IRT equating methods and the criterion based on traditional equipercentile equating could be biased in favor of the equipercentile equating method.

Patience (1981), when comparing the relative efficiency of vertical equating methods, used plots and correlations obtained from the pairing of equating score scales (based on only part of the items) with "criterion" score scales, where the "criterion" was the scores of examinees on the scale when all items were taken. Since Petersen, Cook and Stocking (1981) equated a base test back to itself through a chain of equatings, their criterion was the base test itself.

Except in the case of equating a test to itself, in which case the criterion is the test itself, there appears to be no basis for these arbitrary choices of an equating result as a standard for comparison of other equating methods. Lord (1977) states that "the results obtained by conventional methods cannot be justified as criterion" (p.132) especially in the situation where the two tests to be equated are not parallel, are given to non-equivalent groups and everyone takes an anchor test, in which case traditional methods are not applicable. This is exactly the case in vertical equating.

Yen (1982), summarizing the inadequacy of the equipercentile method as a criterion, states that "It is not necessarily reasonable to attempt to equate or calibrate tests of unequal difficulty or reliability by performing an equipercentile equating of observed scores, and the result of such an equipercentile equating cannot be assumed to be the criterion against which other procedures are to be judged." (p. 9).

Various methods of cross-validation have been used when evaluating equating methods. Slinde and Linn (1978) divided their samples into high, middle and low ability groups. Parameter estimates from the two extreme groups were applied to the middle group. Kolen (1981) used a cross-validation group to establish a criterion for comparing the results of two traditional and seven IRT equating methods. Their main evaluative technique was to estimate the stability of the results when applied to a new, independent sample. The cross-validation criterion was the mean-squared difference over examinees in the cross-validation distribution between scores on one test and equated scores with identical percentile ranks in randomly equivalent cross-validation distributions. Kolen and Whitney (1982) also used cross-validation analyses.

The choice of a criterion or criteria to be used to judge the adequacy of an equating remains a problem that has not been solved. The need for an appropriate criterion is obvious. Questions remain. How can the stability/reliability and the precision/validity of an equating be measured? What criterion or criteria should be used to judge the adequacy of an equating? How can the "best" or "most correct" equating procedure for a specific equating situation be chosen?

Angoff (1963) states four sources of error applicable to equating situations: the unreliability of the measuring instruments, the design of equating and method of treating the data, the choice of samples to estimate the conversion line, and error associated with the size of the study sample.

The first source of error occurs in all testing situations and results in no unique problems for equating. The second source of error is a crucial one in judging the adequacy of equatings and/or in making a choice among methods since equating methods are not equally reliable. None of the evaluative methods used to date seem to give a definitive answer to the problem of judging the adequacy of equating and/or making a logical choice among the available methods.

A second area of major concern relating to all equating techniques is that of the samples chosen for specific equating situations. The larger the sample, the smaller the error. What is the effect of sample size on equating results? How large must the samples be for adequate equating? Ree and Jensen (1983) found that both estimating item parameters and equating require sample sizes of 1000 to 2000. What is the effect of too small a sample? Different samples yield conversion lines that differ randomly from one another in traditional techniques. Theoretically, IRT equating is independent of the ability distributions of the group taking the test. However, in practice, it has been seen that differences do occur (Beard & Pettie, 1979; Cook, Dunbar & Eignor, 1981; Holmes, 1982; Loyd & Hoover, 1980; Marco, Petersen, & Stewart, 1979; Slinde & Linn, 1977, 1978, 1979). Therefore, the effects of using different samples must be taken into consideration. What is the effect of using a random sample of similar ability versus a random sample of dissimilar ability? In particular, how stable are the equating results obtained from one sample of examinees when applied to a different sample? These and other questions about the samples used when equating have not been fully answered.

A third important area relating to equating and especially of importance in vertical equating is the composition of the tests to be equated. How diverse in difficulty can the tests to be equated be? What is the effect of one test being much more difficult than the other? What is the effect if the overall or average difficulty is the same, while the range of difficulties on one test is very narrow and on the other test is very wide? What is the effect of overlap of item difficulties and what percent of overlap is allowable? The same questions can be asked about the discrimination and (pseudo-) guessing parameters.

If an anchor test design is used, the composition of the anchor test and its relationship to the tests to be equated is an area of concern. How many anchor items are needed for adequate equating? What is the effect of too small an anchor? What range of difficulty (discrimination, guessing) is allowable for the anchor? What distribution of difficulties is most effective? If anchor items are plotted according to their difficulties from respective calibrations, outliers from a line with unit slope are known to decrease the stability of the equating constant. How many outliers can be tolerated and how discrepant from a unit slope can they be without detrimental effects?

The tests to be equated and the anchor test can be similar or dissimilar in content (or difficulty). If the tests to be equated are similar, the anchor can be similar to both or dissimilar to both. If the tests to be equated are disimilar, the anchor can be similar to one or the other or dissimilar to both. How similar to the tests to be equated must the anchor be? What is the effect of the anchor being dissimilar to the tests? How much can the anchor test depart in content from the tests to be equated? How diverse can the anchor difficulty distribution be from the tests to be equated? The individual effects of content and difficulty between anchor and total tests need to be separated. What is the effect of an anchor of similar content but dissimilar difficulty? On anchor tests with similar content and difficulty to total tests, an internal anchor has less total error than an external anchor (Marco, Petersen & Stewart, 1979). Is the same true for equating tests of different content and/or different difficulty? Is there bias in favor of an internal anchor?

Most of the questions raised so far are applicable to all equating methods. A fifth area of concern relates specifically to IRT equating. The list can be broken down into problems related to unidimensionality, model-data fit, calibration/equating samples, the introduction of new groups of examinees, items, parameters and scores. The unidimensionality question is: How robust is (three-parameter) equating to violations of the unidimensionality assumption? The question about model-data fit is: How poor can model-data fit be before the equating becomes unacceptable? What is the effect if model-data fit is too poor? The calibration/equating samples question is: How diverse can the sample for calibrating the item parameters be from the equating samples before the equating becomes unacceptable? What is the effect of using equating results when these samples are diverse? Lord (1975, pp 21-24) describes four methods for dealing with new groups of examinees. How adequate are these methods? IRT equatings are based on predictions of what happens if there are no omitted items. In relation to items, what is the effect of omitted items and how many can be tolerated? In relation to item parameters, what is the effect of inadequate estimation of the item parameters, specifically the lower asymptote parameter?

Many problems relating to IRT scores remain unanswered: how to handle missing, zero and perfect scores, since ability cannot be estimated for zero or perfect scores; how to deal with observed scores below the lowest possible number right true score. What is the effect of condensing the score scale in the three-parameter estimated true-score equivalent method? Since true formula scores below the chance score level are undefined for the three-parameter model, a method is needed to obtain the relation between scores below chance level on the forms to be equated (Lord, 1980, p. 210; Petersen, Cook & Stocking, 1981, p. 16). A sixth area of problems concerning equating contains miscellaneous problems that do not seem to fit the other categories. How many items are required to perform an adequate equating? What is the effect of too few items? Once equating is done, how much linking error can be tolerated? Equating errors are worse at the extremes of the distribution. How little is too little at the extremes?

CONCLUSIONS

Clearly, further research is needed to clarify the remaining issues. However, many of these issues can be avoided or minimized by use of appropriate designs and/or sufficiently large sample sizes. The choice of a criterion for judging the adequacy of an equating is not clear. However, the combination of several of the indices currently in use along with appropriate cross validation techniques should be sufficient to detect major problems in the equatings. Sample sizes large enough to ensure good estimation of all three item parameters should also minimize the problems of discrepancies in the tails of the distributions. Equating tests with extreme differences in difficulty should be avoided, as should outlier anchor items (anchor items that deviate excessively from a line with unit slope through a plot of the anchor item difficulties). The anchor test should be as similar as possible to the two tests to be equated in terms of content, should contain items with good discriminations, and should contain a range of difficulties.

As pointed out earlier, research has mainly indicated that equating with the three-parameter model is reasonably robust to violations of the unidimensionality assumption. Problems concerning omitted items and zero or perfect scores are handled by the LOGIST program. If the tests to be equated and the anchor tests are carefully constructed with the above precautions in mind and adequate sample sizes are used, obtaining good equatings should not be a problem. Nevertheless, multidimensionality and other problems affecting equating exist and while it is hypothesized that vertical equating is robust to some violations of the assumptions, examination of specific violations is still needed.

REFERENCES

Allen, M. J. & Yen, W. M. (1979). Introduction to measurement theory. Monterey, Calif.: Brooks-Cole Publishing Company.

Angoff, W. H. (1963). Can useful general-purpose equivalency tables be prepared for different college admission tests? Personnel and Guidance Journal, 41, 792-797.

Angoff, W. H. (1984). Scales, norms and equivalent scores. Princeton, NJ: Educational Testing Service.

Beard, J. G., & pettie, A. L. (1979). A comparison of linear and Rasch equating methods results for basic skills assessment tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Bejar I. I., & Wingersky, M.S. (1981). An application of item response theory to equating the Test of Standard Written English. (College Board Report No. 81-8). New York: College Entrance Examination Board.

Cook, L. L. (1982, March). The application of item response theory to test equating. Paper presented at the American Educational Research Association IRT Pression, New York.

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

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

Cowell, W. R. (1979). ICC pre-equating in the TOEFL testing program. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, San Francisco.

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

Cowell, W. R. (1982). Item-response-theory pre-equating in the TOEFL testing program. In P. W. Holland & D. B. Rubin (Eds.), Test equating. Princeton, NJ: Academic Press.

Dinero, T. E. & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. Applied Psychological Measurement, 1, 581-592.

Divgi, D. R. (1981). Model-free evaluation of equating and scaling. Applied Psychological Measurement, 5, 203-208.

Doody-Bogan, E., & Yen, W. M. (1983, April). Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Gialluca, K. A., Crichton, L. I., Vale, C. D., & Ree, M. J. (1984, November). Methods for equating mental tests (AFHRL-TR-84-35). Brooks Air Force Base, Texas: Air Force Systems Command, Air Force Human Resources Laboratory.

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

Guskey, T. R. & Katims, M. (1978, March). Use of the Rasch model in calibrating the reading comprehension test of the Iowa Tests of Basic Skills. Chicago: Department of Research and Evaluation, Chicago Board of Education.

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

Gustafsson, J. E. (1979b). Testing and obtaining fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hambleton, R. K., & Traub, R. E. (1971). Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 24, 273-281.

Hieronymus, A. N. & Linquist, E. F. (1974). Manual for administrators, supervisors and counselors. Forms 5 & 6. Iowa Tests of Basic Skills. Boston: Houghton-Mifflin.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. Journal of Educational Measurement, 19, 139-147.

Jaeger, R. M. (1981). Some exploratory indices for selection of a test equating method. Journal of Educational Measurement, 18, 23-38.

Kingston, N. M. & Dorans, N.J. (1982). Equating of GRE aptitude tests using item response theory: Effects of multidimensionality. Paper presented at the annual meeting of the American Psychological Association.

Kolen, M. J. (1979). Comparison of equipercentile, linear and selected latent trait methods for equating forms and levels of the seventh edition of the Iowa Tests of Educational Development. Unpublished doctoral dissertation, University of Iowa.

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

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of General Educational Development. Journal of Educational Measurement, 19, 279-293.

Linn, R. L. (1975). Review of Anchor Test Study: The long and short of it. Journal of Educational Measurement, 12, 201-214.

Lord, F. M. (1950). Notes on comparable scales for test scores (RB-48). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1975). A survey of equating methods based on item characteristic curve theory (ETS RB 75-13). Princeton, NJ: Educational Testing Service.

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

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1982). Standard error of an equating by item response theory. Applied Psychological Measurement, 6, 463-472.

Lord, F., M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. (1974). Anchor test study: Equivalence and norms tables for selected reading achievement tests (Grades 4,5,6). Washington, DC: U.S. Government Printing Office.

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

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

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. J. Weiss, (Ed.), New horizons in testing. New York: Academic Press.

Modu, C. C. (1982, March). The robustness of latent trait model for achievement test score equating. Paper presented at annual meeting of the American Educational Research Association, New York.

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

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 137-156.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P.W. Holland & D.B. Rubin, (Eds.), Test equating. New York: Academic Press.

Rasch, G. (1966). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Ree, M. J. & Jensen, H. E. (1983). Effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss (Ed.), New horizons in testing. New York: Academic Press.

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

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

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 15, 23-25.

Slinde, J. A., & Linn, R. L. (1979a). 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, 16, 159-165.

Slinde, J. A., & Linn, R. L. (1979b). The Rasch model, objective measurement, equating and robustness. Applied Psychological Measurement, 3, 437-452.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

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

Wood, R. L., & Lord, F. M. (1976). A User's guide to LOGIST (Research Memorandum RM-76-4). Princeton, NJ: Educational Testing Service.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). LOGIST: A computer program for estimating examinee's ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, NJ: Educational Testing Service.

Woods, E. M., & Wiley, D. E. (1977). An application of item characteristic curve equating to single-form tests. Paper presented at the annual meeting of the Psychometric Society, Chapel Hill, SC.

Woods, E. M., & Wiley, D. E. (1978). An application of item characteristic curve equating to item sampling packages on multi-form tests. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Yen, W. M. (1982, March). Obtaining some degree of correspondence between unequitable scores: A comparison of item response theory and equipercentile equating methods. Paper presented at the annual meeting of the American Educational Research Association, New York.

Yen, W. M. (1983). Use of the three-parameter logistic model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), Applications of item theory. Vancouver, British Columbia: Educational Research Institite of British Columbia.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.