APPLIED PSYCHOLOGICAL MEASUREMENT

APPLIED PSYCHOLOGICAL MEASUREMENT

Volume 25, Number 4 December 2001

APPLIED PSYCHOLOGICAL MEASUREMENT publishes empirical research on the application of techniques of psychological measurement to substantive problems in all areas of psychology and related disciplines. For submission information, see the two-page Information for Authors at the end of each issue.

APPLIED PSYCHOLOGICAL MEASUREMENT (ISSN 0146-6216) is published four times annually—in March, June, September, and December—by Sage Publications, 2455 Teller Road, Thousand Oaks CA 91320. Telephone: (800) 818-SAGE (7243) and (805) 499-9774; Fax/Order line: (805) 375-1700.

Subscriptions: Annual subscription rates for institutions and individuals are based on the current frequency. Prices quoted are in U.S. dollars and are subject to change without notice. Canadian subscribers add 7% GST (and HST as appropriate). Outside U.S. subscription rates include shipping via air-speeded delivery. Institutions: $364 (within the U.S.) / $380 (outside the U.S.) / single issue: $105 (worldwide). Individuals: $57 (within the U.S.) / $73 (outside the U.S.) / single issue: $24 (worldwide). Orders with ship-to addresses in India and South Asia should be sent to the New Delhi address (below). Noninstitutional orders must be paid by personal check, VISA, Discover, or MasterCard.

This journal is abstracted in **Abstract Journal of the Educational Resources Information Center, Australian Education Index, Current Contents: Social & Behavioral Sciences, Current Index to Journals in Education (CIJE), Mathematical Reviews, Psychological Abstracts, PsycINFO, PsycLIT, Social Sciences Citation Index,** and **SRM Database of Social Research Methodology,** and is microfilmed by **UMI**.

Back Issues: Information about availability and prices of back issues may be obtained from the publisher's order department (address below). Single-issue orders for five or more copies will receive a special adoption discount. Contact the order department for details. Write to the London office for sterling prices.

Inquiries: All subscription inquiries, orders, and renewals with ship-to addresses in North America, South America, Australia, China, Indonesia, Japan, Korea, New Zealand, and the Philippines must be addressed to Sage Publications, 2455 Teller Road, Thousand Oaks CA 91320, U.S.A., telephone: (800) 818-SAGE (7243) and (805) 499-9774, fax: (805) 375-1700. All subscription inquiries, orders, and renewals with ship-to addresses in the U.K., Europe, the Middle East, and Africa must be addressed to Sage Publications Ltd, 6 Bonhill Street, London EC2A 4PU, England, telephone +44 0(20) 7374 0645, fax +44 0(20) 7374 8741. All subscription inquiries, orders, and renewals with ship-to addresses in India and South Asia must be addressed to Sage Publications Private Ltd, P.O. Box 4215, New Delhi 110 048 India, telephone (91-11) 641-9884, fax (91-11) 647-2426. Address all permissions requests to the Thousand Oaks office.

Claims: Claims for undelivered copies must be made no later than six months following month of publication. The publisher will supply missing copies when losses have been sustained in transit and when the reserve stock will permit.

Change of Address: Six weeks' advance notice must be given when notifying of change of address. Please send old address label along with the new address to ensure proper identification. Please specify name of journal. POSTMASTER: Send address changes to APPLIED PSYCHOLOGICAL MEASUREMENT, c/o 2455 Teller Road, Thousand Oaks CA 91320.

Printed on acid-free paper

New *APM* Editor Now Accepting New Manuscript Submissions

David J. Weiss
University of Minnesota

I am pleased to announce the appointment of Mark D. Reckase as *APM's* second editor. Mark is well-known to most of *APM's* readers for his contributions to several important developments in item response theory (IRT) and other aspects of applied measurement. Mark's initial term as editor begins immediately and extends through December, 2006.

Mark Reckase received his B. A. in Psychology from the University of Illinois in 1963. He received both his M. A. (1971) and Ph. D. (1972) in Psychology from Syracuse University, under the guidance of Eric Gardner. His Ph. D. dissertation, "Development and Application of a Multivariate Logistic Latent Trait Model," was a major contribution to the developing field of IRT, especially given the relatively undeveloped state of IRT in the early 1970s.

Following his formal academic training, Mark took a position as Assistant Professor of Educational Psychology at the University of Missouri at Columbia. He remained there as an Associate Professor until 1981. While at Missouri, Mark developed a very active research program in adaptive/tailored testing, which contributed considerably to the early knowledge base in that developing field. From Missouri, Mark moved to ACT, Inc., where he supervised the development of the ACT Assessment Program, the design and development of ACT-owned testing, and performed research on computerized adaptive testing, multidimensional item response theory, and performance assessment. He left ACT, Inc., where he had risen to the position of Assistant Vice President of the Assessment Innovations Area, in 1998. Mark then joined the faculty of Michigan State University, where he is currently Professor of Education in the Measurement and Quantitative Methods Program.

Mark's current interests and contributions range across almost all areas of applied and theoretical psychometrics. He has consulted with organizations worldwide on numerous measurement problems and has authored or coauthored more than 80 publications. He also comes into this position with considerable editorial experience. In addition to serving on *APM's* Editorial Board since 1995, during which time I have depended on him quite heavily for evaluations of manuscripts concerned with adaptive testing, multidimensional IRT, and other advanced IRT issues, Mark has been on the editorial boards of a half dozen other measurement journals. He also was editor of APA Division 5's *Score* newsletter for four years, and was editor of the *Journal of Educational Measurement* from 1993 to 1995. Mark Reckase is extremely well qualified to move *APM* forward to the next level.

Manuscript Submissions

Mark will begin accepting new manuscripts submissions immediately. Authors should send three copies of new manuscripts to: Mark D. Reckase, Michigan State University,

Applied Psychological Measurement, Vol. 25 No. 4, December 2001, 315–316

461 Erickson Hall, East Lansing MI 48824-1034, U.S.A.; Phone: 517-355-8537; Fax: 517- 353-6393; Email: reckase@msu.edu. Mark has already begun by taking full responsibility for Volume 26 with publication of the March 2002 issue, using manuscripts that I have accepted and that are currently in the publication queue. Manuscripts that are now being reviewed will be forwarded to Mark for his action when the reviews are complete, including manuscripts that have been accepted with revision. I will attempt to inform authors by email when I forward their manuscripts to Mark.

A Personal Note

I would like to extend my thanks to all those who contributed to the development of *APM* over the last 25 years. All of those currently on the Editorial Board, as well as those who formerly served, have been largely responsible for maintaining *APM's* high standards of quality over the years. In addition, many of you have contributed your expertise to developing and maintaining these standards by serving as ad hoc reviewers of manuscripts. Manuscript reviewing is generally a thankless—but essential—task. Without the thousands of hours of time contributed by all of these reviewers, *APM* would not exist as it does today.

Many others have contributed to *APM*, as well. For its first four years, *APM* was published by the new College Department of West Publishing Company. West's staff developed the original design for *APM*, which we still use today. Its staff launched the journal and started it on its way. For the next 16 years, *APM* was published by a nonprofit corporation created exclusively for publishing *APM*. During that time, and through this issue, I was assisted by a succession of dedicated University of Minnesota students who served as my technical editors, and a number of people who typeset the issues. For the last five volumes, the staff at Sage Publications has contributed in many ways to *APM's* continuing success. Finally, *APM* would not exist without its authors. They deserve special thanks for the excellence of their research, for their conscientiousness in facilitating the manuscript flow for the issues, and for their cooperation in revising their manuscripts to meet our exacting standards.

In 1977, when we started *APM*, the field of psychometrics was stagnating and mired in relatively unproductive methods that had been in use for 75 years. I believe that *APM* has been instrumental in providing a forum for the introduction and investigation of new methodologies that have provided the field with a new vitality and sense of direction that perhaps it has not seen in the almost 100 years since psychological measurement began.

Precision of Warm's Weighted Likelihood Estimates for a Polytomous Model in Computerized Adaptive Testing

Shudong Wang, Assessment Systems, Inc.

Tianyou Wang, ACT, Inc.

This monte carlo study evaluated the relative accuracy of Warm's (1989) weighted likelihood estimate (WLE) compared to the maximum likelihood estimate (MLE), expected a posteriori (EAP) estimate, and maximum a posteriori (MAP) estimate. The generalized partial-credit model was used under a variety of computerized adaptive testing (CAT) conditions. The results indicated that WLE was more accurate than MLE with a fixed-length CAT, consistent with previous findings. WLE and MLE had smaller bias and larger standard errors than EAP and MAP. EAP was more accurate than MAP in a variety of CAT conditions. Although root mean squared errors were different among the four estimation methods, no statistically significant mean differences were found. EAP and MAP had advantages over WLE and MLE in terms of test efficiency. These results suggest that the test termination rule has more impact on the accuracy of θ estimation methods than does the item bank size. *Index terms: adaptive test termination, computerized adaptive testing, item response theory, polytomous responses, θ estimation methods.*

Computerized adaptive testing (CAT) is of considerable interest to the measurement and research community because of its advantages over traditional paper-and-pencil tests (Kingsbury & Weiss, 1983; Lord, 1977; McBride & Martin, 1983; Urry, 1977; Wainer et al., 1990). Most CAT applications have been based on dichotomous items. CAT with polytomous models can be applied to items with rating scales, or in some situations, with multiple-choice items. However, polytomous CAT has not been used widely because machine scoring of polytomous items still is difficult to achieve. Bennett, Steffen, Singley, Morley, & Jacquemin's (1997) research in the computer scoring of open-ended items, however, suggests that the application of CAT with polytomous scoring of open-ended items might soon be practical.

Trait (θ) estimation is one of the most important components in a CAT system. The accuracy of θ estimation has a significant impact on the quality of CAT—it affects not only the final score reported, but also item selection and test termination. The properties of various θ estimation methods have been studied extensively for CAT with dichotomous item response theory (IRT) models (e.g., Bock & Mislevy, 1982; Wang, Hanson, & Lau, 1999; Wang & Vispoel, 1998).

Among the methods studied, Warm's (1989) weighted likelihood estimate (WLE) was found particularly promising, because it not only reduces the bias of maximum likelihood estimates (MLEs), but also reduces standard errors (SEs) for CAT with fixed test length. Bayesian methods [e.g., expected a posteriori (EAP) and maximum a posteriori (MAP)], although having smaller SEs than non-Bayesian methods, have large bias toward the prior mean. Warm claimed that the WLE method removed the first-order bias term from the MLE. Samejima (1998) expanded the WLE method to include general polytomous IRT models.

Applied Psychological Measurement, Vol. 25 No. 4, December 2001, 317–331

In general, unbiased parameter estimation is desirable. Reducing bias in θ estimation can be important in several situations (Wang et al., 1999). For example, when comparability between CAT and paper-and-pencil tests is needed, unbiased θ estimates might help to reduce the need to equate these different versions (Eignor & Schaeffer, 1995; Segall, 1995; Segall & Carter, 1995; Wang & Kolen, 2001). Another situation is when cut scores are set on the θ scale in domain-referenced, certification, or licensure testing. Bias can systematically affect the precision of the cut score and, consequently, the validity of the classification decisions. However, bias might be of little concern when the main purpose of CAT is to rank order examinees. This study investigated the properties of the WLE in CAT using a generalized partial-credit model (GPCM) and compared it to those of MLE, EAP, and MAP.

Reducing MLE Bias

Firth (1993) identified two approaches to reducing MLE bias—corrective and preventive.

The Corrective Approach

The corrective approach includes two methods that have been extensively studied in the literature. The first includes computationally intensive procedures, such as the jackknife and bootstrap methods (Quenouille, 1949, 1956). The second simply subtracts an estimate of the first-order bias, $Bias_1[\mathrm{MLE}(\hat{\theta})]$, from the MLE. The bias-corrected estimate is then

$$\hat{\theta} = \hat{\theta}_{\mathrm{MLE}} - Bias_1\left[\mathrm{MLE}(\hat{\theta})\right] . \tag{1}$$

Both methods might succeed in removing $Bias_1[\mathrm{MLE}(\hat{\theta})]$ (Warm, 1989). A common feature of the two methods is that $\hat{\theta}$ first is calculated and then is corrected. For this reason, both methods require the existence of a finite $\hat{\theta}$. (MLEs are infinite for all correct or all incorrect responses.)

The Preventive Approach

The preventive approach (Firth, 1993) modifies the score function before the MLE is calculated. In general, the MLE is derived as a solution to the score equation,

$$S(\theta) = \frac{\partial}{\partial\theta} l(\theta) = 0 , \tag{2}$$

where $l(\theta) = \ln L(\theta)$ is the log-likelihood function for any given model.

The bias in $\hat{\theta}$ can be reduced by introducing a (first-order) bias term into the score function (Firth, 1993). For a given bias, $B(\theta)$, $S(\theta)$ can be modified to be $S^*(\theta)$ by simple triangle geometry (see Figure 1). The Fisher information function is

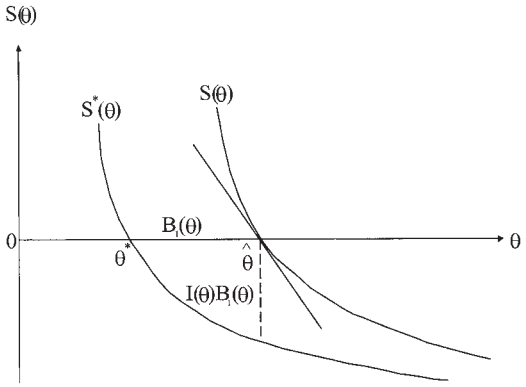
$$I = I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln L(\mathbf{u}|\theta)\right] = -E\left[\frac{\partial^2}{\partial\theta^2} l(\theta)\right] = -E\left[\frac{\partial}{\partial\theta} S(\theta)\right] , \tag{3}$$

which is the expectation of the negative value of $S(\theta)$ tangent at $\hat{\theta}$; that is, $-I(\theta) = S'(\theta)$. For any given bias $B(\theta)$, $S(\theta)$ then can be shifted by $I(\theta)B(\theta)$. The modified score function then is

$$S^*(\theta) = S(\theta) - I(\theta)B(\theta) . \tag{4}$$

Hence, a modified estimate, θ^*, is given by solving for $S^*(\theta) = 0.0$. In general, $O(n^{-1})$ bias (where n is the number of items) can be removed from the MLE by introducing an appropriate bias term into

Figure 1
Modification of the Score Equation



the score function. However, bias reduction is not always desirable. The merits of bias reduction in any particular application depend on a number of factors (Copas, 1988).

Although Warm (1989) did not have a general rationale in deriving his WLE method for reducing the bias of MLE, his method actually is a special case of Firth's (1993) preventive approach (Wang et al., 1999). Warm derived the WLE method based on a conjecture from Lord's (1983, 1986) MLE bias functions and Bayesian modal estimation method. His final result, however, is equivalent to Equation 4.

WLE for Polytomously Scored Responses

Samejima (1998) expanded Warm's WLE method to include polytomously scored responses. Samejima used the correct adjusting term for reducing bias (see Equation 4), apparently without being aware of Firth's (1993) general rationale. To express this adjusting term, Samejima (1993a, 1993b) first generalized Lord's MLE bias function for dichotomous responses to the case of polytomous responses. She then applied Warm's WLE to the polytomous responses.

The MLE bias function for general discrete responses is

$$Bias\left[\mathrm{MLE}(\theta,\hat{\theta}_v)\right] \cong E\left(\hat{\theta}_v - \theta|\theta\right) \cong -\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k=0}^{m_j}\frac{\frac{\partial}{\partial\theta}P_{jk}(\theta)\frac{\partial^2}{\partial\theta^2}P_{jk}}{P_{jk}(\theta)}\ , \tag{5}$$

where

$P_{jk}(\theta)$ is the probability of a correct response in category k to item j, assuming that $P_{jk}(\theta)$ is at least five times differentiable with respect to θ;

$\hat{\theta}_v$ is the MLE of θ based on the response pattern v; and

m_j is the number of discrete responses to item j.

When $P_{jk}(\theta)$ is modeled by Samejima's (1969) graded response model (GRM), the bias function (Equation 5) can be expressed as Equation 14 (see the Appendix). When $P_{jk}(\theta)$ is modeled by Muraki's (1992) GPCM, the bias function can be expressed as Equation 17 (see the Appendix). A straightforward expansion of Warm's WLE for the three-parameter logistic model to general discrete responses is

$$\sum_{j=1}^{n} \frac{\partial}{\partial\theta} \ln P_{jk}(\theta) - Bias\left[\mathrm{MLE}(\theta, \hat{\theta})\right] I(\theta) \equiv 0 \,, \tag{6}$$

where $I(\theta)$ is the test information function and $w(\theta)$ is the weight function. Muraki & Bock (1999) provided a computer program to calculate the WLE for the GPCM and GRM.

Method

Data

Monte carlo methods were used to evaluate the polytomous θ estimation method. 21 true θ values, ranging from -4.0 to 4.0 in increments of .4, were used for the GPCM. CAT was simulated for 500 examinees at each of the 21 true θ parameter points. Both descriptive methods and inferential procedures were used to analyze the results from the simulation. The descriptive methods provided global summaries of the results. The inferential procedure overcame some of the deficiencies of the descriptive methods by conceptualizing the study as a statistical sampling experiment (Harwell, 1997; Spence, 1983).

Independent Variables

θ estimation methods. The primary independent variable was θ estimation method in the context of the GPCM. The four θ estimation methods were MLE, WLE, EAP, and MAP. The relationships among this independent variable and other independent variables (described below) were considered. For each method, the values of the true θ parameter used in this study were spaced equally across a fixed range.

Test termination rules. The second independent variable was the test termination rule. The types of test termination rules investigated were fixed number of items and fixed test reliability.

1. *Fixed test length.* CAT was terminated after a specified number of items had been administered. The four test lengths used were 5, 10, 15, and 20 items. Previous studies (De Ayala, 1992; Dodd, Koch, & De Ayala, 1989; Koch & Dodd, 1985) used fixed test lengths from 10 to 30 items. To search for the smallest acceptable test length, a test length of 5 items was also included in the present study.
2. *Fixed test reliability.* CAT was terminated when specified levels of estimated reliability were reached. Because the relationship between reliability (ρ) and the SE of θ estimation can be expressed as $\rho = 1 - SE^2$, given a θ variance of 1.0 (Wang et al., 1999), estimated reliability and estimated SE have the same effect on CAT termination. MLE and WLE use the square root of the reciprocal of test information as the SE. EAP and MAP Bayesian methods use the standard deviation of the posterior distribution as the SE. The three values of reliability used were .7, .8, and .9, which correspond to SEs of .55, .45, and .32 and test information values of 3.33, 5, and 10, respectively. A maximum test length of 33 items was used to terminate the test for a given simulee if the prespecified levels of reliability could not be reached. This length was chosen because the smallest bank size was 33 items.

Item banks. The third independent variable was the item bank size. Although some research has found that polytomous item banks with 30 items might be sufficient for accurate θ estimation with few nonconvergence problems (Dodd & De Ayala, 1994; Dodd et al., 1989; Dodd & Koch, 1987; Dodd, Koch, & De Ayala, 1993; Koch & Dodd, 1989), these findings do not imply that any item bank composed of 30 or more items will be sufficient for polytomous CAT (Dodd, De Ayala, & Koch, 1995).

The item bank, 1GP, consisted of 263 polytomously scored items taken from the 1996 National Assessment of Educational Progress Science Assessment. The items had 3 to 5 categories for each of Grades 4, 8, and 12. Item parameters for the GPCM were calibrated by Educational Testing Service, using a very complex sampling design and scaling methodology. (For details, see Allen, Carlson, & Zelenak, 1999.) A matrix sampling procedure, balanced incomplete block spiraling, required examinees to take different subsets of items from the item bank. Item subsets were balanced and spiraled among the examinees in test administrations. Item parameters were calibrated using a special combination of BILOG and PARSCALE. The dimensionality analyses (Allen et al., 1999, pp. 191–192) indicated that the response data were close to unidimensional. For purposes of the present study, item parameter estimates were treated as if they were true parameters.

Two additional item banks were created form the 1GP bank. These banks had 66 items (2GP) and 33 items (3GP). Items were randomly selected from the 1GP bank using the proportional stratified random sampling method (Gall, Borg, & Gall, 1996), based on the same proportions of items with different category numbers in the 1GP bank. Table 1 provides summary descriptive statistics for the item parameters of these three item banks. The initial $\hat{\theta}$ for all CAT was 0.0. Items were selected by maximum information and infinite MLE $\hat{\theta}$s were set at -5 or 5.

Table 1
Descriptive Statistics for Item Parameter Estimates From the 1GP, 2GP, and 3GP Item Banks Under the GPCM

Bank/ Parameter	No. Items	Mean	Median	SD	Minimum	Maximum
1GP	263					
a		.549	.522	.229	.105	1.871
b_1		.713	.720	2.011	−6.972	11.746
b_2		1.270	1.264	2.640	−17.381	13.926
b_3		1.034	1.004	2.371	−6.369	7.187
b_4		.822	.822	2.546	−3.159	4.924
2GP	66					
a		.539	.527	.171	.171	1.200
b_1		1.066	1.000	1.728	−3.204	7.399
b_2		1.679	1.491	2.519	−2.665	13.926
b_3		1.832	1.412	1.656	−.856	5.506
b_4		4.270	4.270	.535	.535	4.925
3GP	33					
a		.560	.523	.190	1.900	1.055
b_1		.752	.631	1.384	−2.738	3.437
b_2		1.695	1.684	2.495	−3.638	7.293
b_3		1.467	1.680	3.480	−6.369	7.187
b_4		2.000	2.000	.000	2.000	2.000

Dependent Variables

Based on previous literature (Wang & Vispoel, 1998), four criterion variables (or their log transformations) were used: (1) bias, (2) SEs, (3) root mean squared errors (RMSEs) for estimated and true θ parameters, and (4) administrative efficiency (mean numbers of items needed to reach a criterion SE level). Each criterion provided complementary evidence. Bias, SE, and RMSE were determined conditionally and averaged across the θ distribution. Conditional indices were computed at each θ point, and mean (overall) indices were computed by taking the absolute values of the conditional

indices and integrating them over a normally distributed θ for a population of examinees, using the numerical integration method (Wang & Vispoel, 1998).

Conditional indices. The conditional indices were

$$Bias(\hat{\theta}) = \frac{1}{N}\sum_{r=1}^{N}\left(\hat{\theta}_r - \theta\right) , \tag{7}$$

$$SE(\hat{\theta}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(\hat{\theta}_r - \frac{\sum_{t=1}^{N}\hat{\theta}_t}{N}\right)^2} , \tag{8}$$

and

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(\hat{\theta}_r - \theta\right)^2} , \tag{9}$$

where

θ is true (generating) trait level,

$\hat{\theta}_r$ is the estimated θ for the rth replication, and

N is the number of replications.

The number of replications was the analogue of sample size. Because the primary goal was to assess the relative accuracy of θ estimation methods, statistical significance was tested and the empirical sampling distributions for the statistics were generated. To minimize the sample variance and increase the power to detect the effects of interest, a large number of replications was desired; 500 replications were considered sufficient (Stone, 1993). The RMSE can be separated into two components, bias and SE—$RMSE^2 = Bias^2 + SE^2$.

Overall indices. The overall indices were

$$\text{mean}(BIAS) = \sum_{i=1}^{21} |Bias(\hat{\theta})| \Big| \theta_i \times q(\theta_i) , \tag{10}$$

$$\text{mean}(SE) = \sqrt{\sum_{i=1}^{21} SE^2(\theta) \big| \theta_i \times q(\theta_i)} , \tag{11}$$

and

$$\text{mean}(RMSE) = \sum_{i=1}^{21} RMSE(\hat{\theta}) \Big| \theta_i \times q(\theta_i) , \tag{12}$$

where $q(\theta_i)$ are quadrature weights based on the standard normal distribution, and θ_i are the 21 equally spaced true θ levels. Descriptive statistics are provided for all conditional and overall indices. Inferential statistics were used only for the overall indices.

Experimental Design

Two experimental designs were used in the analyses of the overall indices. For the fixed-length tests, a 4 θ estimation methods $\times$ 3 bank sizes $\times$ 4 test lengths completely crossed analysis of variance (ANOVA) design was used. For the fixed-reliability tests, a 4 θ estimation methods $\times$ 3 item bank sizes $\times$ 3 reliability levels completely crossed ANOVA design was used. Effect sizes were computed and tabulated with significance level statistics. [Partial results are presented here; for complete results, see Wang (1999).]

Results

Conditional Indices

Fixed test length. Figure 2 shows bias, SE, and RMSE of the four θ estimation methods as a function of θ for 5- and 10-item CATs using the 1GP bank. Figures 2a and 2b show that WLE had the smallest bias over the largest θ range for all test lengths. This result agrees with those for a dichotomous model (Wang, Hanson, & Lau, 1989).

WLE and MLE had considerably less bias than EAP and MAP. Although WLE had almost no bias, MLE had "outward bias"—that is, its bias was positively correlated with θ. EAP and MAP had "inward bias"—the bias was negatively correlated with θ.

Figures 2c and 2d show that WLE had lower SEs than MLE at almost all θ levels for both fixed test lengths, although the differences were small. Thus, WLE reduced both the SE and bias of MLE. For both θ extremes, EAP and MAP had substantially lower SEs than WLE and MLE. However, this reduction of SE was at the expense of increased bias. This also is consistent with previous findings for dichotomous models (Wang, Hanson, & Lau, 1999; Warm, 1989). Figures 2e and 2f show that, for both test lengths, there was a difference in RMSE patterns between Bayesian and the MLE/WLE methods and little differences within these two types of methods.

Fixed reliability. Figures 3a and 3b show the effects of fixing the test reliability at .7 and .9 with the 1GP bank. The fixed reliability changed the bias direction of MLE from outward to slightly inward. WLE further increased that inward bias, which is the reverse of what happened with a fixed test length. This means that, under a fixed reliability rule, WLE failed to reduce the bias of MLE.

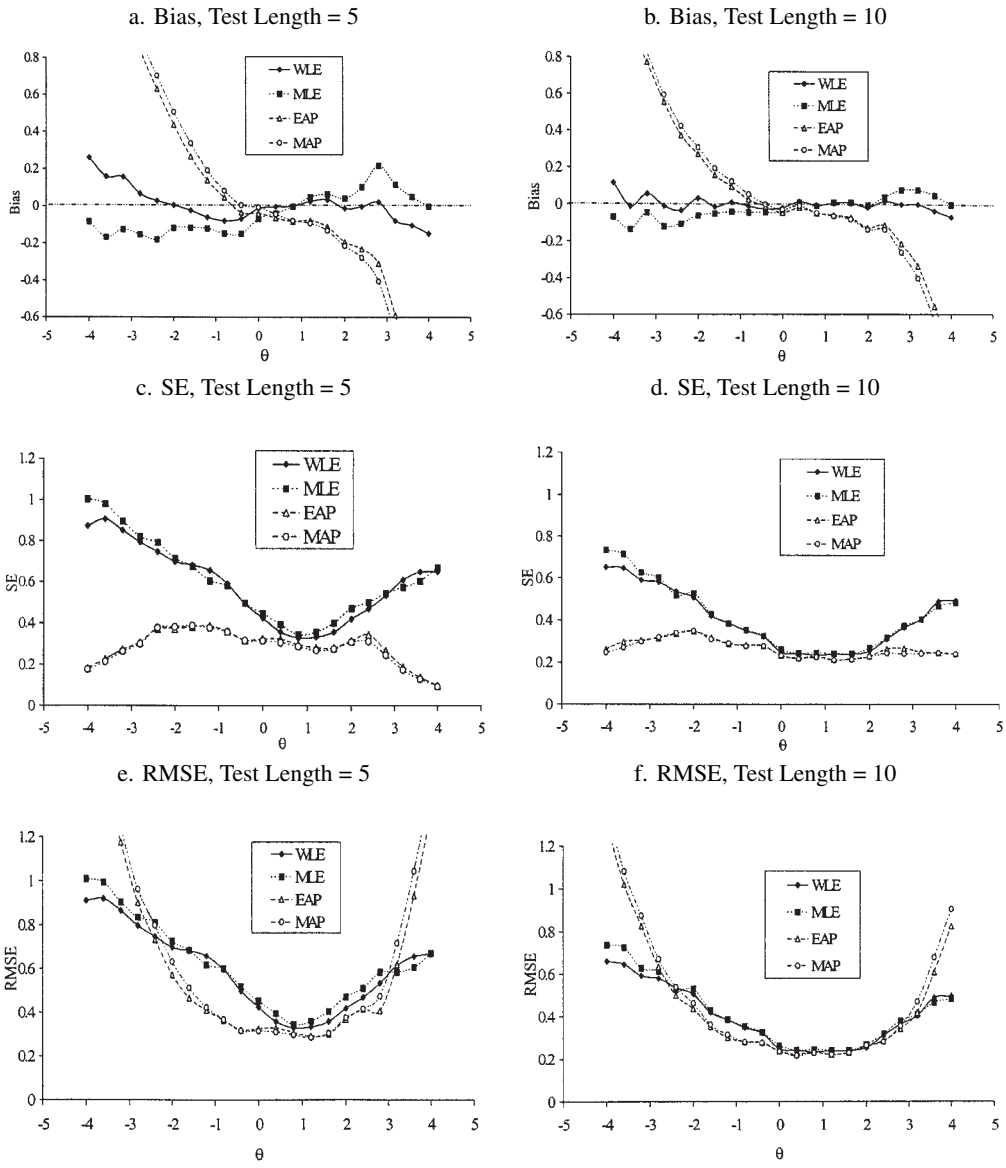
MLE and WLE had remarkably smaller bias than EAP and MAP. Figures 3c and 3d show that WLE and MLE had larger SEs than EAP and MAP; WLE had smaller SE than MLE. Figures 3e and 3f show that MLE and WLE had lower RMSEs than EAP and MAP at extreme θ levels, but slightly higher RMSEs at midrange θs. WLE performed slightly better than MLE. All bias, SEs, and RMSEs decreased as the test reliability increased, except for the results in Figure 3c for MLE and WLE at $\theta = 2.0$ and above.

Figure 4 shows the mean number of items required for achieving reliabilities of .7 and .9. Averaged across θ levels for reliability of .7, the number of items required for MLE and WLE were approximately twice those for EAP and MAP. For reliability of .9, MLE and WLE had smaller mean numbers of items than EAP and MAP, but differences were not as large as for reliability of .7. Item bank size had slight effects on all conditional indices (see Wang, 1999).

Overall Indices

Table 2 shows the results of the three-way ANOVA of absolute bias, SE, and RMSE (averaged across θ levels) for the fixed test length termination and the fixed reliability conditions. In general, the results for the overall indices further support the results of the conditional indices. θ estimation methods had the most influence on absolute bias—they accounted for 54.0% of the total variance of absolute bias for the fixed test length termination condition and 74.0% of the total variance of

Figure 2
Bias, SE, and RMSE of θ Estimation Methods For 5- and 10-Item CAT

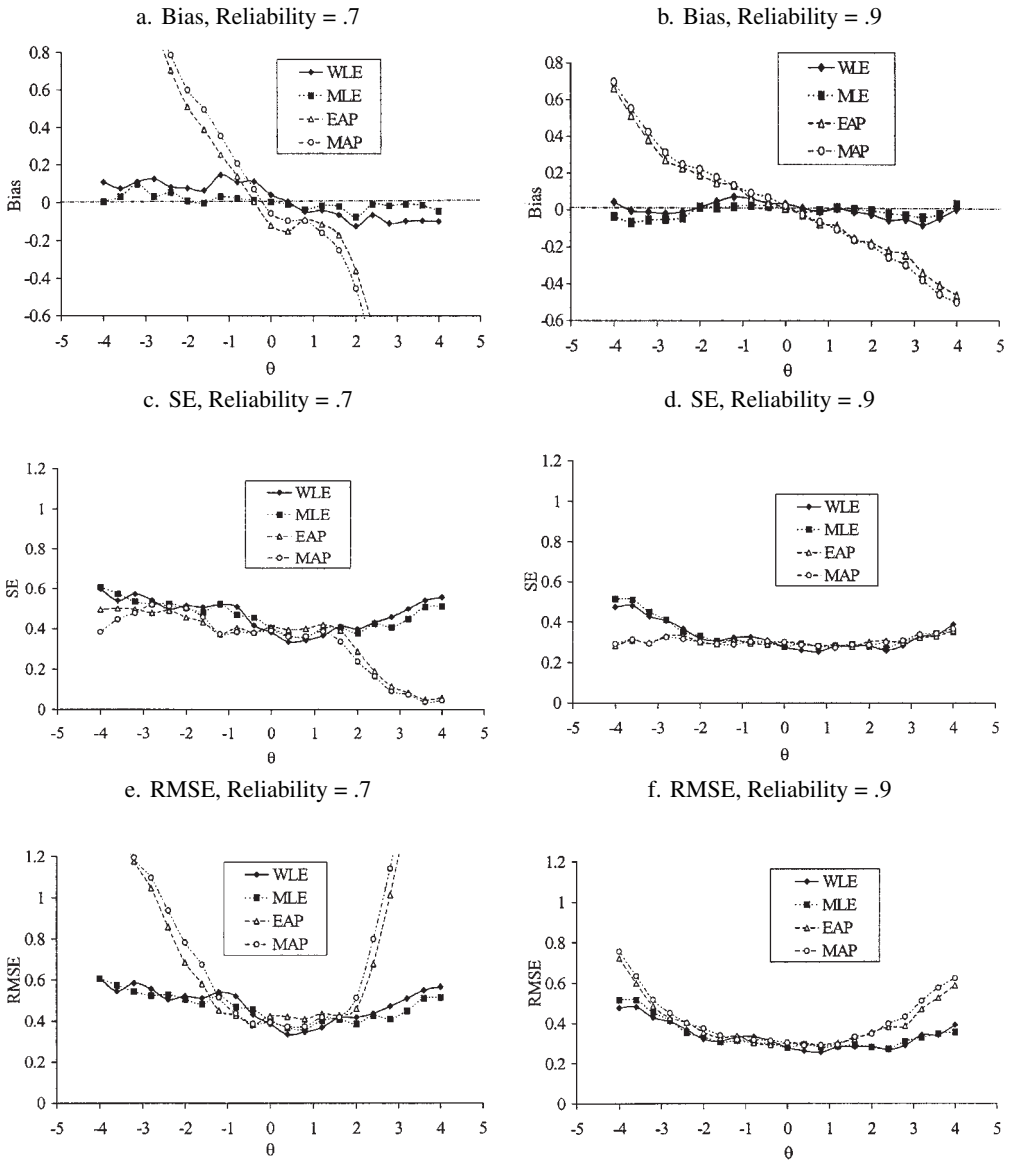


absolute bias for the fixed reliability termination condition. WLE and MLE showed significantly less absolute bias than Bayesian methods, with WLE performing best. For 5-item tests, the overall mean absolute bias for WLE, MLE, EAP, and MAP were .032, .078, .102, and .094, respectively.

For the fixed test termination criterion, the test length factor accounted for 65.9% of the total variance in SE and 31.8% of the variance in the estimated absolute bias. For fixed test reliability termination, the reliability factor accounted for 15.3% of the total variance.

SE was most affected by the test termination rule. When test length was fixed, it accounted for 65.9% of the total variance in mean SE; when reliability was fixed, it accounted for 59.1%. The

Figure 3
Bias, SE, and RMSE of θ Estimation Methods for CAT With Reliability of .7 and .9

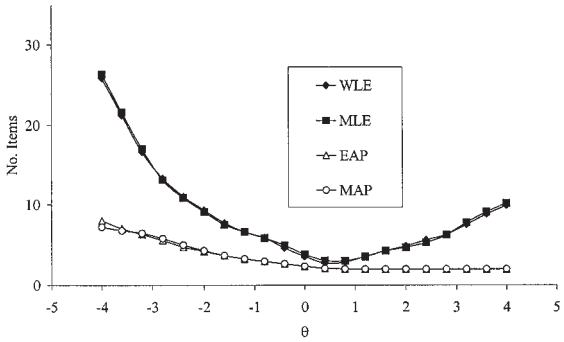


second most influential factor for SE varied by termination rule. For the fixed test length condition, θ estimation method accounted for 16.7% of the variance in SE. The interaction between item bank size and level of reliability accounted for 14.3% of the variance in SE. For fixed test length, WLE had significantly lower SEs (mean SE was .066) than MLE (mean SE was .071). For fixed test reliability, EAP had significantly higher SEs (mean SE was .076) than MAP (mean SE was .073).

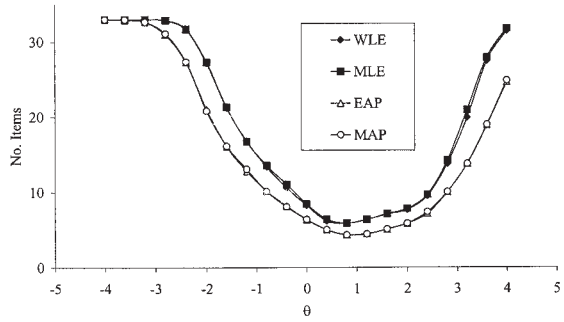
There were no significant difference in RMSE among θ estimation methods, except the difference between WLE (mean RMSE was .279) and MLE (mean RMSE was .319) for fixed-length tests. RMSE

Figure 4
Mean Number of Items Required for Test Termination

a. Reliability = .7



b. Reliability = .9



(log of RMSE) was affected most by the test termination rule. When test length was fixed, it accounted for 31.2% of the total variance of RMSE, and the three-way interaction of method × size × length accounted for 16.1%. When reliability was fixed, it accounted for 87.6% of the total variance of RMSE and θ estimation accounted for 6.3%. There were no significant differences in RMSE among θ estimation methods, except the difference between WLE and MLE.

Conclusions

In general, for all four θ estimation methods, conditional and overall bias, SE, and RMSE decreased as the test length, test reliability, and item bank size increased. The magnitudes of the differences among the dependent variables decreased as the values of independent variables increased.

WLE performed better than MLE on all the dependent variables studied for fixed-length tests, and WLE performed better than EAP and MAP in terms of bias. MLE had less bias than both Bayesian methods. EAP and MAP had lower SEs than WLE or MLE. EAP had better results than MAP for almost all conditions. Test termination rules had a significant impact on the dependent variables, especially for WLE and MLE.

Although in prior research the quality of item banks had effects on the conditional distribution of bias, SE, RMSE, and test efficiency (Wang & Vispoel, 1998), in the present study the item bank size had less impact on the differences among the dependent variables than test termination rules.

Table 2
Results of ANOVA With Fixed Test Length and Fixed Test Reliability Termination Rules

Rule/		Absolute Bias			Average SE			Average ln(RMSE)		
Effect	DF	F	p	η^2	F	p	η^2	F	p	η^2
Fixed Test Length										
Main Effects										
Method (M)	3	1039.952	.000	.540	122.258	.000	.167	3.361	.026	.052
Bank Size (S)	2	59.298	.000	.021	66.079	.000	.060	1.473	.238	.015
Test Length (L)	3	613.137	.000	.318	483.836	.000	.659	20.135	.000	.312
Interactions										
M × S	6	37.496	.000	.039	.605	.725	.002	1.731	.134	.054
M × L	9	38.191	.000	.059	18.696	.000	.076	1.559	.155	.072
S × L	6	2.336	.046	.002	1.904	.099	.005	2.650	.020	.082
M × S × L	18	4.076	.000	.013	1.059	.419	.009	1.733	.066	.161
Error	48									
Fixed Test Reliability										
Main Effects										
Method (M)	3	5774.783	.000	.740	60.737	.000	.021	124.574	.000	.062
Bank Size (S)	2	7.463	.002	.001	590.280	.000	.139	8.993	.000	.003
Reliability (R)	2	1805.544	.000	.154	2518.213	.000	.591	2632.117	.000	.876
Interactions										
M × S	6	33.933	.000	.009	29.867	.000	.021	6.910	.001	.007
M × R	6	341.405	.000	.087	37.894	.000	.027	9.523	.000	.010
S × R	4	13.018	.000	.002	304.547	.000	.143	7.438	.000	.005
M × S × R	12	11.723	.000	.006	37.996	.000	.054	3.564	.002	.007
Error	36									

This study confirmed Warm's conclusions that (1) WLE is unbiased to first order bias for fixed-length tests, whereas MLE, EAP, and MAP are biased; and (2) the WLE method has small variance over the entire range of θ for fixed-length CAT.

A limitation of this study is that the item parameters were regarded as true item parameters and free of error. Van der Linden & Glas (2000) suggested that errors in item parameter estimates can have a dramatic effect on CAT θ estimation. Further research on the impact of item parameter error on the accuracy of polytomous CAT θ estimation is needed.

Appendix

The MLE bias function for the GRM is

$$B(\theta,\hat{\theta}_v) \cong E\left[\hat{\theta}_v - \theta|\theta\right] \cong -\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k_j}\frac{\frac{\partial}{\partial\theta}P_{jk}(\theta)\frac{\partial^2}{\partial\theta^2}P_{jk}(\theta)}{P_{jk}(\theta)}$$

$$= -\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k_j}\left\{\frac{Da_j\exp[Da_j(\theta-b_{jk})]}{\{1+\exp[Da_j(\theta-b_{jk})]\}^2} - \frac{Da_j\exp[Da_j(\theta-b_{jk+1})]}{\{1+\exp[Da_j(\theta-b_{jk+1})]\}^2}\right\}$$

$$\times\left\{\frac{D^2a_j^2\dfrac{\exp[Da_j(\theta-b_{jk})]-\exp\{2[Da_j(\theta-b_{jk})]\}}{\{1+\exp[Da_j(\theta-b_{jk})]\}^3}}{P_{jk}(\theta)}\right.$$

$$\left.-\frac{\dfrac{\exp[Da_j(\theta-b_{jk})]-\exp\{2[Da_j(\theta-b_{jk+1})]\}}{\{1+\exp[Da_j(\theta-b_{jk+1})]\}^3}}{P_{jk}(\theta)}\right\}, \tag{13}$$

where $I(\theta)$ is test information,

$$I(\theta) = E[I_v(\theta)|\theta] = \sum_v I_v(\theta)P_v(\theta) = \sum_{j=1}^{n} I_j(\theta) = \sum_{j=1}^{n}\sum_{k_j} I_{k_j}(\theta)P_{k_j}(\theta)\ , \tag{14}$$

and $I_j(\theta)$ is item information,

$$I_j(\theta) = \sum_{k=0}^{m_j}\frac{\left[P'_{jk}(\theta)\right]^2}{P_{jk}(\theta)} - 0 = \sum_{k=0}^{m_j}\frac{\left[P'_{jk}(\theta)\right]^2}{P_{jk}(\theta)} = \sum_{k=0}^{m_j}\frac{\left(P^{*\prime}_{jk} - P^{*\prime}_{j,k+1}\right)^2}{\left(P^{*}_{jk} - P^{*}_{j,k+1}\right)}\ . \tag{15}$$

When $P_{jk}(\theta)$ is modeled by the GPCM, the MLE bias function becomes

$$B(\theta,\hat{\theta}_v) \cong E\left[\hat{\theta}_v - \theta|\theta\right] \cong -\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k_j}\frac{\frac{\partial}{\partial\theta}P_{jk}(\theta)\frac{\partial^2}{\partial\theta^2}P_{jk}}{P_{jk}(\theta)}$$

$$= -\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k_j} D^3a_j^3P_{jk}\left(k - \sum_{c=0}^{m} cP_{jk}\right)$$

$$\times \left[k^2 - 2k \sum_{c=0}^{m} c P_{jk} + 2 \left(\sum_{c=0}^{m} c P_{jk} \right)^2 - \sum_{c=0}^{m} c^2 P_{jk} \right] , \tag{16}$$

where the relationship between $I(\theta)$ and $I_j(\theta)$ is the same as in Equation 14. $I_j(\theta)$ becomes

$$\begin{aligned} I_j(\theta) &= \sum_{j=1}^{m_j} P_{k_j}(\theta) I_{k_j}(\theta) = \sum_{k=1}^{m_j} P_{k_j}(\theta) \left[-\frac{\partial^2}{\partial \theta^2} \log P_{k_j}(\theta) \right] \\ &= \sum_{k=1}^{m_j} P_{k_j}(\theta) \left\{ \left[\frac{\left[P'_{jk}(\theta) \right]}{P_{jk}(\theta)} \right]^2 - \frac{P''_{jk}(\theta)}{P_{jk}(\theta)} \right\} \\ &= D^2 a_j^2 \sum_{k=1}^{m_j} P_{k_j}(\theta) \left\{ \sum_{c=1}^{m_j} T_c^2 P_{jc}(\theta) - \left[\sum_{c=1}^{m_j} T_c P_{jc}(\theta) \right]^2 \right\} \\ &= D^2 a_j^2 \left\{ \sum_{c=1}^{m_j} T_c^2 P_{jc}(\theta) - \left[\sum_{c=1}^{m_j} T_c P_{jc}(\theta) \right]^2 \right\} = D^2 a_j^2 \sum_{c=1}^{m_j} \left[T_c - \overline{T}_j(\theta) \right]^2 P_{jk}(\theta) , \end{aligned} \tag{17}$$

where T_c is a linear function. That is, $T_c = (1, 2, \ldots, c, c+1, \ldots, m_j)$. Then,

$$\overline{T}_j(\theta) = \sum_{c=1}^{m_j} T_c P_{jc}(\theta) . \tag{18}$$

References

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report, NCES 1999-452.* Washington DC: National Center for Educational Statistics.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34,* 162–176.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

Copas, J. B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society, B, 50,* 225–265.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16,* 327–343.

Dodd, B. G., & De Ayala, R. J. (1994). Item information as a function of threshold values in the rating scale model. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 301–317). Norwood NJ: Ablex.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19,* 5–22.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement, 11,* 371–384.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13,* 129–143.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53,* 61–77.

Eignor, D. R., & Schaeffer, G. A. (1995, April). *Comparability studies for GRE General CAT and the NCLEX using CAT.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco CA.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80,* 27–38.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction.* New York: Longman.

Harwell, M. R. (1997). Analyzing the results of Monte Carlo studies in item response theory. *Educational and Psychological Measurement, 57,* 266–279.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–238). New York: Academic Press.

Koch, W. R., & Dodd, B. G. (1985, April). *Computerized adaptive attitude measurement.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Koch, W. R., & Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education, 2,* 335–357.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1,* 95–100.

Lord, F. M. (1983). Unbiased estimation of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48,* 233–245.

Lord, F. M. (1983, August). *Memorandum for: Ms. Stocking, Ms. M. Wang, Ms. Wingersky. Subject: Sampling variance and bias for MLE and Bayesian estimation of* θ [Internal Memorandum]. Princeton NJ: Educational Testing Service.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23,* 157–162.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 224–236). New York: Academic Press.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Muraki, E. & Bock, D. (1999). *PARSCALE: Parameter scaling of rating data (Version 3.5)* [Computer software]. Lincolnwood IL: Scientific Software International.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, B, 11,* 68–84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika, 43,* 353–360.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No. 17,* 34.

Samejima, F. (1993a). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika, 58,* 119–138.

Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika, 58,* 195–209.

Samejima, F. (1998, April). *Expansion of Warm's weighted likelihood estimator of ability for the three-parameter logistic model to general discrete responses.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego CA.

Segall, D. O. (1995, April). *Equating the CAT-ASVAB: Experiences and lessons learned.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Segall, D. O., & Carter, G. (1995, April). *Equating the CAT-GATB: Issues and approach.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Spence, I. (1983). Monte carlo simulation studies. *Applied Psychological Measurement, 7,* 405–425.

Stone, C. A. (1993, July). *The use of multiple replications in IRT based Monte Carlo research.* Paper presented at the European meeting of the Psychometric Society, Barcelona, Spain.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14,* 181–196.

van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13,* 35–53.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale NJ: Erlbaum.

Wang, S. (1999). *The accuracy of ability estimation methods for computerized adaptive testing using*

the generalized partial credit model. Unpublished doctoral dissertation, University of Pittsburgh.

Wang, T. (1995). *The precision of ability estimation methods in computerized adaptive testing.* Unpublished doctoral dissertation, University of Iowa, Iowa City (UM No. 9945102).

Wang, T., Hanson, B. A., & Lau, C. M. (1999). Reducing bias in computerized adaptive testing trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23,* 263–278.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria, and an example. *Journal of Educational Measurement, 38,* 19–49.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 109–135.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450.

Author's Address

Send requests for reprints or further information to Shudong Wang, Suite 300, Three Bala Plaza West, Bala Cynwyd PA 19004-3481, U.S.A. Email: shudong_wang@asisvcs.net.

Computer Program Exchange

EQUIPERCENT: A SAS Program For Calculating Equivalent Scores Using the Equipercentile Method

Larry R. Price, Southwest Texas State University
Anna Lurie and Chuck Wilkins, The Psychological Corporation

In test development and use, it is often desirable for the user to relate scores measuring the same construct from one test form to another. The linear equating method traditionally has been used for this purpose when the score distributions of two groups are the same or very close. However, when the distributions of the examinees are different, this procedure is inappropriate because the relationship between the scores of the two groups is curvilinear, rather than linear. Equipercentile equating provides a useful alternative in this situation. In the equipercentile approach, test scores are converted to percentile rank scores. Those scores with the same percentile rank on two different forms are considered equivalent.

Although the major statistical software packages (SAS, SPSS, BMDP, SYSTAT) are able to perform the mathematical aspects of equipercentile equating, sample- and problem-specific programs must be written by the researcher in the program's "native language." Given the widespread use of SAS in statistical computing, an equating program written in SAS provides an easily accessible platform for applied and theoretical researchers who need to conduct equipercentile equating.

Description

EQUIPERCENT can be run on SAS Windows, Versions 6.0–8.0. Using the SAS MACRO language within the program enables the user to call large numbers of variables into the program, which is advantageous in large-scale research and development. It can compute equivalent standard scores across different test forms with a large number of variables. The statistical procedure, which uses equipercentile equating with observed scores, is based on the work of Kolen (1995, pp. 40–47). EQUIPERCENT can be modified easily to incorporate the use of scaled scores instead of number-correct scores.

Availability

Information on EQUIPERCENT can be obtained from Larry Price, Southwest Texas State University, Department of Educational Administration and Psychological Services, San Marcos TX 78666, U.S.A. Copies of the source code for the program can be obtained by sending a self-addressed 3.5-inch disk mailer to the above address or by sending an email to lrprice@mindspring.com.

Reference

Kolen, M. J. (1995). *Test equating.* New York: Springer.

a-Stratified Multistage Computerized Adaptive Testing With b Blocking

Hua-Hua Chang, University of Texas
Jiahe Qian, Educational Testing Service
Zhiliang Ying, Rutgers University

Chang & Ying's (1999) computerized adaptive testing item-selection procedure stratifies the item bank according to a parameter values and requires b parameter values to be evenly distributed across all strata. Thus, a and b parameter values must be incorporated into how strata are formed. A refinement is proposed, based on Weiss' (1973) stratification of items according to b values. Simulation studies using a retired item bank of a Graduate Record Examination test indicate that the new approach improved control of item exposure rates and reduced mean squared errors. *Index terms: adaptive testing, a stratification, b blocking, item exposure rate, item selection, test security.*

Chang & Ying (1999) proposed an a-stratified (AS) method for item selection in computerized adaptive testing (CAT). They demonstrated that, when it was used for certain types of item banks, the item exposure rates were automatically controlled while maintaining the accuracy in trait (θ) estimation. Item exposure control and the related test security issues have been a major concern in the development and implementation of CAT. CAT typically selects an item to maximize precision in estimating an examinee's θ. In doing so, certain items tend to be used more often than others, making item exposure rates uneven. Various remedies to control high item exposure rates have been proposed (e.g., McBride & Martin, 1983; Parshall, Davey, & Nering, 1998; Stocking & Lewis, 1995; Sympson & Hetter, 1985; Thomasson, 1995; van der Linden, 1998).

When the three-parameter logistic model (3PLM) is used, the probability of an examinee with latent trait θ_j giving a correct response to item i ($Y_i = 1$) is

$$P(Y_i = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + \exp[-a_i(\theta - b_i)]}, \quad (1)$$

where

a_i is the item discrimination parameter,
b_i is the difficulty parameter, and
c_i is the guessing parameter.

A standard approach to item selection in CAT has been to select the item with the maximum Fisher item information as the next item (Lord, 1980, pp. 151–153) at the currently estimated trait level, $\hat{\theta}$. Using the Fisher information function, items with high a values have high information, provided that b is close to $\hat{\theta}$. Consequently, items with high a values tend to be exposed more frequently than items with low information.

The AS method directly controls item exposure rates by altering the item-selection process. In this method, items are stratified into K strata based on their a values. Accordingly, the item selection process is divided into K stages. In the first stage, items are selected from the first stratum,

Applied Psychological Measurement, Vol. 25 No. 4, December 2001, 333–341

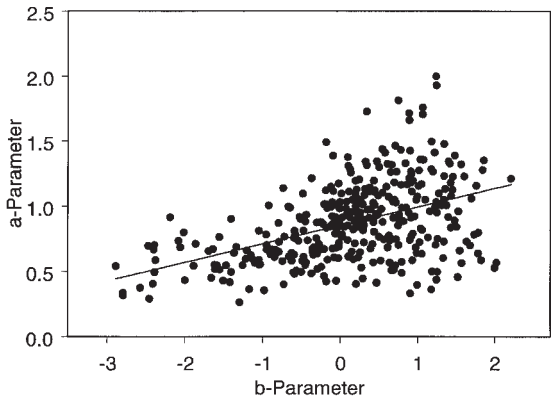
which corresponds to the items with the lowest a values. In the second stage, items are selected from the second stratum. In the Kth stage, items are selected within the Kth level. The rationale behind this approach is:

1. Because estimation of θ could be inaccurate early in the test, it is more appropriate to use low a items. Likewise, items with high a can be more efficiently used later in the test (Chang & Ying, 1996).
2. Selection based largely on item information typically leads to the over-exposure of highly discriminating items. a stratification forces a balanced exposure for all items.

The AS method assumes that the examinee's θ can be matched closely with suitable items at every stage. In other words, at the kth stage, there should be items available with b values that are sufficiently close to $\hat{\theta}$. This entails that the distribution of b should not be influenced by the stratification—i.e., a and b should be uncorrelated. In practice, however, this assumption rarely holds. In fact, a and b parameter estimates often are positively correlated (Lord & Wingersky, 1984).

Figure 1 displays pairs of a and b estimates for 360 items from a retired item bank of a Graduate Record Examination (GRE) quantitative test. The correlation between a and b for these data was .44. When this happens, the strata consisting of items with high a values tend to have high b values. Shortage of lower b items in those strata could cause low b items to be selected more frequently. Parshall, Hogarty, & Kromrey (1999) reported that maximum exposure rates were very high when the AS method was used with some operational item banks.

Figure 1
Relationship Between a and b for 360 GRE Items



An obvious modification to the AS selection method is to balance the distributions of b values among all strata. This can be done by either reassigning items across strata after stratification or by pre-partitioning the item bank before stratification. The latter approach is simpler to implement and follows ideas developed by Weiss (1973) in his stratified adaptive test.

AS CAT With *b* Blocking

A refinement of Chang & Ying's (1999) AS method is introduced. In the AS with b blocking (BAS) method, the basic idea is to force each stratum to have a balanced distribution of b values to ensure a good match of θ for different examinees. This is important, because one of the major goals of CAT is to provide such matching. The BAS method is implemented in the following steps:

1. Divide the item bank into M blocks according to b values. All blocks should have the same number of items (they can differ at most by 1 if the total number of items is not divisible by M). The blocks are arranged in ascending order—the first block contains items with the lowest b values, and the Mth block contains items with the highest b values.
2. Partition each of the M blocks into K strata according to their a values. Thus, for the mth block, the first stratum contains those items with the lowest a values within the block, and the Kth stratum contains items with the highest a values. This stratification procedure is essentially the same as that of Chang & Ying (1999), except that it is performed within a b block.
3. For $k = 1, 2, \ldots, K$, recombine the kth stratum items across M blocks into a single stratum. There are now K strata.
4. Divide the test into K stages.
5. In the kth stage, select items from the kth stratum based on the closeness of b values to the current estimate of θ for an examinee.
6. Repeat Step 5 for $k = 1, 2, \ldots, K$.

Each stratum formed in Step 3 covers approximately the same range of b values. When a and b are uncorrelated, all the strata generated in Steps 1, 2, and 3 should be similar to those formed with Chang & Ying's (1999) AS method, because Steps 1 and 2 essentially yield the same two-way partition as the two-way cross-classification given by a and b under the assumption of no correlation. The two methods will result in different kinds of stratification when a and b are correlated.

The BAS method stipulates even distributions of b values across all strata. However, it also might increase the sample variances of the within-stratum a values. Thus, it is possible that low strata (small values of k) could contain items with high a values, and high strata (large values of k) could contain items with low a values. On average, though, a values will increase in k. Ideally, it is preferable to keep high a items for use in later stages of a test. The increased variability of a values could lead to less-efficient use of items. This is not a very serious problem. It is more important to match b with θ than to adjust a. Also, in the event that θ is not well matched with b, the efficiency could become worse if an item with a higher a value is used (Chang & Ying, 1996).

Simulation Study

A simulation study was conducted to investigate the performance of the BAS method in terms of estimation efficiency, effectiveness in item bank utilization, and test security. Item parameters for the 3PLM from a retired quantitative test from the GRE were used as the item bank. The bank contained 360 items.

Method

The methods used to design and evaluate the simulation study were similar to those of Chang & Ying (1999). However, because the main objective here was to assess possible improvements for the modified stratification, the comparisons were made only between the BAS and the original AS methods.

Design. A fixed test length of 40 items was used. 3,000 θ values were generated from a standard normal N(0, 1) distribution. The 3PLM was used to estimate the item parameters. At each θ value, the item response for each selected item was generated based on the item response function at θ and the corresponding item parameters. Figure 1 shows the scatterplot of estimated as and bs for the 360-item bank.

For the AS method, the item bank was partitioned into four strata in ascending order of a values, where the first stratum contained items with the lowest as and the fourth stratum contained items with the highest as.

For BAS, the item bank was also partitioned into four levels, but the item bank was first blocked into 90 groups that were homogeneous in b. Each group consisted of four items that had the most similar bs. In other words, the first group contained four items with the four lowest bs and the 90th group contained four items with the four highest bs. Next, the item with the lowest a value was taken from each of the 90 groups to form the first level. The item with the second-lowest a then was taken from each of the 90 groups to form the second level. The third and fourth levels were created in the same way, containing items with the third- and fourth-lowest as, respectively. Each level contained 90 items.

Table 1 gives summary statistics for the two stratification methods. In both cases, the mean as for the four levels were naturally ordered. As expected, the standard deviations (SDs) of the as for the AS method were smaller than those for BAS. An important feature in the BAS method is that the means and SDs for the bs were approximately the same across all levels and similar to the overall mean and SD given by the first column of Table 1. This corresponds with the original objective of the BAS method. On the other hand, for the AS method, the means of b values noticeably varied across strata.

Table 1
Item Bank Statistics

		Level 1		Level 2		Level 3		Level 4	
Statistic	Test	AS	BAS	AS	BAS	AS	BAS	AS	BAS
No. Items	360	90	90	90	90	90	90	90	90
a									
Mean	.87	.52	.60	.74	.80	.95	.94	1.28	1.15
SD	.31	.09	.17	.06	.20	.06	.23	.20	.30
Minimum	.26	.26	.26	.64	.34	.85	.38	1.07	.54
Maximum	2.00	.64	1.13	.84	1.36	1.01	1.71	2.00	2.00
b									
Mean	.14	−.39	.14	−.08	.12	.30	.15	.74	.14
SD	.99	1.19	.99	1.02	.98	.65	.99	.57	1.00
Minimum	−2.89	−2.89	−2.78	−2.47	−2.80	−2.19	−2.58	−.73	−2.89
Maximum	2.21	2.02	2.00	1.79	1.80	1.70	2.21	2.21	1.86

Maximum likelihood estimation was used to estimate θ in both methods. The estimates are $\hat{\theta}_{j,AS}$ and $\hat{\theta}_{j,BAS}$, respectively, for the AS and BAS stratification methods.

Item selection. For both methods, the initial three items were selected as described by Chang & Ying (1999). That is, the initial item was selected with $(a_1, b_1, c_1) = (1, b_0, .2)$, where b_0 was randomly selected from N(0, 1). If the first item was answered correctly, then the b parameter for the second item became $b_2 = b_1 + 2$; otherwise, $b_2 = b_1 - 2$. a_2 and c_2 remained unchanged. The remaining items were selected according to Step 5 above, but two items with the closest b values were selected first, and then one of the two was randomly selected. This procedure guaranteed randomized item selection.

Evaluation. The evaluation criteria used by Chang & Ying (1999) were used. Bias and mean squared error (MSE) were computed for AS and BAS. The bias and MSE are, respectively,

$$Bias = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \theta_j\right) , \tag{2}$$

and

$$MSE = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \theta_j\right)^2 , \tag{3}$$

where $N = 3{,}000$ simulees, and $\hat{\theta}_j$ and θ_j are the estimated and true trait levels for the jth simulee.

Chang & Ying (1999) proposed a χ^2 statistic to measure the skewness of the exposure rate distribution,

$$\chi^2 = \sum_{i=1}^{n}\frac{(er_i - L/n)^2}{L/n} , \tag{4}$$

where

er_i is the observed exposure rate for the jth item,

$L = 40$ is the test length, and

$n = 360$ is the number of items in the item bank.

Test overlap rate, which is the expected number of common items encountered by two randomly selected examinees divided by L, was also measured. (For the rationale behind this criterion, see Chang & Ying, 1999.)

Results

Table 2 summarizes the simulation results. Correlations between θ and $\hat{\theta}$ were comparable for the two methods (approximately .96). However, for examinees with true θs less than -1.95, the BAS method had a higher $\rho_{\theta,\hat{\theta}}$ (.607, compared to .321 for the AS method). These results suggest that the BAS method might improve θ estimation for examinees with extreme values. In addition, the BAS method performed better than the AS method in terms of reducing bias and MSE. The BAS method also made more efficient use of the item bank. Of the 360 items, only eight had exposure rates below 5% when the BAS method was used. When the AS method was used, there were 48 such items. The χ^2 measure for BAS was approximately one-fifth of that for AS,

$$F_{BAS,AS} = \chi^2_{BAS}/\chi^2_{AS} = .26 . \tag{5}$$

Thus, approximately 74% of the skewness in the BAS method was reduced relative to the AS method. The test overlap rates were 17.4% and 12.4% for the AS and BAS methods, respectively. Specifically, the average number of overlapping items was seven for the AS method and five for the BAS method. Following Chang & Zhang (in press), a lower bound for the overlap rate for this dataset should be 11%. The BAS method resulted in a rate quite close to that theoretical lower bound.

Figure 2 shows item exposure rates for the 360 items. For the AS method (Figure 2a), there were many items that had unacceptably high exposure rates (greater than .2). These items all came from Strata 3 and 4 (items with high as), which had low bs. The overexposure apparently was due to a lack of low b items within those strata. This can be seen from Figure 1, which shows the positive

Table 2
MSE, Bias, and Other Performance Statistics for AS and BAS

Statistic	AS	BAS
MSE	.081	.076
Bias	.012	−.002
Overlap rate	17.4%	12.7%
χ^2	22.769	5.813
Exposure rate $\leq 5\%$	48	8
$\rho_{\theta,\hat{\theta}}$	.962	.964
$\rho_{\theta,\hat{\theta}}$ for $\theta \leq -1.95$	.321	.607

correlation between a and b. Table 1 shows that the mean of the bs was .14. In Stratum 3, it was .30; in Stratum 4, mean b was .74. Because the 3,000 θs were generated from a normal distribution, a lack of low b items in Strata 3 and 4 resulted in the overexposure of low b items within the strata (Figure 2a). Such overexposure did not occur for the BAS method (Figure 2b).

Figure 3 shows scatterplots of θ and $\hat{\theta}$ for both methods. The solid line represents the 45° line. The plots indicate that the AS and BAS methods essentially were unbiased. Also, they show that the BAS method resulted in substantial improvements for examinees with θs below -1.95. This is consistent with the correlations in Table 2.

Conclusions

The AS method performs well for ideal item banks with uncorrelated a and b parameters. However, it leads to problems when a and b are correlated. In particular, the AS method can result in overexposure of certain items, as well as loss of efficiency due to the inability to match bs with θs. The BAS method provides a simple and effective solution through a two-stage stratification. The first stage can be thought of as a preemptive measure to force a balanced distribution of b values. As a result, each stratum formed at the second stage covers a wide spectrum of b values. Simulation results showed that the new method performed well in a dataset with moderately correlated as and bs. The BAS method improved item exposure rates and reduced MSE.

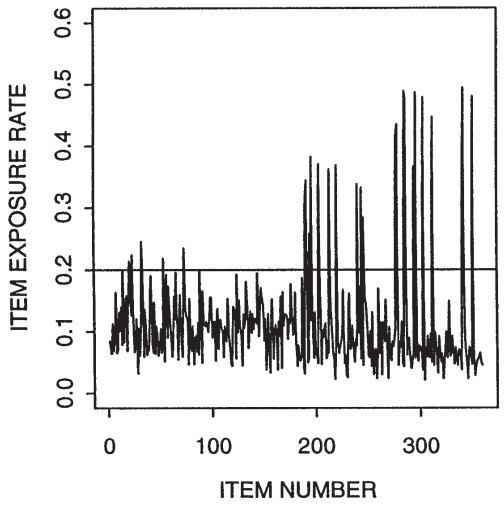
The two-stage stratification can be generalized to a multi-stage process to deal with many other practical issues in CAT designs. For example, it can be used to achieve content balance. The item bank could first be partitioned into blocks of equal size according to their content similarity, in addition to their bs. Then, within each block, items would be stratified into K levels according to as. Items of the same level then would be combined to form a stratum so that K strata will be formed.

Because most overexposure problems are caused by over-selecting certain types of items, preventing over-selection and making more efficient use of item banks could be a more direct approach to solving these problems. This general principle can be used in resolving many important issues (e.g., item exposure control, content balance, and item selection under multiple constraints) in CAT without resorting to complex solutions.

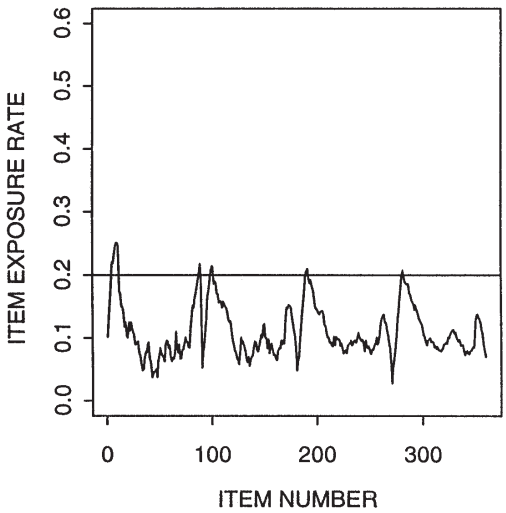
Many further improvements on the BAS method are possible. Some hybrids with other existing methods are fruitful, especially for certain types of item banks that would not satisfactorily work in the BAS method (e.g., Leung, Chang, & Hau, 2001). The BAS method incorporates Weiss's (1973) idea of stratification according to b values into the AS scheme of Chang & Ying (1999). To this end, it could be helpful to combine the proposed method with some existing methods (e.g., Chang & Ying, 1996; Eggen, 1999; Sympson & Hetter, 1985; van der Linden, 1998).

Figure 2
Item Exposure Rates for the 360 Items (Ordered by Stratum)

a. AS Method



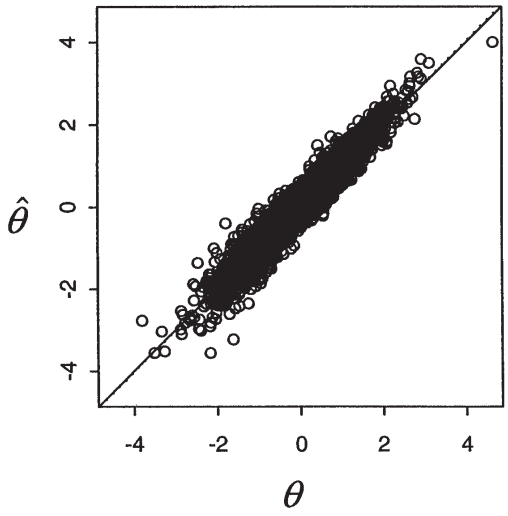
b. BAS Method



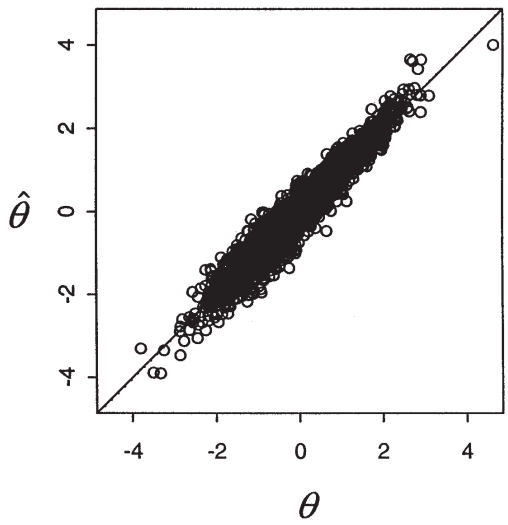
To apply the BAS method to operational item banks, many other issues need to be addressed. Some of these issues include: (1) the number of strata to be used, (2) the size of the a value range, (3) the minimum a value that is acceptable for item bank stratification, and (4) a determination of the kinds of item banks that might not work well with the method. Further research is needed and general guidelines should be developed for practitioners who would like to apply the AS method.

Figure 3
Relationship Between Estimated and True θ

a. AS Method



b. BAS Method



References

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213–229.

Chang, H.-H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23,* 211–222.

Chang, H.-H., & Zhang, J. (in press). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika.*

Eggen, T. J. H. M. (1999) Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249–261.

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2001, April). *Integrating stratification and information approaches for multiple constrained CAT.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle.

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

Lord, F. M., & Wingersky, M. S. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive verbal ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223–236). New York: Academic Press.

Parshall, C., Davey, T., & Nering, M. (1998, April). *Test development exposure control for adaptive tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego CA.

Parshall, C., Hogarty, K., & Kromrey, J. (1999, June). *Item exposure in adaptive tests: An empirical investigation of control strategies.* Paper presented at the annual meeting of the Psychometric Society, Lawrence KS.

Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (Research Report No. 95-25). Princeton NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego CA: Navy Personnel Research and Development Center.

Thomasson, G. L. (1995, June). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis MN.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22,* 195–211.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report No. 73-3). Minneapolis MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Acknowledgments

This paper was written while the first author was on the staff of the National Board of Medical Examiners.

Author's Address

Send requests for reprints or further information to Hua-Hua Chang, Department of Educational Psychology, SZB 504, University of Texas, Austin TX 78712-1296, U.S.A. Email: hua.chang@mail.utexas.edu.

Computer Program Exchange

RWEIGHT: Computing the Relative Weight of Predictors in a Regression Equation

Jeff W. Johnson, Personnel Decisions Research Institutes

It is often desirable to have a quantitative measure of the relative importance of each predictor variable in a multiple regression analysis. Relative importance is defined as the proportionate contribution each predictor makes to R^2, considering the unique contribution of it alone and when combined with other variables. The determination of relative importance is straightforward when predictor variables are uncorrelated—the squared zero-order correlations with the criterion sum to R^2 and represent the unique contribution of each variable. This simple relationship does not exist when predictors have non-zero intercorrelations, however, and zero-order correlations or standardized regression coefficients are inadequate for assessing the relative importance of predictors.

Description

Johnson (2000) presented a procedure for proportionately distributing the predictable criterion variance among intercorrelated predictors. He considered the direct effect of each predictor and its joint effect with other variables. This procedure has a distinct advantage over other procedures for determining relative importance, because it handles any number of predictor variables with equal efficiency. RWEIGHT is an SPSS syntax file for computing Johnson's (2000, p. 9, Equation 10) relative weight (ϵ).

The required input is a positive-definite correlation matrix with one dependent variable and any number of predictor variables. This can be calculated from a dataset or input directly by the user. The output includes R^2, the relative weight for each predictor, and relative weights converted to a percentage of R^2.

Availability

The program runs on any version of SPSS for Windows that includes the advanced statistics module. To receive the program files and a reprint of Johnson (2000) at no cost, write Jeff W. Johnson, Personnel Decisions Research Institutes, 43 Main Street SE, Suite 405, Minneapolis MN 55414, U.S.A.; or email jeff.johnson@personneldecisions.com.

Reference

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35,* 1–19.

Applied Psychological Measurement, Vol. 25 No. 4, December 2001, 342

Computerized Adaptive Testing With Equated Number-Correct Scoring

Wim J. van der Linden
University of Twente

A constrained computerized adaptive testing (CAT) algorithm is presented that can be used to equate CAT number-correct (NC) scores to a reference test. As a result, the CAT NC scores also are equated across administrations. The constraints are derived from van der Linden & Luecht's (1998) set of conditions on item response functions that guarantees identical observed NC score distributions on two test forms. An item bank from the Law School Admission Test was used to compare the results of the algorithm with those for equipercentile observed-score equating, as well as the prediction of NC scores on a reference test using its test response function. The effects of the constraints on the statistical properties of the θ estimator in CAT were examined. *Index terms: computerized adaptive testing, item response theory, number-correct scores, observed-score equating, optimal test assembly, 0-1 linear programming.*

Three practical situations exist in which a method for equating number-correct (NC) scores from computerized adaptive testing (CAT) would be useful:

1. To accommodate preferences among its examinees, a testing program [e.g., the Armed Services Vocational Aptitude Battery (ASVAB); Segall, 1997] might offer the choice between two different versions—CAT and paper-and-pencil—of the same test. This choice is fair only if examinees can be guaranteed comparable scores on both versions of the test.
2. To enhance the interpretation of the scores on a CAT for its examinees, a testing organization [e.g., the Scholastic Assessment Test (SAT); Lawrence & Feigenbaum, 1997] might want to equate them to the NC scores on a paper version of the test created to help examinees interpret their CAT scores.
3. Although a CAT is scored using θ estimates ($\hat{\theta}$), some examinees tend to focus on the number of items they have answered correctly. These examinees might get confused if they receive a lower score than examinees with fewer items correct than they had. CAT with NC scores of different examinees automatically equated to each other would resolve this problem.

One approach to solving the problem of score comparability between CAT and paper-and-pencil testing is to equate the $\hat{\theta}$s to the NC score scale of the paper-and-pencil test using observed-score equating. Equipercentile equating of $\hat{\theta}$s and observed scores, in combination with a randomly equivalent groups design, has been used for the ASVAB (Segall, 1997). The same method, combined with a nonequivalent group common-items design, was used by Lawrence & Feigenbaum (1997).

In the SAT, the $\hat{\theta}$s of the examinees are used to predict their scores on a released version of the SAT using its test response function (TRF). Let the released version of the test have items indexed by $j = 1, 2, \ldots, n$, with item response functions (IRFs) defined by the three-parameter logistic model (3PLM):

$$P_i(\theta) \equiv c_j + (1 - c_j)\left\{1 + \exp\left[-a_j(\theta - b_j)\right]\right\}^{-1} , \quad (1)$$

Applied Psychological Measurement, Vol. 25 No. 4, December 2001, 343–355

where

$\theta \in (-\infty, \infty)$ is a parameter for examinee trait level,
$b_j \in (-\infty, \infty)$ is the difficulty of item j,
$a_j \in (-\infty, \infty)$ is the discrimination of item j, and
$c_j \in (-\infty, \infty)$ is the guessing parameter of item j.

A prediction of the NC score on the released version of the test, Y, for this examinee can be obtained using the TRF as

$$\hat{\tau}_Y \equiv \tau_Y(\hat{\theta}) = \sum_{j=1}^{n} P_j(\hat{\theta}) \; . \tag{2}$$

The SAT uses a modification of this transformation to correct for guessing (see below). Note that Equation 2 yields an estimated "true score," not an observed score. Thus, for longer CATs, the practice of predicting scores on a released test form through its TRF can be viewed as an attempt at item response theory (IRT) true-score equating (Kolen & Brennan, 1995; Lord, 1980).

Stocking (1996; see also Chen & Thissen, 1999; Yen, 1984) dealt with the complexity of IRT-based test scoring by proposing to modify the likelihood equation so that the solution becomes a monotonic function of the examinee's NC score. Combined with the transformation through the TRF in Equation 2, this modification produces estimated NC scores on the reference test that have the same ranking as the NC scores on the CAT.

A practical disadvantage of equipercentile equating in CAT scores is the need for a separate empirical study before CAT can become operational. Such studies typically involve a large amount of resources and must be repeated each time the item bank or another feature of the CAT is changed. Potential threats to the validity of observed-score equating include the difficulty of realizing common administration conditions between two tests and the inability to deal with scores at the lower end of the scale because of guessing.

The TRF approach avoids the practical problems involved in observed-score equating. The TRF of the released test is obtained directly from the item calibration process. However, it is difficult to claim full comparability of the transformed CAT scores and the NC scores on the reference test. If a true (versus an estimated) θ value were used in Equation 2, the transformation would amount to IRT true-score equating. Substituting $\hat{\theta}$ for true θ does not result in observed-score equating. The error distributions in the NC scores in CAT are not identical to those in the $\hat{\theta}$s transformed by Equation 2.

Although the modified-likelihood approach is attractive because it does not require an expensive equating study, it has the disadvantage of introducing an estimator of θ that does not belong to any familiar class of estimators. Unlike the maximum-likelihood (ML) or the Bayesian estimators currently used in CAT, the estimator does not have known (asymptotic) properties. In fact, the nonmonotonic relationship of the estimator to ML estimators in the two-parameter logistic model and the 3PLM shows that information in the data is lost. Also, no proof exists for the consistency of this estimator.

Purpose

This paper addresses the above problems of score equating by imposing a set of constraints on item selection in CAT that automatically equates the NC scores to those on a reference test. The constraints are derived from a set of conditions on the IRFs that guarantees that the NC score distributions on two test forms will be identical (van der Linden & Luecht, 1998). To impose the item selection constraints, the method of constrained CAT with shadow tests (van der Linden, 2000a; van der Linden & Reese, 1998) is used.

Because the algorithm selects the items to automatically have the same NC score distribution as that of a reference test, no additional score transformation is needed. On the other hand, the method can only be used for a fixed-length CAT and a reference test that has the same length. Further, as in any other IRT-based method, the method relies on the assumption that the item bank fits the item response model.

"Reference test" is used here to refer to the test to which CAT is equated. Three different types of reference tests are distinguished:

1. A paper-and-pencil version of the CAT with exactly the same specifications that the examinee can select as an alternative to the CAT.
2. A paper version of the test that is created to help examinees interpret their CAT scores.
3. A dummy test with a conveniently selected set of IRFs.

The first two types have already been discussed. In either case, because the NC scores of examinees are equated to the same reference test, they also are equated mutually. Thus, for these tests it is not possible for an examinee to answer more items correctly on a CAT but receive a lower score than an examinee with fewer items correct.

The same principle can be exploited if equating across CAT administrations is needed, but no external reference test is available. The proposed algorithm then can be used with a subset of items from the bank as a dummy reference test. In fact, because only the IRFs in the reference test matter (see Equation 3 below), it is possible to select a set of item parameter values for the dummy test that do not belong to actual items, but just represent a useful target.

A CAT Algorithm With Equated NC Scores

Let X be the NC score on a test with items $i = 1, 2, \ldots, n$, and let Y be the NC score on another test with items $j = 1, 2, \ldots, n$. Van der Linden & Luecht (1998) proved that, for any common distribution $h(\theta)$, the distributions of X and Y are identical if and only if

$$\sum_{i=1}^{n} P_i^r(\theta) = \sum_{j=1}^{n} P_j^r(\theta) , \quad -\infty < \theta < \infty , \tag{3}$$

for $r = 1, 2, \ldots, n$.

These constraints require the sums of the first through nth powers of the IRFs in the two tests to be equal. However, van der Linden & Luecht (1998) showed that, for $n \to \infty$, the constraints for $r \geq 2$ become negligible. Their empirical examples indicated that use of the first 2–3 constraints gives excellent approximations for realistic test lengths. Note that when $r = 1$ in Equation 3, the true scores on the two test are equated. For other features of these constraints, see van der Linden (2000b).

Applications to CAT

For linear tests (i.e., conventional tests in which all examinees receive the same items in the same order), the constraints in Equation 3 would require the sums of the powers of the IRFs to be identical over the full range of θ. However, for a situation in which a CAT is to be equated to a linear reference test, each examinee answers an individual item set from the bank and the sums in Equation 3 need to be identical only for the examinee's true θ. Under certain conditions, the $\hat{\theta}$s in CAT converge to this true value (see below). Because each examinee's (conditional) NC score distribution is equated, the marginal distributions of the NC scores on the two tests are equated for any population of examinees.

To create identical observed-score distributions for the examinees, the constraints in Equation 3 are imposed on the item selection procedure in CAT for small values of r. Implementation of this

idea is possible through the method of constrained CAT with shadow tests (van der Linden, 2000a; van der Linden & Reese, 1998).

At each step in this method, items are not selected directly from the bank, but from a full (shadow) test assembled from the bank. The shadow test for the administration of the kth item must meet the following specifications:

1. Maximum information at the current $\hat{\theta}$.
2. Length equal to the number of items in the (fixed-length) CAT.
3. Meet all constraints imposed on the CAT.
4. Contain the $k-1$ items already administered.

From the unused items in the shadow test, the item with maximum information at the current $\hat{\theta}$ is administered as the kth item in the CAT. The procedure is repeated until the CAT is complete.

Because each shadow test must meet all constraints, so must the CAT. Also, because both the shadow tests and the individual items are selected to have maximum information, the CAT tends to be maximally informative. For further technical details on constrained CAT with shadow tests, see van der Linden (2000a).

A Model for Selection of Shadow Tests

Because the conditions in Equation 3 are linear in the items, shadow tests can be selected using 0-1 linear programming (LP; for an introduction, see van der Linden, 1998a). Let there be $i = 1, 2, \ldots, I$ items in the CAT bank. Let $k = 1, 2, \ldots, n$ be the items in the CAT. Thus, i_k is the index of the item in the bank administered as the kth item in the CAT. The set indices of the first $k-1$ items in the CAT is thus $S_{k-1} \equiv (i_1, , , ., i_{k-1})$. The items in the reference test are denoted $j = 1, 2, \ldots, n$. $\hat{\theta}_{k-1}$ is the estimated value of θ after $k-1$ items have been administered.

To formulate the model, binary variables x_i are used to denote whether item i is selected in the shadow test. If it is, $x_i = 1$; if it is not, $x_i = 0$. For the selection of the kth shadow test, maximize

$$\sum_{i=1}^{I} I_i(\hat{\theta}_{k-1})x_i \ , \tag{4}$$

subject to

$$\sum_{i=1}^{I} P_i^r(\hat{\theta}_{k-1})x_i - \sum_{j=1}^{n} P_j^r(\hat{\theta}_{k-1}) \leq \delta \ , \quad r = 1, 2, \ldots, R \ , \tag{5}$$

$$\sum_{i=1}^{I} P_i^r(\hat{\theta}_{k-1})x_i - \sum_{j=1}^{n} P_j^r(\hat{\theta}_{k-1}) \geq -\delta \ , \quad r = 1, 2, \ldots, R \ , \tag{6}$$

$$\sum_{i=1}^{I} x_i = n \ , \tag{7}$$

$$\sum_{i \in S_{k-1}} x_i = k-1 \ , \tag{8}$$

and

$$x_i \in \{0, 1\} , \quad i = 1, 2, \ldots , I . \tag{9}$$

The objective function in Equation 4 maximizes the information in the shadow test at $\hat{\theta}_{k-1}$. Because of the constraints in Equations 5 and 6, the differences between the sums of the first R powers of the reference items' probabilities of success and those of the shadow test items must be identical up to a small constant δ (selected by the CAT administrator). If necessary, δ can be selected dependent on r. The test length is set equal to n by the constraint in Equation 7, whereas the constraint in Equation 8 forces the previous $k-1$ items to be in the shadow test for the kth item. The constraints in Equation 9 define the range of the decision variables. Other constraints can be added, for example to make the CAT administrations content balanced (for a review, see van der Linden, 1998a). An application of the method to the problem of controlling for differential speededness in CAT is given in van der Linden, Scrams, & Schnipke (1999).

The tolerance factor δ in Equations 5 and 6 is introduced for technical reasons only; imposing an exact equality would generally lead to an infeasible test assembly problem. Asymptotic consequences of using $\hat{\theta}$ in Equations 5 and 6 rather than θ are discussed below.

Models such as Equations 4–9 can be solved for optimal values of their decision variables using one of the algorithms or heuristics available in general LP software (e.g., CPLEX; ILOG, 1998) or a test assembly package (e.g., CONTEST; Timminga, van der Linden, & Schweizer, 1996).

Extensions and Special Cases

Because identity of distributions is maintained under identical transformation of their variables, the conditions in Equation 3 guarantee the equating of any monotonic function of NC scores. A transformation often used with multiple-choice items is formula scoring to correct for guessing,

$$\frac{AX - n}{A - 1} , \tag{10}$$

where X is the NC score on the test, and A is the (common) number of alternatives per item (Lord & Novick, 1968, Eq. 14.3.4). Because the relationship in Equation 10 is linear in X, formula scores automatically are equated under the conditions in Equation 3.

An alternative to the transformation through the TRF in Equation 2 is Lord's (1980, Eq. 15.6) true formula score; θ is replaced by its estimated value,

$$\frac{A\sum_{j=1}^{n} P_j(\hat{\theta}) - n}{A - 1} . \tag{11}$$

Lawrence & Feigenbaum (1997) used this transformation as an analogue to the transformation in Equation 2.

Note that the use of true θ in Equation 11 results in equated true formula scores. However, for $\hat{\theta}$, the transformation equates neither true nor observed formula scores. Nevertheless, true formula score equating is possible by selecting a shadow test such that the formula score in Equation 11 is equated to that of the reference test. Requiring these formula scores to be equal up to a tolerance factor δ gives the following constraints:

$$\sum_{i=1}^{I} P_i(\hat{\theta}_{k-1})x_i - \sum_{j=1}^{n} P_j(\hat{\theta}_{k-1}) \leq (A - 1)\delta/A , \tag{12}$$

and

$$\sum_{i=1}^{I} P_i(\hat{\theta}_{k-1})x_i - \sum_{j=1}^{n} P_j(\hat{\theta}_{k-1}) \geq -(A-1)\delta/A \ . \tag{13}$$

The constraints in Equations 12 and 13 are identical to those in Equations 5 and 6, except for a rescaling of the tolerance factor δ. If exact equality were required (i.e., $\delta = 0$), true score equating and true formula score equating would yield the same results.

Discussion

The constraints in Equations 5 and 6 equate sums of powers of success probabilities, not powers of individual probabilities. In fact, as follows from Proposition 4 in van der Linden & Luecht (1998), equating the individual response probabilities in CAT to those in the reference test is implied if all n conditions in Equation 3 are imposed. Because CATs typically have 25–30 items and only the sums of the first 2–3 powers need to be equated, compensation can occur across items of the terms in these sums. Also, only sums of powers of probabilities for individual examinees are equated; these constraints do not require the items to have identical IRFs across CAT administrations.

The conditions in Equation 3 are formulated for true θ. However, they are implemented for the current $\hat{\theta}$ in the constraints in Equations 4 and 5. As Chang & Ying (in press) showed for the one-parameter logistic model and an infinitely large item bank, the ML estimator of θ in CAT with maximum information item selection is strongly consistent. Consistency holds for the two-parameter logistic model, provided that realistic bounds on the discrimination parameters are met. For the 3PLM, the same results hold, provided an additional realistic bound on the guessing parameter is met and the likelihood equations do not have multiple solutions. Because the conditions in Equation 3 are based on continuous functions of θ, it follows that the differences in the left-hand sides of Equations 5 and 6 also converge to their true θ equivalents (e.g., Lehmann, 1999, theorem 2.1.4). These results are expected to be closely approximated for CATs from well-designed finite item banks.

The presence of the constraints in Equations 5 and 6 in the CAT algorithm results in a reduction of the effective size of the item bank. However, because the reduction is for the most-informative subset of items, the effect of these constraints can be expected to be only a slightly slower rate of convergence. Convergence can be accelerated using good initial θs (see the empirical example below).

Thus, in a typical CAT session, the first items, selected at $\hat{\theta}$s that likely will not correspond to the true θ, cumulate partial sums of powers of success probabilities at true θ that do not match those of the reference test. When $\hat{\theta}$ converges to its true value, contributions by items later in the process tend to compensate for earlier contributions, and the differences between the sums in Equations 5 and 6 converge to their true equivalents. Because the constraints in Equations 5 and 6 are imposed at different values of $\hat{\theta}$ at each step, the process in no way requires the CAT to select sets of items with identical IRFs across administrations.

Empirical Example

Method

The proposed CAT algorithm was studied for an item bank from the Law School Admissions Test, consisting of 753 items calibrated using the 3PLM in Equation 1. The bank was assembled from previously administered paper-and-pencil versions of the test. An arbitrary form was identified

from which a nested set of reference tests ($n = 10, 20, \ldots, 50$) was randomly selected, to which the CATs had to be equated. Target NC score distributions on the reference tests were generated using the algorithm for the generalized binomial distribution (Lord & Wingersky, 1984). For a description of the bank, see van der Linden & Reese (1998).

The following conditions were simulated, all with $n = 10, 20, \ldots, 50$ items:

1. Unconstrained CAT (UCAT).
2. UCAT with true NC scores on the reference test estimated through its TRF (Equation 2).
3. Constrained CAT with shadow tests selected for $R = 1, 2, 3, 4$ (Equations 4–9).

Condition 1 was used only to assess the impact of the constraints on the NC score distribution.

Test length was varied to examine the speed of convergence of the observed-score distributions to the target. For each condition, the following data were collected:

1. Observed NC scores (estimated true NC scores in Condition 2).
2. Estimated bias in $\hat{\theta}$.
3. Estimated mean squared error (MSE) in $\hat{\theta}$.

The θs of the simulees were randomly drawn from N(0, 1). 30,000 values were selected for each condition. CAT was simulated using the procedure for Bayesian initialization of the θ estimator (van der Linden, 1999). In this procedure, the initial $\hat{\theta}$ is the regressed value of θ on one or more background variables, Z. The bivariate distribution of θ and Z was assumed to be standard normal, with $\rho_{\theta Z} = .60$ [approximately the same correlation as found in van der Linden's (1999) empirical example]. The first shadow test was assembled to gave maximum Fisher information at the regressed value of θ on the value for the background variables drawn for the examinee. Next, θ estimates were obtained using the expected a posteriori (EAP) estimator. This estimator is known to perform generally well with a smaller MSE than the ML estimator and a slight inward bias (van der Linden, 1998b). Also, unlike the ML estimator, it always exists. The procedure was terminated after n items were selected.

Implementation of the Algorithm

Trial runs with the algorithm showed an occasional case of infeasibility when δ was too small. In such cases, the first items administered appeared to be highly informative at off-target $\hat{\theta}$s. This resulted in the sum of the IRFs for the first part of the CAT becoming steep at incorrect θs. It therefore became difficult for the full CAT to meet the constraints in Equations 5 and 6 closer to true θ for the same small value of δ.

To deal with such cases, the algorithm was implemented as follows:

1. The CAT IRFs were constrained to satisfy Equations 5 and 6—not only at the current $\hat{\theta}$, but also at slightly lower and slightly higher values ($\hat{\theta} - .5$ and $\hat{\theta} + .5$). As a consequence, the algorithm behaved more robustly with respect to estimates that were off target; the solution still met the constraints in the original problem.
2. The algorithm was started with a small δ value. When a case of infeasibility was met, the algorithm tested whether the infeasibility was caused by the additional constraints. If so, they were removed from the model. If not, the value of δ was increased slightly. The algorithm started with $\delta = .5$ and the increase was set at .2.

All simulations were run on a personal computer with a 166 Mhz Pentium processor. The LP models for the shadow tests were solved using CPLEX 6.0 (ILOG, 1998). Solutions to 0-1 LP models for test assembly were obtained iteratively. A shadow test optimal at $\theta = 0$ calculated prior to the simulations was used as the initial solution. The CPU time needed to calculate a shadow test and select an item was 6–8 seconds per item.

Results

Figure 1 shows the NC score distributions for the UCAT, constrained CAT, and their target distributions. The distributions for UCAT showed their typical peaked form at an NC score slightly higher than $n/2$. The target distributions on the reference tests were much wider. For $n = 10$ (Figure 1a), all distributions for the constrained CAT were between those for UCAT and the target distribution, but much closer to the latter. For increasing test length, the target distribution was more closely approximated. For $n = 30$ (Figure 1c), the approximation was very close. For $n = 50$ (Figure 1e), there were no systematic differences between the distributions and the target. The value of R did not appear to have much impact.

Figure 1
Distributions of Observed NC Scores for UCAT (Bold Solid Line), Reference Test (Bold Dashed Line), and Constrained CAT ($R = 1, 2, 3, 4$; Other Lines)

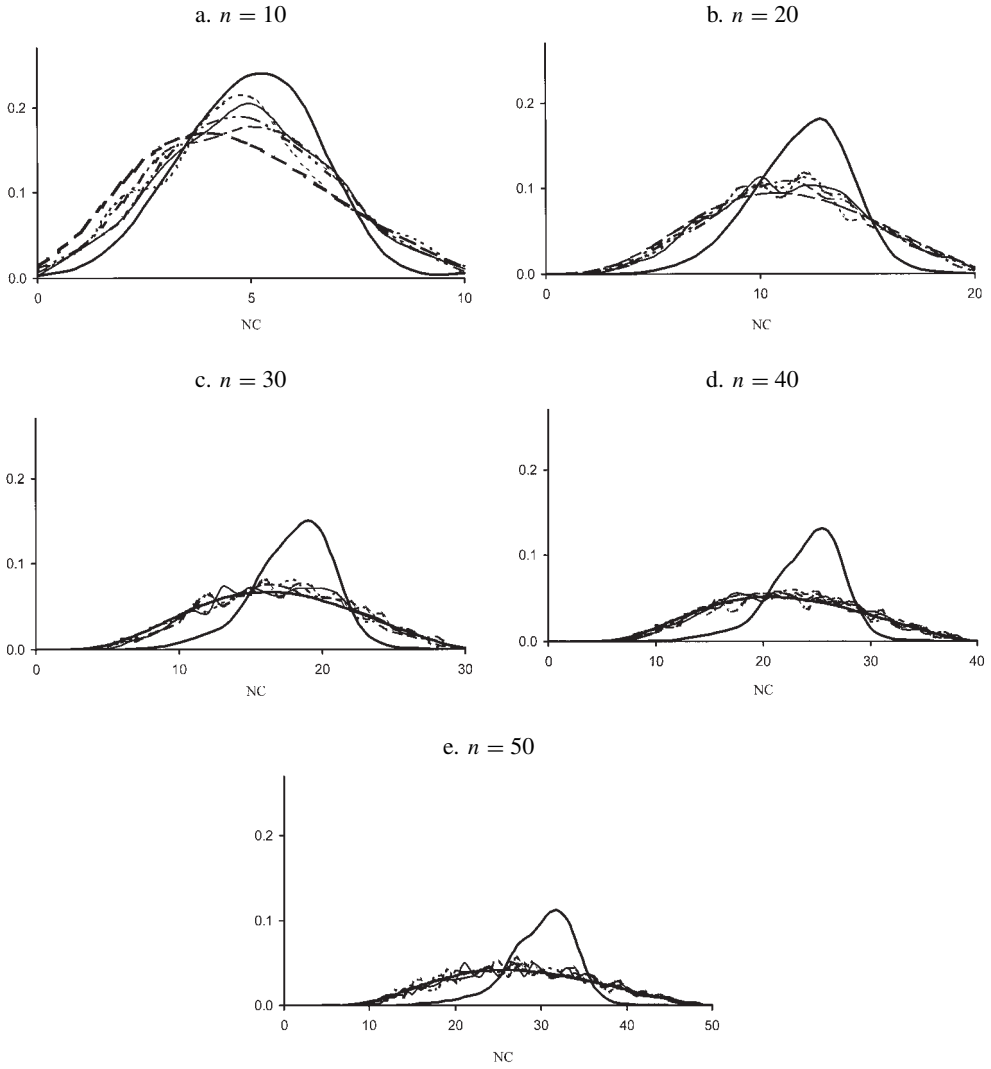


Figure 2 shows the distributions of the true NC scores on the reference test estimated through the TRF (Condition 2). Also shown are the observed-score distributions for UCAT and the target distributions on the reference tests. For all test lengths, the distributions of the estimated true NC scores were more peaked than the target distributions. However, the approximation improved considerably with increasing test length because of the convergence of observed to true NC scores. Further, a distortion of the lower tail of the estimated true-score distributions was observed. This resulted because of the lower asymptote of the TRF introduced by the guessing parameter in the 3PLM, which made it difficult to estimate true scores below the expected guessing level.

Figure 2
Distributions of Observed NC Scores for UCAT (Bold Solid Line) and Reference Test (Bold Dashed Line), With Distributions of True NC Scores Estimated Through TRF (Dotted Line)

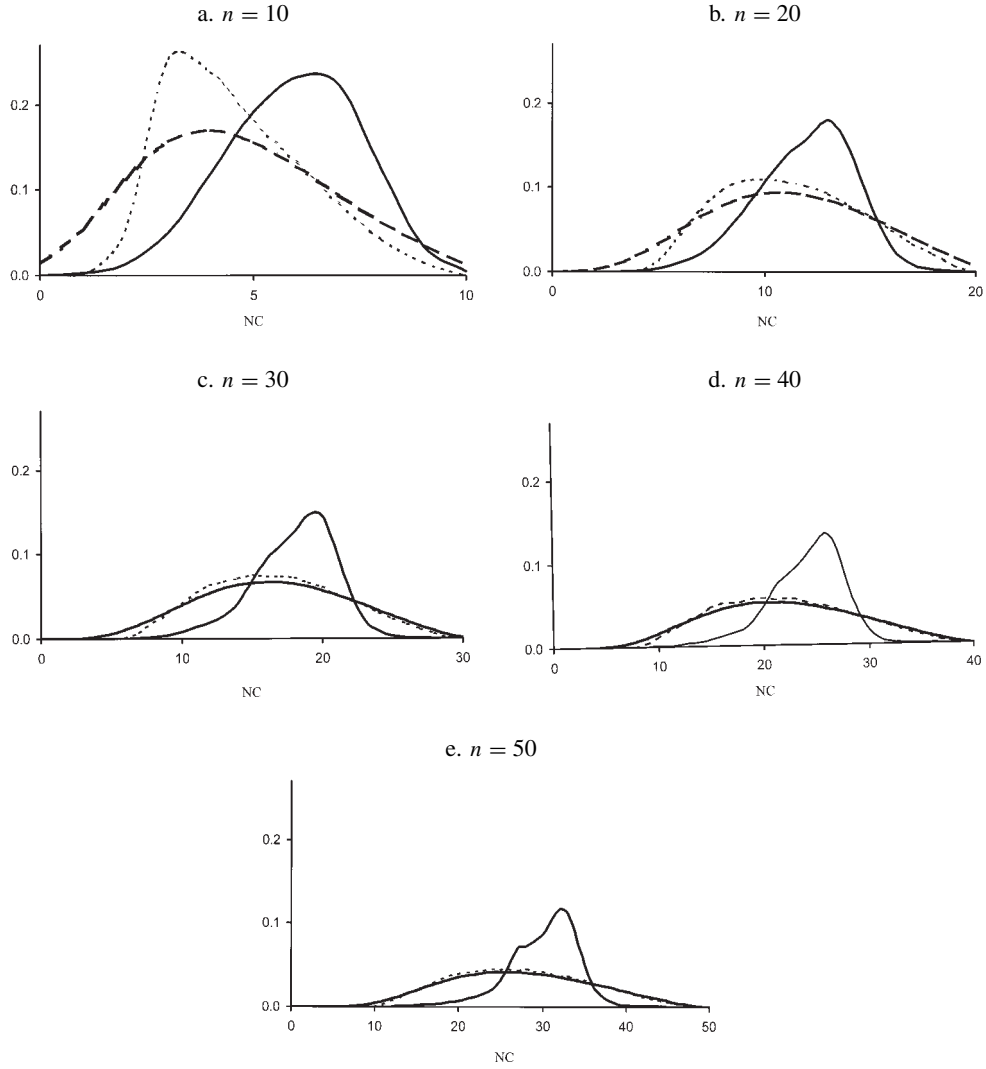


Figure 3 shows estimated bias in the θ estimator for UCAT and the constrained CATs as a function of θ (Condition 3). The general shape of these plots shows the well-known inward bias for the EAP estimator. Except for $n = 10$ (Figure 3a), the presence of the constraints in Equations 5 and 6 did not introduce any additional bias in the estimator. For $n = 10$, an increased bias at the lower end of the θ scale and more variation for the values of R at the upper end of the scale were found.

In Figure 4, for the same conditions, the estimated MSE in the θ estimators are plotted as a function of θ. For all test lengths, the MSE for constrained CAT was systematically larger than for UCAT. The largest loss of efficiency occurred for $n = 10$ (Figure 4a) at lower θs. The worst cases

Figure 3
Bias Functions for UCAT (Bold Solid Line) and Constrained CAT ($R = 1, 2, 3, 4$; Other Lines)

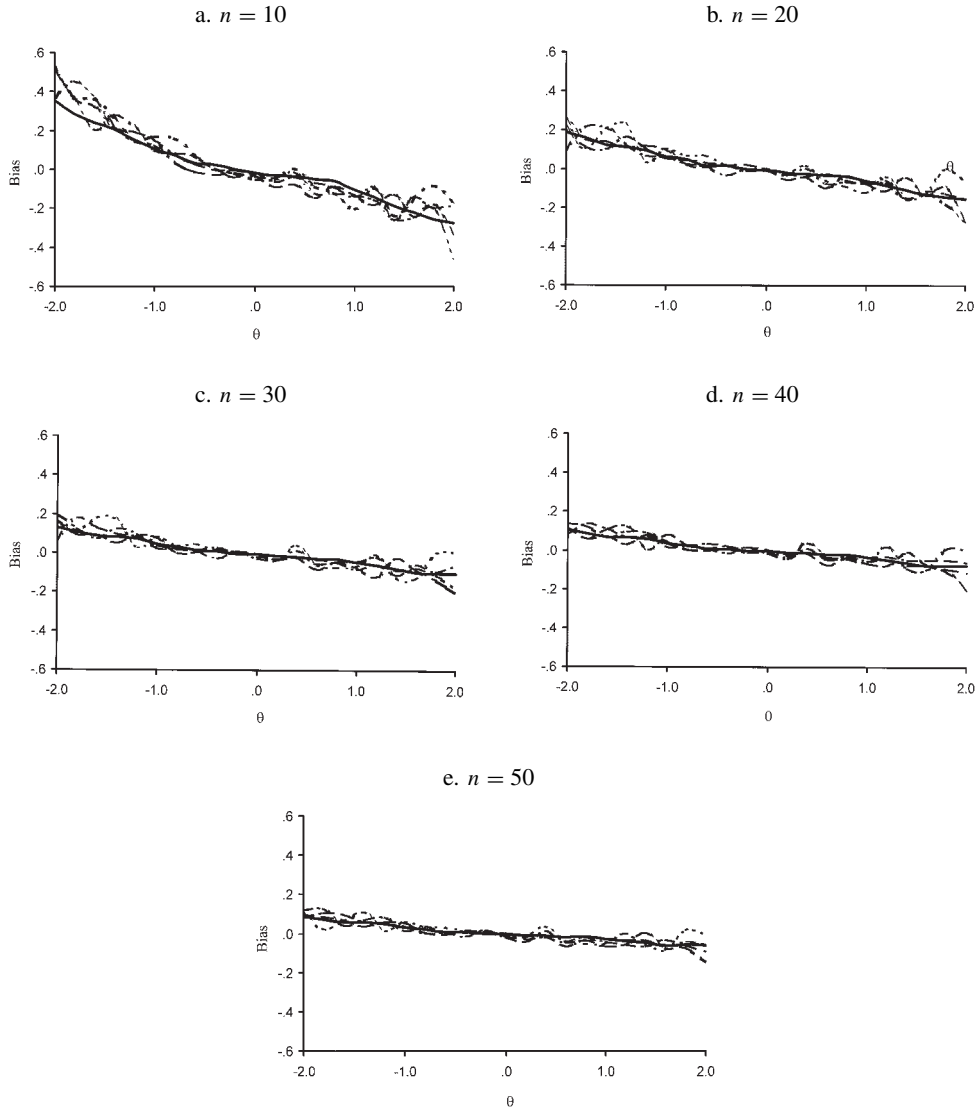
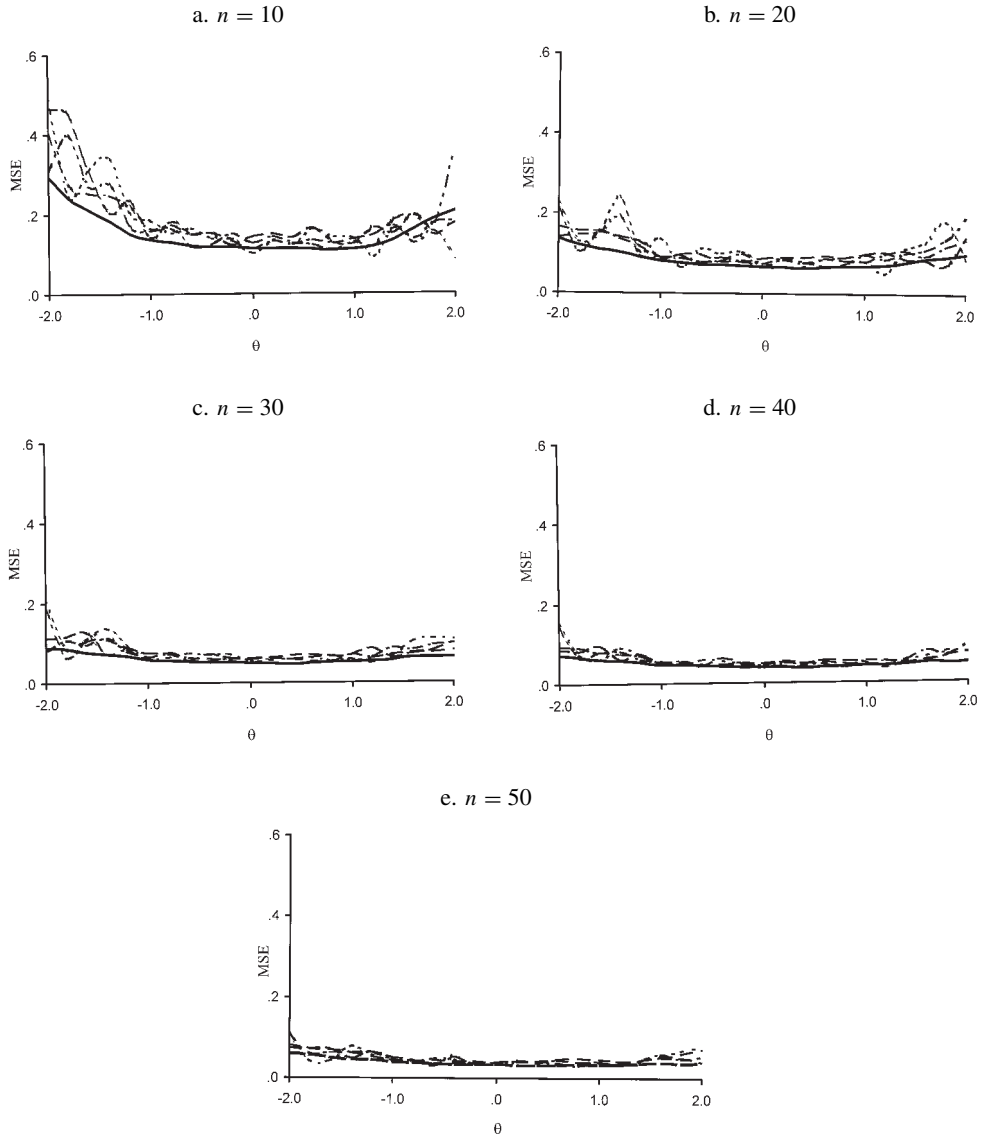


Figure 4
MSE Functions for UCAT (Bold Solid Line) and
Constrained CAT ($R = 1, 2, 3, 4$; Other Lines)



occurred for test lengths shorter than the typical CATs used in practice (25–35 items) and for the most extreme values of θ with few examinees. More importantly, however, the differences were very small in terms of the impact on the standard errors. For example, when MSE = .20 (least efficient) and MSE = .15 (most efficient), the corresponding standard errors were .45 and .39, respectively.

Conclusions

In the empirical example, the constrained CAT algorithm did not show any unexpected behavior. The CPU times of 6–8 seconds per item are small enough for practical applications in which the

computer selects two alternative items for item k (one based on an incorrect answer and the other based on a correct answer) to be presented to the examinee after each update of $\hat{\theta}$, while the examinee works on item $k - 1$. However, with faster computers becoming more common, CPU times will become shorter (e.g., Veldkamp, 2001).

For longer tests, the algorithm produced an NC score distribution that did not differ systematically from the target on the reference test. Also, it did not introduce any systematic bias in the θ estimator. However, the estimator did lose some of the efficiency associated with UCAT. For shorter tests, the empirical example yielded NC score distributions that were more peaked than the distributions on the reference test. In particular, for $n = 10$, additional equating seemed necessary. However, for such cases, use of the algorithm is still recommended, because it minimizes the distortion of the NC scale involved in additional equating.

For all test lengths, the constrained CAT algorithm clearly outperformed the TRF transformation. The algorithm realized this improvement for the actual numbers of items answered correctly by the simulees, rather than a post hoc transformation of their $\hat{\theta}$s with an indirect relation to the response vectors.

More importantly, the algorithm meets Lord's (1980) criterion of second-order equity. Because it equates the full (conditional) distributions of the NC scores on the CAT and the reference test for each examinee, the examinees have identical error distributions as well as comparable scores (van der Linden, 2000b). Predicted scores on a reference test based on estimates of θ with unequal error variances do not have this property.

References

Chang, H., & Ying, Z. (in press). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Annals of Statistics.*

Chen, W.-H., & Thissen, D. (1999). Estimation of item parameters for the three-parameter logistic model using the marginal likelihood of summed scores. *British Journal of Mathematical and Statistical Psychology, 52,* 19–37.

ILOG (1998). *CPLEX 6.0 user's manual.* Incline Village NV: Author.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Lawrence, I., & Feigenbaum, M. (1997). *Linking scores for computer-adaptive and paper-and-pencil administrations of the SAT* (Research Report No. 97-12). Princeton NJ: Educational Testing Service.

Lehmann, E. L. (1999). *Elements of large-sample theory.* New York: Springer-Verlag.

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

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 452–461.

Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 181–198). Washington DC: American Psychological Association.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21,* 365–389.

Timminga, E., van der Linden, W. J., & Schweizer, D. A. (1996). *ConTEST 2.0: A decision support system for item banking and optimal test assembly* [Computer program and manual]. Groningen, The Netherlands: ProGAMMA.

van der Linden, W. J. (1998a). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22,* 195–211.

van der Linden, W. J. (1998b). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63,* 201–216.

van der Linden, W. J. (1999). A procedure for empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement, 23,* 21–29.

van der Linden, W. J. (2000a). Constrained adaptive testing with shadow tests. In W. J. van der Linden

& C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Boston: Kluwer-Nijhoff.

van der Linden, W. J. (2000b). A test-theoretic approach to observed-score equating. *Psychometrika, 65,* 437–456.

van der Linden, W. J., & Luecht, R. M. (1998). Observed-score equating as a test assembly problem. *Psychometrika, 62,* 401–418.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22,* 259–270.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement, 23,* 195–210.

Veldkamp, B. P. (2001). *Principles and methods of constrained test assembly* (Chapter 4). Doctoral dissertation, University of Twente, The Netherlands.

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

Acknowledgments

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of LSAC. The author thanks Wim M. M. Tielen for his computational assistance.

Author's Address

Send requests for reprints or further information to Wim J. van der Linden, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: vanderlinden@edte.utwente.nl.

Computer Program Exchange

Read_FastTestPro_Log: Extraction of Examinee Data From FastTEST Pro Examinee Files

Christine De Mars, James Madison University

Description

Read_FastTestPro_Log is a SAS macro that reads the log files from the FastTEST Professional Testing System (Assessment Systems Corporation, 1999) and organizes the data into a form that can be analyzed easily. FastTEST Pro is commercial software designed to implement computerized adaptive testing (CAT). The test output in FastTEST Pro is in the form of a log file for each examinee, which has a record of every action he/she took (e.g., moving between items, selecting an answer, selecting a different answer, scrolling within the item, starting and stopping video and audio clips) during a test. The result is a very detailed record that could be used to study a multitude of research questions about CAT. However, a test administrator setting up a simple test or survey might find it difficult to work with this information to obtain a quick list of how each examinee answered each item.

Read_FastTestPro_Log reads the log files and outputs a single file for the test, with three records for each examinee. The first record for each examinee includes: his/her name, identification number, test date, and total score. The final answer chosen then is listed for each item. The second record contains a score for each item. The third record includes the ordinal position in which each item was administered. If items were administered randomly or adaptively, they are re-ordered in a sequence specified by the user, so that a given column in the output contains data for the same item for all examinees (with a missing value if an item was not administered to an examinee). This information then can be imported into data analysis software.

There are many items of information the macro does not collect, such as answer changes and time spent on each item. Because the program is source code (not compiled), users familiar with SAS can modify it to collect other information from the log files (or to output less information).

Read_FastTestPro_Log was intended primarily for FastTEST Pro's structured-response items (e.g., multiple-choice, true-false, survey). It works with matching items under restricted conditions (i.e., treating each element in the left list as an item and the element selected from the right list as a response).

Availability

Read_FastTestPro_Log runs under Windows with the base SAS modules installed. The macro, documentation, and sample files can be downloaded from http://www.jmu.edu/assessment/readlog/readlog.htm. They also can be obtained by sending an email to demarsce@jmu.edu.

Reference

Assessment Systems Corporation. (1999). *The FastTEST Professional Testing System* [Computer program]. St. Paul MN: Author.

Comparison of Dichotomous and Polytomous Item Response Models in Equating Scores From Tests Composed of Testlets

Guemin Lee, CTB/McGraw-Hill

Michael J. Kolen, David A. Frisbie, and Robert D. Ankenmann
University of Iowa

The performance of two polytomous item response theory models was compared to that of the dichotomous three-parameter logistic model in the context of equating tests composed of testlets. For the polytomous models, testlet scores were used to eliminate the effect of the dependence among within-testlet items. Traditional equating methods were used as criteria for both. The equating methods based on polytomous models were found to produce results that more closely agreed with the results of traditional methods. *Index terms: dichotomous item response theory, equating, item response theory assumption, polytomous item response theory, testlet.*

When item response models are applied in test-score equating situations, strong statistical assumptions—unidimensionality and local item independence (LI)—must be made. Because unidimensional dichotomous item response models frequently are used for equating, it is important to study the robustness of these models to violations of the assumptions and to investigate model-data fit (Kolen & Brennan, 1995).

This study was concerned with the application of item response theory (IRT) equating procedures to tests composed of testlets (i.e., tests that are small enough to manipulate, but large enough to carry their own context; Wainer & Kiely, 1987; Wainer & Lewis, 1990). Reading comprehension tests, containing sets of passages with a set of items pertaining to each passage, are examples of tests composed of testlets. Previous studies dealing with test scores obtained from testlet-based tests have indicated that LI is likely to be violated, making it difficult to satisfy unidimensionality, which is required by IRT. That is, when several items in a test are related to a common passage or other common stimulus material, conditional dependence is present among those items (Lee, Dunbar, & Frisbie, 1999; Lee & Frisbie, 1999; Wainer & Thissen, 1996; Yen, 1993). In this situation, the application of dichotomous IRT (DIRT) models to equate testlet-based tests might cause problems. Because there is little evidence in the literature about how the violation of IRT assumptions affects equating relationships involving testlets, it is not clear how serious the degree of distortion of equated scores might be.

When testlets are used, testlet scores can be used instead of item scores to eliminate the influence of dependence among within-testlet items (Lee, 2000; Lee & Frisbie, 1999; Sireci, Thissen, & Wainer, 1991; Wainer & Thissen, 1996). Polytomous IRT (PIRT) models might be considered as alternatives to DIRT models if the problem is serious.

Applied Psychological Measurement, Vol. 25 No. 4, December 2001, 357–372

The objectives of this study were to:

1. Assess the local item dependence and dimensionality of testlet-based tests to determine the appropriateness of DIRT and PIRT models in the context of equating these test forms.
2. Compare equating results from traditional equating methods with those from the dichotomous three-parameter logistic model and several PIRT models to investigate the implications of using them for equating testlet-based tests.
3. Investigate the generalizability of equating results for various types of testlet-based tests, using three tests from the Iowa Tests of Basic Skills (ITBS): Reading Comprehension, Maps and Diagrams, and Math Problem Solving and Data Interpretation (Hoover, Hieronymous, Frisbie, & Dunbar, 1994).

IRT Equating With Testlets

Three traditional equating methods—mean, linear, and equipercentile—were applied to the datasets. Because they use total test scores, these equating methods are not influenced by the violation of LI. This made them reasonable baseline methods, because DIRT and PIRT models require LI.

In mean equating, the scores on a form are adjusted by adding or subtracting a constant (i.e., the difference between the mean scores of two forms) so that the adjusted scores have the same mean as another form. Linear equating allows equated scores to have the same standard deviation (SD) and mean as the original scores. Equipercentile equating can be developed by identifying a score transformation that makes the equated score from one form have the same percentile rank as a corresponding score from another form (for more detailed explanations of traditional equating methods, see Kolen, 1988; Kolen & Brennan, 1995).

IRT linking/equating refers to the process of placing item parameter estimates from two test forms onto a common scale. Once this is done, there is no further need to develop an equating relationship between trait (θ) estimates for examinees from two forms. However, as Kolen & Brennan (1995) indicated, there are several practical problems in using IRT θ estimates: (1) examinees with the same number-correct (NC) score often receive different θ estimates; (2) IRT θ estimates are difficult to compute and cannot be obtained by hand; and (3) for high and low θ examinees, relatively greater amounts of measurement error can be anticipated. For these practical reasons, tests often are scored with NC scores that must be equated from two forms.

In IRT "true-score" equating, the NC "true score" associated with a particular θ on one form is considered to be equivalent to that on another form. In IRT observed-score equating, the IRT model is used to estimate distributions of observed NC scores on two forms for a population of examinees. These estimated observed-score distributions then are equated using equipercentile methods. [Theoretical explanations of DIRT equating methods are presented in Cook & Eignor (1991) and Kolen & Brennan (1995).]

Two PIRT models were used to equate testlet-based tests. PIRT models introduced during the last three decades include the graded response model (GRM; Samejima, 1969), the rating scale model (Andrich, 1978), the partial-credit model (Masters; 1982), and the nominal model (NM; Bock, 1972). With respect to testlet applications, the NM has been used most often (Sireci et al., 1991; Wainer, 1995; Wainer, Sireci, & Thissen, 1991; Wainer & Thissen, 1996) because "the testlet scores are nominal (or at most semi-ordered) responses. . . [because] a score of 1 may not always reflect higher proficiency than a score of 0, due to guessing" (Thissen, Steinberg, & Mooney, 1989, p. 259). Although the GRM is based on ordered response categories, its use in testlet-based equating applications might be appropriate. There would be an ordered quality to testlet-based scores if such scores corresponded to the extent of completeness of the examinee's reasoning process within a

specific testlet. This a priori rationale is reasonable for reading comprehension testlets, in which several dichotomously scored items relate to a single reading passage. The more of these items within a testlet that an examinee answers correctly, the more extensive his/her reasoning process is. Therefore, the GRM and NM were studied and compared in the present study.

Equating With the NM

To apply PIRT models, testlet scores are obtained by summing the dichotomous scores of the items that comprise the testlet. Let testlet j contain n_j items. Then, the polytomous testlet score would be an integer between 0 and n_j. In other words, a testlet consisting of n_j dichotomous items can be reconceptualized and treated as a single polytomous item having $n_j + 1$ response categories. Each of the response categories $(1, 2, \ldots, n_j + 1)$ corresponds to one of the polytomous testlet scores.

Under the NM, the probability of an examinee with a given θ responding to category k in testlet j is

$$P_{jk}(\theta) = \frac{\exp[a_{jk}\theta + c_{jk}]}{\sum_{k=1}^{K} \exp[a_{jk}\theta + c_{jk}]} , \tag{1}$$

where

$j = 1, 2, \ldots, J$ is the number of testlets;
$k = 1, 2, \ldots, K$ is the number of categories;
a_{jk} and c_{jk} are the discrimination and intercept parameters, respectively, associated with the kth category of testlet j, that identify the shape of the testlet category response functions; and
$\sum_k a_{jk} = \sum_k c_{jk} = 0$ are constraints imposed on the model.

The parameters are rescaled using centered polynomials of the associated scores to represent the category-to-category changes in a_k and c_k:

$$a_{jk} = \sum_{p=1}^{P} \alpha_{jp} \left(k - \frac{K}{2}\right)^p , \tag{2}$$

and

$$c_{jk} = \sum_{p=1}^{P} \gamma_{jp} \left(k - \frac{K}{2}\right)^p , \tag{3}$$

where $(\alpha_p, \gamma_p)_j (p = 1, 2, \ldots, P$, for $p \leq K)$ are the free parameters to be estimated from the data (Thissen et al., 1989).

True-score equating. The "true score" is defined by

$$T(\theta) = \sum_{j=1}^{J} \sum_{k=1}^{K} u_{jk} P_{jk}(\theta) , \tag{4}$$

where u_{jk} is a weight allocated to response category k of testlet j. The IRT true-score (TS) equating method outlined by Cook & Eignor (1991) and Kolen & Brennan (1995) was applied to related NC scores from two forms. That is, the true score on one form associated with a given θ is considered to correspond to the true score on another form associated with the same θ.

Observed-score equating. After item and θ parameters have been estimated, IRT observed-score (OS) equating can be conducted using the following steps:

1. Estimate the NC score distribution for each of two forms using item parameter estimates and estimated θ distributions.
2. Use equipercentile methods to equate scores from two forms so that the percentile rank of an NC score from one form is the same as that from another form. The conventional equipercentile method can be used to find score equivalents (Kolen & Brennan, 1995; Zeng & Kolen, 1995).

In DIRT OS equating, the compound binomial distribution can be used to generate the distribution of observed NC scores for examinees with a given θ (Lord & Wingersky, 1984). The extended algorithm of Lord and Wingersky's recursive formula can be applied to polytomous items (Wang, Kolen, & Harris, 1996).

For the first item,

$$P_1(X = x|\theta) = P(U_1 = x|\theta) , \quad x = 0, 1, 2, \ldots , n_1 . \tag{5}$$

For items $k = 2, 3, \ldots , K$,

$$P_k(X = x|\theta) = \sum_{u=0}^{n_k} P_{k-1}(X = x - u)P(U_k = u|\theta) , \quad x = 0, 1, 2, \ldots , \sum_{k=1}^{k} n_k , \tag{6}$$

where U_k represents a random variable for the score on item k, ranging from 0 to n_k.

After obtaining the observed NC score distribution for examinees of a given θ, the observed-score distribution of one test form (New Form) for examinees of various θs can be found by accumulating the observed-score distribution for examinees at each θ. If the θ distribution is characterized by a discrete distribution on a finite number of equally spaced points, the observed-score distribution for examinees of various θs can be approximated by summing over θs:

$$f(x) = \sum_{\theta} f(x|\theta)\psi(\theta) , \tag{7}$$

where $\psi(\theta)$ is the θ distribution, and $f(x|\theta)$ is the conditional NC score distribution, given θ, which can be obtained by Equations 5 and 6. The observed-score distribution of the old form, $g(y)$, can be found using Equations 5–7 and replacing x with y.

IRT Equating With the GRM

Under the GRM, the NC score corresponding to dichotomous items within testlet j can be classified into categories $(1, 2, \ldots , K)$. Then, the probability that a graded response to testlet j is classified into category k or higher, given θ, is

$$P_{jk}^{*}(\theta) = \begin{cases} 1 & k = 1 \\ \dfrac{1}{1 + \exp[-a_j(\theta - b_{j,k-1})]} & 2 \le k \le K \\ 0 & k > K \end{cases} , \tag{8}$$

where a_j is the testlet discrimination parameter and $b_{j,k-1}$ is the difficulty parameter of the category boundary $k-1 (2 \le k \le K)$ for testlet j. a_j is constant across the response categories of a particular testlet (i.e., constant throughout the entire reasoning process). $b_{j,k-1}$ is free to vary among the category boundaries of a particular testlet such that $b_{j,k-1} < b_{j,k}$. It is the θ value at which the probability of the response being classified into category k or higher is .5.

The probability that a graded response is classified in category k, given θ, is defined by $P_{jk}(\theta) = P^*_{jk}(\theta) - P^*_{j,k+1}(\theta)$. That is,

$$P_{jk}(\theta) = \begin{cases} 1 - \dfrac{1}{1+\exp[-a_j(\theta - b_{j1})]} & k = 1 \\ \dfrac{1}{1+\exp[-a_j(\theta - b_{j,k-1})]} - \dfrac{1}{1+\exp[-a_j(\theta - b_{jk})]} & 2 \le k \le K-1 \\ \dfrac{1}{1+\exp[-a_j(\theta - b_{j,k-1})]} & k = K \end{cases} . \quad (9)$$

An examinee's "true score" can be calculated using Equations 4 and 9, and then the procedures for IRT TS equating can be applied. From Equations 5, 6, and 9, the observed NC distribution for examinees of a given θ can be obtained, and the observed-score distribution for examinees of various θs can be found using Equation 7. The procedures for IRT OS equating then can be applied.

Method

Data

The data for this study were taken from the 1995 ITBS Form M to Form K equating study. Data were used from the sample of eighth-grade students for Reading Comprehension (Reading), Maps and Diagrams (Maps), and Math Problem Solving and Data Interpretation (Math) tests (Hoover et al., 1994). Sample size and characteristics of each test are presented in Table 1.

Table 1
Descriptive Statistics for Datasets

Form and Test	Sample Size	No. Items	No. Passages	No. Items/ Passage	Mean	S_X	Skewness	Kurtosis
Form K								
Reading	663	49	8	9, 4, 7, 5, 5, 6, 3, 10	24.9	9.99	.375	2.216
Maps	632	33	5	7, 7, 6, 6, 7	16.3	6.34	.383	2.306
Math	537	36	8	8, 4, 4, 4, 4, 4, 4, 4	16.5	6.38	.363	2.413
Form M								
Reading	680	49	7	8, 4, 9, 4, 5, 8, 11	25.9	10.53	.235	2.065
Maps	653	33	5	7, 7, 6, 6, 7	15.3	6.31	.408	2.444
Math	561	36	7	8, 6, 6, 4, 4, 4, 4	19.2	6.67	.191	2.257

Analysis

LI assessments. The LI assumption of the DIRT models was evaluated using Yen's (1984) Q_3 statistic, which is based on the correlation of the residuals of an item pair based on IRT models. The computer program IRT_LD (Chen & Thissen, 1997) was used to compute Q_3. The distributional characteristics of the pair of random variables (one for within-testlet Q_3 and one for between-testlet Q_3) for each form of each test were compared.

If there were n items in a test, $n(n-1)/2$ Q_3 statistics could be computed. In a similar way, for k_h items in the hth testlet, there could be $k_h(k_h-1)/2$ Q_3 statistics. Two types of Q_3 statistics—within-testlet and between-testlet—were distinguished for each form of each test. The number of within-testlet Q_3 statistics is

$$\sum_{h=1}^{H} k_h(k_h - 1)/2 \; . \tag{10}$$

The number of between-testlet Q_3 statistics is

$$n(n-1)/2 - \sum_{h=1}^{H} k_h(k_h - 1)/2 \; . \tag{11}$$

Factor analyses. To evaluate the unidimensionality assumption, several exploratory factor analyses were completed for each form of each test. One analysis used a tetrachoric correlation matrix, obtained using PRELIS2 (Jöreskog & Sörbom, 1993), based on individual items. Tetrachoric inter-item correlations frequently are recommended for factor analyses for dichotomously scored items, because other measures of association (e.g., ϕ correlations) might detect a second, spurious factor, which could be identified as a difficulty factor (Carroll, 1945; Hattie, 1985). Other factor analyses used product-moment correlation matrices based on testlet scores. Eigenvalues from the common factor analyses were compared, and the root mean square (RMS) of residuals was compared among several factor models with prespecified numbers of factors.

Equating. The equating designs used in the 1995 ITBS Form M to Form K equating included a single-group design and a random-groups design. For tests used in this study, only a random-groups design was used. Analyses were conducted with the computer program RAGE (Zeng, Kolen, & Hanson, 1995) to find an equating function between both forms for each test using mean, linear, and equipercentile equating methods. For DIRT, item parameters were estimated using BILOG (Mislevy & Bock, 1990); it was not necessary to place item parameters of the two forms on a common scale because a random-groups design had been used for the equating. TS and OS equating relationships were found using the computer program PIE (Hanson & Zeng, 1995). Item parameters under the NM and GRM were estimated using MULTILOG (Thissen, 1991). The test response functions for both test forms, and the TS and OS equating functions, were found using a FORTRAN 90 program written specifically for this purpose.

Evaluation of equating. The equated score distribution moments from each equating method were calculated and compared. For comparing the overall level of discrepancy between each IRT equating method and traditional equating, unweighted RMS (URMS) and weighted RMS (WRMS) were computed (Harris & Crouse, 1993).

URMS generally is defined as

$$URMS = \left[\sum_{i} (A_i - B_i)^2 / k \right]^{1/2} , \tag{12}$$

where

A_i is the equivalent of an NC score of i on the new test,

B_i is another equivalent of an NC score of i on the new test,

k represents the number of items, and

i represents each NC score point.

URMS can be used to examine differences that occur throughout the score scale. However, this index does not take into account the score distribution of the new test (or distribution of equated scores). The degree of equated score distortion within a score range that includes a large proportion of examinees would be more important than that including a fairly small proportion of examinees.

For this reason, WRMS, which is an index weighted by the probability function of examinees at equated score points, was computed as

$$WRMS = \left[\sum_i f_i (A_i - B_i)^2 \Big/ \sum_i f_i \right]^{1/2}, \tag{13}$$

where f_i is the frequency distribution of the NC score of the new test.

Restriction. One restriction of MULTILOG (Thissen, 1991) had to be addressed during the analysis. MULTILOG can accommodate no more than ten categories per item. However, both forms of the Reading test contained one testlet that had more than ten categories. In those cases, two categories were combined into a single category. Because the proportion of examinees in the combined categories was very small in both cases, the influence on the equating relationship of combining categories was not expected to be significant in any practical sense.

Results

LI

The distributional statistics for within- and between-testlet Q_3 local item dependence measures are shown in Table 2. Although Q_3 is a correlation between residuals of an item pair (i.e., zero correlation might be expected for a locally independent item pair), Q_3 has a tendency to be slightly negative in the null case (Chen & Thissen, 1997; Yen, 1984, 1993). Yen (1993) demonstrated that the expected value of Q_3, when LI holds, is approximately $-1/(n-1)$, where n is the number of test items. These approximations of the expected values for Q_3, which also are presented in Table 2, can be used as a criterion for comparing the overall level of local dependence of within- and between-testlet item pairs.

When LI holds, the averages of Q_3 from within- and between-testlet item pairs would be similar to the expected values of the Q_3 measures. Table 2 shows that the averages of between-testlet Q_3 statistics for both forms of Reading, Maps, and Math tests had values similar to the expected values of Q_3, implying that between-testlet item pairs were locally independent. In contrast, the averages of within-testlet Q_3 statistics for both forms of these tests had more positive values compared to the expected values of Q_3. The magnitudes of the differences in the Reading and Maps tests were greater than in the Math test. This suggests that LI was violated. The magnitude of the difference between the observed and expected mean of Q_3 with the SD of the observed Q_3 statistics seemed to be approximately one SD in cases where local item dependence was identified, except in the Math test. When local item dependence was not identified, the magnitude of the difference was near zero—much less than 1 SD.

Unidimensionality

Table 3 shows the RMS of the off-diagonal residuals under each specified number of factors. The difference between the RMS from the one- and two-factor models for both forms of the Reading and Maps tests and Form M of the Math test was approximately 2–6 times greater than the difference between the RMS from the two- and three-factor models. This suggests that one factor was not sufficient to describe the dimensionality of these tests. For Form K in the Math test, the difference between the RMS from the one- and two-factor models was similar to that from the two- and three-factor models. Here, one factor was sufficient to describe dimensionality.

Table 4 shows eigenvalues from the common factor analyses of product-moment correlation matrices among testlet scores. One dominant factor was evident, and the other eigenvalues were

Table 2
Distribution of Q_3 for Between- and Within-Testlet Item Pairs

Test	No. Q_3	$E(Q_3)$	Mean	$\lvert D\rvert$[a]	SD	Skewness	Kurtosis	Range
Form K								
Reading	1176	−.021						
Between	1030		−.021	.000	.043	.006	3.045	−.180 − .121
Within	146		.038	.059	.058	.044	3.328	−.131 − .196
Maps	528	−.031						
Between	435		−.029	.002	.045	−.056	2.838	−.177 − .098
Within	93		.031	.062	.049	.375	2.886	−.064 − .164
Math	630	−.029						
Between	560		−.022	.007	.049	−.013	3.084	−.183 − .137
Within	70		.008	.037	.057	.525	2.858	−.087 − .166
Form M								
Reading	1176	−.021						
Between	1007		−.027	.006	.045	−.103	3.016	−.181 − .111
Within	169		.058	.079	.069	1.014	6.097	−.080 − .400
Maps	528	−.031						
Between	435		−.033	.002	.046	.166	3.002	−.167 − .177
Within	93		.030	.061	.070	1.018	4.905	−.118 − .281
Math	630	−.029						
Between	548		−.024	.005	.045	.028	3.581	−.159 − .186
Within	82		.002	.031	.056	.008	4.440	−.191 − .172

[a]Absolute value of the difference between $E(Q_3)$ and the sample mean.

considered negligible. Unidimensionality was supported for the tests when testlet scores were used as units of analysis.

Comparisons With Traditional Equating Methods

Equated score moments. Table 5 shows the moments of converted scores for each method and the absolute value of the moment difference from the Form K moments ($|D|$). For the Reading test, the mean of the converted scores using NM-TS and NM-OS were 24.82 and 24.78, respectively; differences from the target were .03 and .07. These differences were much smaller than those of DIRT-TS and DIRT-OS (.57 and .60, respectively). Both NM-TS and NM-OS also provided more similar SD, skewness, and kurtosis values to the target than did DIRT-TS or DIRT-OS. GRM-TS and GRM-OS were more similar than DIRT-TS or DIRT-OS in terms of their moments, although they were less similar compared to NM-OS and NM-TS.

For the Maps test, the GRM provided more similar moments to those of the target than did other methods. The NM still provided more similar moments than DIRT. For the Math test, the means of NM and GRM were similar to each other and slightly more similar to the mean of the target than DIRT. However, in terms of SD, skewness, and kurtosis, no specific method provided more similar moments to those of the target. For the Reading and Maps tests, PIRT equating methods (NM or GRM) produced more similar moments than DIRT methods; for the Math test, they provided somewhat more similar moments.

Equating conditional on NC scores. Difference scores were used to compare the various equating methods. The difference scores were calculated by subtracting the equated score of a baseline equating method (mean, linear, or equipercentile) from the equated score of each equating method (DIRT- and PIRT-TS or OS). Difference scores (D) of IRT TS and OS equating methods were graphed

Table 3
RMS (× 100) of Off-Diagonal Residuals for Specified Numbers of Factors (Reading = 49 Items, Maps = 33 Items, and Math = 36 Items) and Difference of RMSs Between n-Factor and $(n + 1)$-Factor Models (Diff.)

Form and No. of Factors	Reading		Maps		Math	
	RMS	Diff.	RMS	Diff.	RMS	Diff.
Form K						
1	7.2	1.3	6.8	1.0	7.1	.8
2	5.5	.4	5.8	.5	6.3	.7
3	5.1	.3	5.3	.4	5.6	.5
4	4.8	.2	4.9	.4	5.1	.3
5	4.6	.3	4.5	.3	4.8	.3
6	4.3	.3	4.2	.3	4.5	.3
7	4.1	.2	3.9	.4	4.2	.3
8	3.9	.2	3.5	.2	3.9	.3
9	3.7	.2	3.3	.3	3.6	.2
10	3.5		3.0		3.4	
Form M						
1	7.7	2.3	7.3	1.3	7.1	1.1
2	5.4	.4	6.0	.5	6.0	.5
3	5.0	.3	5.5	.5	5.5	.4
4	4.7	.3	5.0	.5	5.1	.4
5	4.4	.2	4.5	.3	4.7	.3
6	4.2	.3	4.2	.3	4.4	.3
7	3.9	.2	3.9	.3	4.1	.3
8	3.7	.2	3.6	.3	3.8	.2
9	3.5	.1	3.3	.3	3.6	.3
10	3.4		3.0		3.3	

separately. Because these two equating methods provided similar equating relationships, only the results for the IRT TS equating method are presented for each test.

Difference score results for the Reading test are presented in Figure 1. The vertical axis in Figure 1 represents the difference score of each plotted equating method from the baseline equating equivalents; $D = 0$ represents the equating function most similar to the specified traditional equating method. The NM-TS and GRM-TS equating functions were much more similar to the mean (Figure 1a) and linear (Figure 1b) equating functions than were the DIRT-TS equating functions. For scores below 15, the NM-TS equating method produced equivalents most similar to those of mean and linear equating. In the middle score range (approximately 18–22), the equated scores of the three methods were similar to those from mean and linear equating methods. For scores above 25, GRM-TS provided equivalents most similar to those of mean and linear equating. For the equipercentile baseline (Figure 1c), DIRT-TS provided more similar equivalents only for scores of 30 and 31. Otherwise, NM-TS and GRM-TS provided more similar equivalents to those of equipercentile equating.

Figure 2 presents D plots of the TS equating for the Maps test. As was found in the case of the Reading test, the PIRT methods gave more similar results to the traditional equating methods than did DIRT equating methods. The primary difference in the trends of the Maps test from those of the Reading test was that the GRM equating methods provided similar equating functions in relation to the mean and linear equating methods. For the equipercentile method (Figure 2c), GRM-TS still

Table 4
Eigenvalues (Eig.) of Product-Moment Correlation Matrices Based on Testlet Scores and Difference Between nth and $(n + 1)$th Eigenvalues (Diff.)

Form and	Reading		Maps		Math	
Rank	Eig.	Diff.	Eig.	Diff.	Eig.	Diff.
Form K						
1	3.504	3.330	1.883	1.911	2.632	2.533
2	.174	.173	−.028	.069	.099	.066
3	.001	.043	−.097	.038	.034	.072
4	−.042	.050	−.136	.023	−.038	.031
5	−.093	.005	−.158		−.069	.029
6	−.097	.027			−.098	.044
7	−.124	.031			−.142	.045
8	−.155				−.187	
Form M						
1	3.390	3.230	1.898	1.844	2.657	2.547
2	.160	.174	.053	.144	.109	.118
3	−.014	.052	−.091	.064	−.008	.045
4	−.067	.036	−.155	.030	−.053	.045
5	−.103	.013	−.184		−.098	.050
6	−.115	.050			−.148	.027
7	−.165				−.174	

provided the most similar equating function to that of the baseline method. NM-TS (and NM-OS) were more similar to the traditional equating methods than were DIRT-TS and DIRT-OS, except for scores below 4 and above 26. Because the number of examinees in these score ranges were relatively small, NM-TS and NM-OS would be expected to produce more similar equating functions to those of the traditional equating methods than DIRT-OS and DIRT-TS.

Figure 3 presents the *D* plots of the TS equating for the Math test. DIRT-TS and PIRT-TS provided very similar equating functions, except for scores below 11, where NR-TS and GRM-TS were more similar to the baseline methods. The IRT assumptions were violated less with the Math test than with the Reading or Maps tests. This might have contributed to the relative similarity noted among the three equating methods for the Math test, compared to the dissimilarity found for the Reading and Maps tests. These results suggest that the more the IRT assumptions are violated, the greater the discrepancy among equating methods.

URMS and WRMS. URMS and WRMS between equated scores and the baselines are presented in Table 6. For the Reading test, PIRT equating methods provided more similar equating relationships to the baseline methods than DIRT methods. This was true when either URMS or WRMS was used, but the difference was clearer using WRMS.

For the Maps test, GRM-TS and GRM-OS were more similar to the baseline methods than were DIRT or NM methods, whether URMS or WRMS was used. The differences in URMS between DIRT and NM equating methods were not consistent, although the differences in WRMS between OS equating methods based on the two models were more distinct (WRMS of the NM-OS equating method was smaller than that of DIRT-OS).

For the Math test and URMS, the differences among the three IRT models were not large. However, the WRMS of the PIRT methods had somewhat smaller values than that of the DIRT methods. This

Table 5
Moments for Equating Form M to Form K for Mean, Linear, and Equipercentile (EE) Baseline Methods and DIRT and PIRT Methods

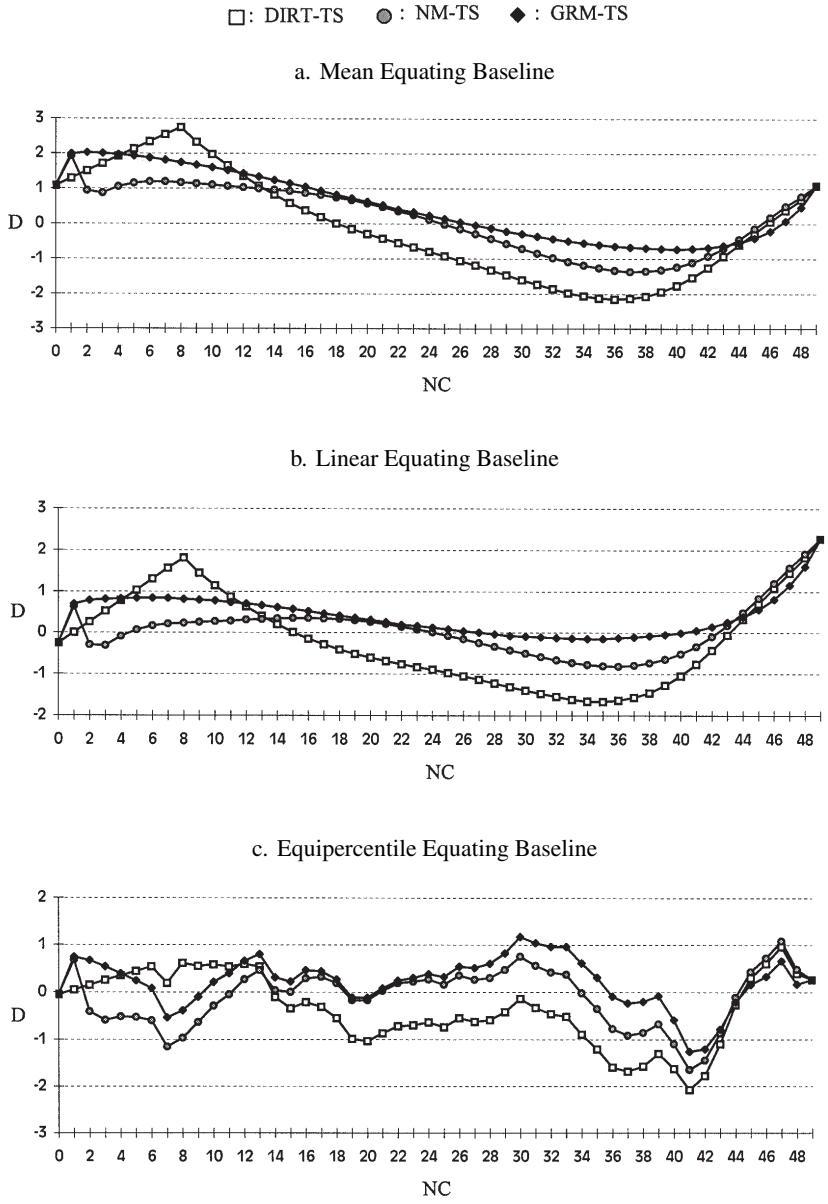
Test	Mean	$\|D\|$	SD	$\|D\|$	Skewness	$\|D\|$	Kurtosis	$\|D\|$
Reading								
Form K	24.85		9.985		.375		2.216	
Form M	25.93		10.529		.235		2.065	
Mean	24.85	.00	10.529	.544	.235	.140	2.065	.151
Linear	24.85	.00	9.985	.000	.235	.140	2.065	.151
EE	24.85	.00	9.986	.001	.372	.003	2.217	.001
DIRT-TS	24.28	.57	9.671	.314	.481	.106	2.396	.180
DIRT-OS	24.25	.60	9.781	.204	.408	.033	2.327	.111
NM-TS	24.82	.03	9.841	.144	.337	.038	2.322	.106
NM-OS	24.78	.07	9.722	.263	.316	.059	2.272	.056
GRM-TS	25.10	.25	9.840	.145	.325	.050	2.198	.018
GRM-OS	25.07	.22	9.773	.212	.303	.072	2.169	.047
Maps								
Form K	16.28		6.337		.383		2.306	
Form M	15.28		6.307		.408		2.444	
Mean	16.28	.00	6.307	.030	.408	.025	2.444	.138
Linear	16.28	.00	6.337	.000	.408	.025	2.444	.138
EE	16.27	.01	6.326	.011	.376	.007	2.291	.015
DIRT-TS	16.03	.25	6.022	.315	.576	.193	2.582	.276
DIRT-OS	15.98	.30	6.085	.252	.487	.104	2.542	.236
NM-TS	16.11	.17	6.071	.266	.428	.045	2.337	.031
NM-OS	16.08	.20	6.172	.165	.389	.006	2.347	.041
GRM-TS	16.19	.09	6.271	.066	.344	.039	2.364	.058
GRM-OS	16.21	.07	6.273	.064	.345	.038	2.363	.057
Math								
Form K	16.48		6.379		.363		2.413	
Form M	19.17		6.674		.191		2.257	
Mean	16.48	.00	6.674	.295	.191	.172	2.257	.156
Linear	16.48	.00	6.379	.000	.191	.172	2.257	.156
EE	16.48	.00	6.390	.011	.366	.003	2.451	.038
DIRT-TS	16.79	.31	6.361	.018	.385	.022	2.352	.061
DIRT-OS	16.73	.25	6.440	.061	.312	.051	2.319	.094
NM-TS	16.68	.20	6.399	.020	.316	.047	2.282	.131
NM-OS	16.63	.15	6.466	.087	.297	.066	2.254	.159
GRM-TS	16.70	.22	6.399	.020	.357	.006	2.406	.007
GRM-OS	16.65	.17	6.432	.053	.335	.028	2.361	.052

might have been caused by the assumptions of DIRT modeling being violated less in the Math test than in the Reading or Maps tests.

Discussion and Conclusion

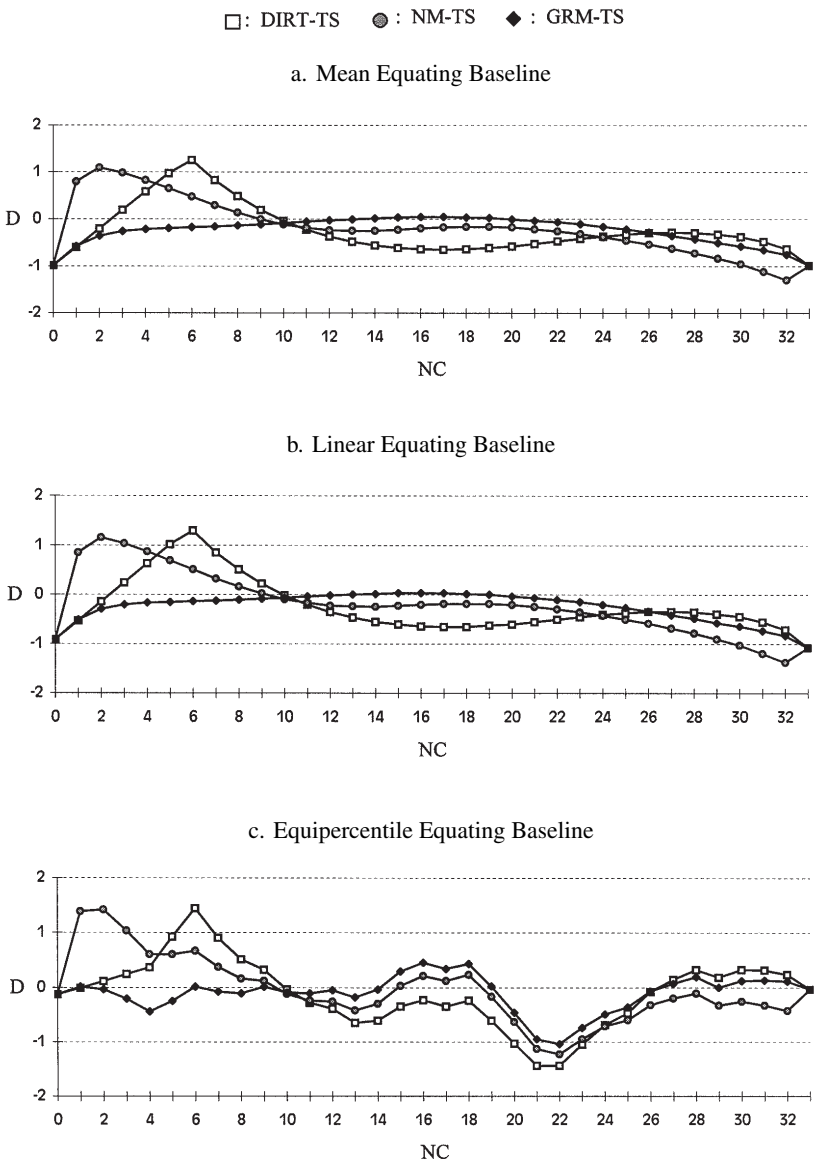
LI and unidimensionality assumptions were violated for testlet-based tests when individual items were used as the unit of analysis, although these assumptions were violated less with the Math test than with the Reading or Maps tests. However, those assumptions were satisfied when using testlets as the unit of analysis. Based on these results, the unidimensional DIRT model might be problematic for equating testlet-based tests. That is, the common use of a unidimensional DIRT model in this equating situation should be suspect because of assumption violations. In contrast, the use of PIRT models instead of DIRT models was appropriate.

Figure 1
Comparison of DIRT-TS, NM-TS, and GRM-TS Equating Using Traditional Equating Methods as Baselines for the Reading Test



The PIRT TS and OS equating methods provided equating relationships that were more similar to mean, linear, and equipercentile equating methods than were DIRT TS or OS equating methods for the Reading and Maps tests. This result was unclear in the Math test, perhaps because of the violations of LI and unidimensionality in DIRT modeling when testlets were used. To evaluate the validity of IRT model equating methods, a comparison of the equating functions derived from

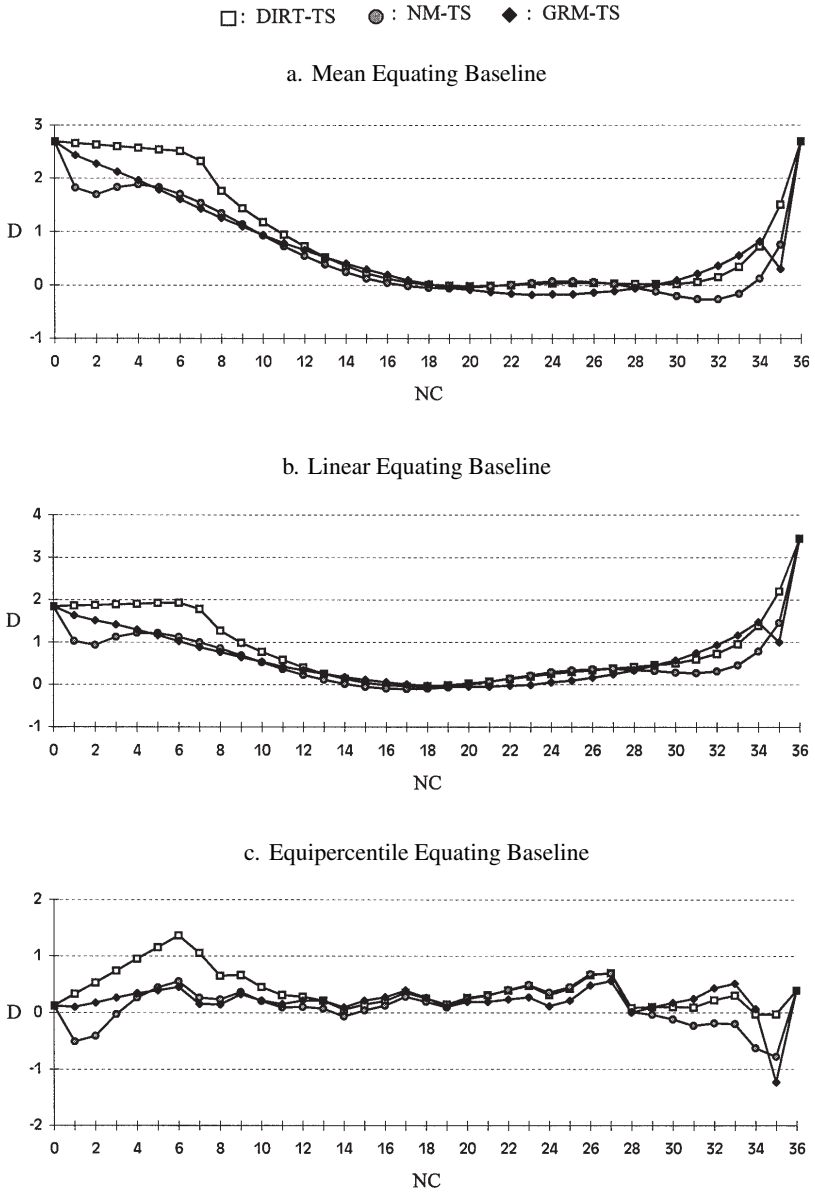
Figure 2
Comparison of DIRT-TS, NM-TS, and GRM-TS Equating Using Traditional Equating Methods as Baselines for the Maps Test



traditional and IRT methods might be examined. Because PIRT models satisfied the assumptions of IRT modeling more closely than DIRT models for testlet-based tests, it is reasonable to expect better equating relationships using PIRT than DIRT. The NM and GRM seem to offer useful approaches for equating testlet-based tests.

The practical effect of the difference between item-based and testlet-based equating methods can be demonstrated with an example to examine how serious the violation of assumptions is for

Figure 3
Comparison of DIRT-TS, NM-TS, and GRM-TS Equating Using Traditional Equating Methods as Baselines for the Math Test



the distortion of the equated scores. Suppose a certain eighth-grade student obtained an NC score of 35 on Form M of the Reading test. That student's NC score equivalent on Form K using an item-based equating method would differ by approximately two points from the result obtained by traditional baseline equating methods. However, the value from using a testlet-based equating method would yield almost the same value as the traditional baseline equating methods. This

Table 6
URMS and WRMS for Each IRT Equating Method Using Mean, Linear, and Equipercentile (EE) Methods as Baselines

Test	URMS			WRMS		
Method	Mean	Linear	EE	Mean	Linear	EE
Reading						
DIRT-TS	1.482	1.115	.834	1.335	1.055	.884
NM-TS	.942	.657	.614	.867	.533	.551
GRM-TS	1.058	.632	.556	.807	.440	.562
DIRT-OS	1.241	.874	.793	1.151	.890	.808
NM-OS	1.085	.589	.644	.919	.503	.613
GRM-OS	1.071	.518	.572	.832	.398	.572
Maps						
DIRT-TS	.577	.602	.630	.514	.533	.663
NM-TS	.618	.656	.605	.349	.374	.478
GRM-TS	.376	.393	.351	.172	.190	.380
DIRT-OS	.426	.462	.547	.426	.447	.592
NM-OS	.444	.484	.471	.287	.309	.435
GRM-OS	.295	.321	.357	.160	.177	.371
Math						
DIRT-TS	1.388	1.200	.513	.607	.515	.380
NM-TS	1.064	.887	.352	.455	.359	.311
GRM-TS	1.142	.977	.339	.485	.396	.271
DIRT-OS	.930	.749	.341	.420	.351	.302
NM-OS	.924	.694	.394	.351	.294	.292
GRM-OS	.990	.796	.306	.410	.331	.235

difference does not appear to be negligible. When using an item-based equating method, that student would obtain a standard score of eight points lower, a national percentile rank of five points lower, and a grade-equivalent score of five months lower.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlation between items or between tests. *Psychometrika, 10,* 1–19.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265–289.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10,* 37–45.

Hanson, B. A., & Zeng, L. (1995). *A computer program for IRT equating (PIE) (Version 1.0)* [Computer program]. Iowa City IA: ACT.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6,* 195–240.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Hoover, H. D., Hieronymous, A. N., Frisbie, D. A., & Dunbar, S. B. (1994). *Iowa tests of basic skills: Interpretive guide for school administrators.* Chicago: Riverside Publishing.

Jöreskog, K. G., & Sörbom, D. (1993). *PRELIS2 user's reference guide.* Chicago: Scientific Software.

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice, 7,* 29–36.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Lee, G. (2000). Estimating conditional standard errors of measurement for tests composed of testlets. *Applied Measurement in Education, 13,* 161–180.

Lee, G., Dunbar, S. B., & Frisbie, D. A. (1999, April). *Measurement models for a testlet-based test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12,* 237–255.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 452–461.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models (2nd ed.)* [Computer program]. Mooresville IN: Scientific Software.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement No. 17.*

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Thissen, D. (1991). *MULTILOG: Multiple categorical item analysis and test scoring using item response theory (Version 6.0)* [Computer program]. Chicago: Scientific Software.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical models. *Journal of Educational Measurement, 26,* 247–260.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8,* 157–186.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1–14.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definition and detecting. *Journal of Educational Measurement, 28,* 197–219.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15,* 22–29.

Wang, T., Kolen, M. J., & Harris, D. J. (1996, April). *Conditional standard errors, reliability and decision consistency of performance levels using polytomous IRT.* Paper presented at the annual meeting of the American Educational Research Association, New York.

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.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213.

Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement, 19,* 231–240.

Zeng, L., Kolen, M. J., & Hanson, B. A. (1995). *Random groups equating program (RAGE) (Version 2.0)* [Computer program]. Iowa City IA: ACT.

Author's Address

Send requests for reprints or further information to Guemin Lee, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey CA 93940, U.S.A. Email: glee@ctb.com.

Least Squares Estimation of Item Response Theory Linking Coefficients

Haruhiko Ogasawara
Otaru University of Commerce

Three types of least squares estimation (generalized, unweighted, and weighted) for item response theory linking coefficients are discussed. Unweighted least squares estimation gives somewhat large asymptotic standard errors. Although generalized least squares has the smallest asymptotic standard errors, it frequently gives biased estimates. Thus, weighted least squares estimation is the preferred method. In weighted least squares estimation, the ordinary weights are replaced with their powers. Results from a monte carlo simulation study showed that the weighted least squares method generally reduced bias without increasing the asymptotic standard errors, in comparison to other least squares methods. *Index terms: asymptotic standard errors, common items, equating, item response theory, linking, weighted least squares.*

Within item response theory (IRT), there is an indeterminacy of location and scale for some parameters. To remove this indeterminacy, examinee (θ) parameters usually are restricted to have a mean of zero and unit variance. Therefore, linking is required when two sets of parameters estimated separately are to be compared with each other. IRT linking typically is implemented by using the means (first moment) of item parameter estimates or the estimated test response functions (TRFs) or item response functions.

The moments method [i.e., the mean/mean (M/M) method] uses two sets of moments from the parameter estimates of the common items in two tests to be linked or equated. For example, a set of linking coefficients (constants) is estimated so that the means of the difficulty and discrimination parameter estimates in one test become equivalent to those in the other test. The TRF method obtains a set of linking coefficients that causes the two TRFs (from different sets of common item parameter estimates) to become as close as possible to each other in least squares over a θ distribution (Hambleton, Swaminathan, & Rogers, 1991, Ch. 9; Kolen & Brennan, 1995, Ch. 6).

Ogasawara (2001) investigated the statistical behavior of these methods, assuming the three-parameter logistic model (3PLM) using marginal maximum likelihood (MML) item parameter estimation. He showed that the standard errors (SEs) of the parameter estimates from the TRF method tended to be smaller than those from the M/M method. One disadvantage of the TRF method is that the formula for the asymptotic SEs of the parameter estimates is complex. Divgi (1985) provided a minimum χ^2 method for estimating linking coefficients. He used the asymptotic variance-covariance matrix of item parameter estimates, given θ parameters (Lord, 1980, p. 191). However, because it is based on the assumption that θ parameters are known, this matrix is an underestimate. In addition, an iterative computation is required for each linking coefficient.

Purpose

Least squares estimation methods for IRT linking coefficients without iterative computation are proposed, and their asymptotic SEs are provided. A simulation study shows the accuracy of the SEs and compares them with those found using the M/M and TRF methods.

Applied Psychological Measurement, Vol. 25 No. 4, December 2001, 373–383

IRT Model and Linking Design

In the 3PLM, the probability of a correct response to item j ($j = 1, 2, \ldots, q_k$) in test k by examinee i with θ_{ki} ($i = 1, 2, \ldots, N_k$) in the kth examinee group is

$$P_{kj}(\theta_{ki}) = c_{kj} + \frac{1 - c_{kj}}{1 + \exp[-Da_{kj}(\theta_{ki} - b_{kj})]} , \tag{1}$$

where

a_{kj} is the discrimination parameter for item j in test k,
b_{kj} is the difficulty parameter for item j in test k,
c_{kj} is the guessing parameter for item j in test k,
N_k is the number of examinees in the kth examinee group,
q_k is the number of items in the kth test, and
D is 1.7.

Assume p common items (an anchor test). That is, the kth test (calibration) is made up of p common items and $q_k - p$ unique items. The common items can be internal or external. If they are external, the kth test is made up of $q_k - p$ unique items and the anchor test is used only for linking.

Because Equation 1 has an indeterminacy, it is assumed that θ_{ki} has a N(0, 1) normal distribution for each calibration. This gives different scales of a_{kj} and b_{kj} from calibration to calibration. Assume that item responses are as in Equation 1 and that the same examinee i responds to p common items in each of two tests. Because the probability of a correct response to an item is unchanged with respect to different calibrations,

$$\theta_{1i} = A\theta_{2i} + B , \quad a_{1j} = a_{2j}/A , \quad b_{1j} = Ab_{2j} + B , \quad c_{1j} = c_{2j} , \tag{2}$$

where $j = 1, 2, \ldots, p$, and A and B are the linking coefficients from the second test (Test 2) to the first test (Test 1). Even with optimal $\hat{A}$ and $\hat{B}$ given the item parameter estimates in a sample, Equation 2 holds only approximately.

Estimation of Linking Coefficients

Generalized Least Squares Estimation

From Equation 2,

$$A = a_{2j}/a_{1j} , \quad B = b_{1j} - Ab_{2j} = b_{1j} - a_{2j}b_{2j}/a_{1j} , \tag{3}$$

in the population, where $j = 1, 2, \ldots, p$. Let

$$u_j = a_{2j}/a_{1j} , \quad v_j = b_{1j} - a_{2j}b_{2j}/a_{1j} , \tag{4}$$

where

$j = 1, 2, \ldots, p$,
$\mathbf{u} = (u_1, u_2, \ldots, u_p)'$,
$\mathbf{v} = (v_1, v_2, \ldots, v_p)'$, and
$\mathbf{w} = (\mathbf{u}', \mathbf{v}')'$.

The generalized least squares (GLS) estimates, $\hat{A}_{\mathrm{GLS}}$ and $\hat{B}_{\mathrm{GLS}}$, then are obtained by minimizing

$$\begin{aligned} f_{\mathrm{GLS}} &= (1/2)\left(A_{\mathrm{GLS}} - \hat{u}_1, \ldots, A_{\mathrm{GLS}} - \hat{u}_p , B_{\mathrm{GLS}} - \hat{v}_1, \ldots, B_{\mathrm{GLS}} - \hat{v}_p\right) \\ &\quad \times \mathrm{acov}^{-1}\left(\hat{\mathbf{w}}\right)\left(A_{\mathrm{GLS}} - \hat{u}_1, \ldots, A_{\mathrm{GLS}} - \hat{u}_p , B_{\mathrm{GLS}} - \hat{v}_1, \ldots, B_{\mathrm{GLS}} - \hat{v}_p\right)' \\ &= (1/2)\left(\mathbf{1}_p'A_{\mathrm{GLS}} - \hat{\mathbf{u}}' , \mathbf{1}_p'B_{\mathrm{GLS}} - \hat{\mathbf{v}}'\right)\mathrm{acov}^{-1}\left(\hat{\mathbf{w}}\right)\left(\mathbf{1}_p'A_{\mathrm{GLS}} - \hat{\mathbf{u}}'\mathbf{1}_p'B_{\mathrm{GLS}} - \hat{\mathbf{v}}'\right)' , \end{aligned} \tag{5}$$

where

$\mathrm{acov}(\hat{\mathbf{w}})$ is the asymptotic variance-covariance matrix for $\hat{\mathbf{w}}$,

$\mathrm{acov}^{-1}(\hat{\mathbf{w}}) = [\mathrm{acov}(\hat{\mathbf{w}})]^{-1}$, and

$\mathbf{1}_p$ is the column vector with p 1s.

$\mathrm{acov}(\hat{\mathbf{w}})$ can be obtained by using the delta method,

$$\mathrm{acov}(\hat{\mathbf{w}}) = \frac{\partial \mathbf{w}}{\partial \boldsymbol{\alpha}'} \mathrm{acov}(\hat{\boldsymbol{\alpha}}) \frac{\partial \mathbf{w}'}{\partial \boldsymbol{\alpha}} , \tag{6}$$

where $\mathrm{acov}(\hat{\boldsymbol{\alpha}})$ is the asymptotic variance-covariance matrix of the estimate of the parameter vector $\boldsymbol{\alpha}$ for p common items. That is,

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')' , \tag{7}$$

with

$$\boldsymbol{\alpha}_k = \left(\boldsymbol{\alpha}_{k1}', \ldots, \boldsymbol{\alpha}_{kp}'\right)' , \tag{8}$$

and

$$\boldsymbol{\alpha}_{kj} = (\alpha_{kj}, b_{kj}, c_{kj})' , \tag{9}$$

where $k = (1, 2)$ and $j = 1, 2, \ldots, p$.

$\mathrm{acov}(\hat{\boldsymbol{\alpha}})$ is assumed to be estimated using MML (Bock & Aitkin, 1981; Bock & Lieberman, 1970). The two examinee groups are assumed to be selected independently from possibly nonequivalent populations. By this assumption, $\mathrm{acov}(\hat{\boldsymbol{\alpha}})$ becomes a block diagonal matrix with two nonzero submatrices, $\mathrm{acov}(\hat{\boldsymbol{\alpha}}_1)$ and $\mathrm{acov}(\hat{\boldsymbol{\alpha}}_2)$.

The partial derivatives in Equation 6 (see also Equation 4) are obtained easily—their nonzero elements are

$$\begin{aligned} &\frac{\partial u_j}{\partial a_{1j}} = -\frac{a_{2j}}{a_{1j}^2} , \quad \frac{\partial u_j}{\partial a_{2j}} = \frac{1}{a_{1j}} , \quad \frac{\partial v_j}{\partial a_{1j}} = \frac{a_{2j} b_{2j}}{a_{1j}^2} , \\ &\frac{\partial v_j}{\partial b_{1j}} = 1 , \quad \frac{\partial v_j}{\partial a_{2j}} = -\frac{b_{2j}}{a_{1j}} , \quad \frac{\partial v_j}{\partial b_{2j}} = -\frac{a_{2j}}{a_{1j}} , \end{aligned} \tag{10}$$

where $j = 1, 2, \ldots, p$. Solving $\partial f_{\mathrm{GLS}} / \partial (A_{\mathrm{GLS}}, B_{\mathrm{GLS}}) = (0, 0)$,

$$\begin{pmatrix} \hat{A}_{\mathrm{GLS}} \\ \hat{B}_{\mathrm{GLS}} \end{pmatrix} = W^{-1} \begin{pmatrix} \mathbf{1}_p' & \mathbf{0}' \\ \mathbf{0}' & \mathbf{1}_p' \end{pmatrix} \mathrm{acov}^{-1}(\hat{\mathbf{w}})\, \hat{\mathbf{w}} , \tag{11}$$

where

$$W = \begin{pmatrix} \mathbf{1}_p' & \mathbf{0}' \\ \mathbf{0}' & \mathbf{1}_p' \end{pmatrix} \mathrm{acov}^{-1}(\hat{\mathbf{w}}) \begin{pmatrix} \mathbf{1}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_p \end{pmatrix} . \tag{12}$$

It follows from Equation 11 that

$$\begin{aligned} \mathrm{acov} \begin{pmatrix} \hat{A}_{\mathrm{GLS}} \\ \hat{B}_{\mathrm{GLS}} \end{pmatrix} &= W^{-1} \begin{pmatrix} \mathbf{1}_p' & \mathbf{0}' \\ \mathbf{0}' & \mathbf{1}_p' \end{pmatrix} \mathrm{acov}^{-1}(\hat{\mathbf{w}})\, E\left[(\hat{\mathbf{w}} - \mathbf{w})(\hat{\mathbf{w}} - \mathbf{w})'\right] \times \mathrm{acov}^{-1}(\hat{\mathbf{w}}) \begin{pmatrix} \mathbf{1}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_p \end{pmatrix} W^{-1} \\ &= W^{-1} W W^{-1} = W^{-1} , \end{aligned} \tag{13}$$

where E[·] is taken in large samples—that is, E[·] = acov$(\hat{\mathbf{w}})$. The asymptotic SEs of $\hat{A}_{\mathrm{GLS}}$ and $\hat{B}_{\mathrm{GLS}}$ are the square roots of the diagonal elements of Equation 13. Note that acov$(\hat{\mathbf{w}})$ in Equations 11 and 13 usually is unknown, but the matrix is estimated by replacing the population values of the parameters in Equation 6 with their estimates.

Unweighted Least Squares (ULS) Estimation

The ULS estimates, $\hat{A}_{\mathrm{ULS}}$ and $\hat{B}_{\mathrm{ULS}}$, are obtained by minimizing

$$f_{\mathrm{ULS}} = (1/2)\left(\mathbf{1}'_p A_{\mathrm{ULS}} - \hat{\mathbf{u}}' \, , \mathbf{1}'_p B_{\mathrm{ULS}} - \hat{\mathbf{v}}'\right)\left(\mathbf{1}'_p A_{\mathrm{ULS}} - \hat{\mathbf{u}}' \, , \mathbf{1}'_p B_{\mathrm{ULS}} - \hat{\mathbf{v}}'\right)' \, , \tag{14}$$

which gives

$$\left(\hat{A}_{\mathrm{ULS}}, \hat{B}_{\mathrm{ULS}}\right) = \left(\mathbf{1}'_p \hat{\mathbf{u}} \, , \mathbf{1}'_p \hat{\mathbf{v}}\right)/p \, , \tag{15}$$

with

$$\mathrm{acov}\begin{pmatrix}\hat{A}_{\mathrm{ULS}} \\ \hat{B}_{\mathrm{ULS}}\end{pmatrix} = \begin{pmatrix}\mathbf{1}'_p & \mathbf{0}' \\ \mathbf{0}' & \mathbf{1}'_p\end{pmatrix} \mathrm{acov}\left(\hat{\mathbf{w}}\right) \begin{pmatrix}\mathbf{1}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_p\end{pmatrix} / p^2 \, . \tag{16}$$

Equations 11 and 15 show that the asymptotic covariance matrix of $(\hat{A}_{\mathrm{GLS}}, \hat{B}_{\mathrm{GLS}})$ with respect to $(\hat{A}_{\mathrm{ULS}}, \hat{B}_{\mathrm{ULS}})$ is

$$\mathrm{acov}\left[\left(\hat{A}_{\mathrm{GLS}}\hat{B}_{\mathrm{GLS}}\right)' \, ; \left(\hat{A}_{\mathrm{ULS}}\hat{B}_{\mathrm{ULS}}\right)\right] = W^{-1} \, , \tag{17}$$

which is equivalent to Equation 13.

Weighted Least Squares (WLS) Estimation

GLS estimates of equating coefficients have asymptotic efficiency; however, they tend to be biased in samples (see below). ULS estimates, although having larger variances than GLS estimates, are less biased. It is natural to compromise these two estimates. WLS estimation, with some extension, is a good alternative. WLS minimizes

$$f_{\mathrm{W}m} = (1/2)\left(\mathbf{1}'_p A_{\mathrm{W}m} - \hat{\mathbf{u}}', \mathbf{1}'_p B_{\mathrm{W}m} - \hat{\mathbf{v}}'\right)\left[\mathrm{Diag\ acov}\left(\hat{\mathbf{w}}\right)\right]^{-m} \times \left(\mathbf{1}'_p A_{\mathrm{W}m} - \hat{\mathbf{u}}', \mathbf{1}'_p B_{\mathrm{W}m} - \hat{\mathbf{v}}'\right)' \, , \tag{18}$$

where Diag(·) is a diagonal matrix. That is, WLS estimation with $m = 1$ uses only the variances of $\hat{\mathbf{w}}$ among the elements of acov$(\hat{\mathbf{w}})$. When $m = 0$, it becomes ULS estimation. Because the usual WLS estimates with $m = 1$ are not necessarily optimum, $m(0 \leq m \leq 1)$ can be used as a reasonable compromise between GLS and ULS (or between standard WLS and ULS). The WLS estimates that minimize Equation 18 are

$$\begin{pmatrix}\hat{A}_{\mathrm{W}m} \\ \hat{B}_{\mathrm{W}m}\end{pmatrix} = \left\{\begin{pmatrix}\mathbf{1}'_p & \mathbf{0}' \\ \mathbf{0}' & \mathbf{1}'_p\end{pmatrix}\left[\mathrm{Diag\ acov}\left(\hat{\mathbf{w}}\right)\right]^{-m}\begin{pmatrix}\mathbf{1}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_p\end{pmatrix}\right\}^{-1}\begin{pmatrix}\mathbf{1}'_p & \mathbf{0}' \\ \mathbf{0}' & \mathbf{1}'_p\end{pmatrix} \times \left[\mathrm{Diag\ acov}\left(\hat{\mathbf{w}}\right)\right]^{-m} \hat{\mathbf{w}} \, . \tag{19}$$

The asymptotic covariance matrix for the estimates of Equation 19 is

$$\begin{aligned}\operatorname{acov}\begin{pmatrix}\hat{A}_{\mathrm{W}m}\\ \hat{B}_{\mathrm{W}m}\end{pmatrix} &= \left\{\begin{pmatrix}\mathbf{1}'_p & \mathbf{0}'\\ \mathbf{0}' & \mathbf{1}'_p\end{pmatrix}\left[\operatorname{Diag}\operatorname{acov}\left(\hat{\mathbf{w}}\right)\right]^{-m}\begin{pmatrix}\mathbf{1}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{1}_p\end{pmatrix}\right\}^{-1}\begin{pmatrix}\mathbf{1}'_p & \mathbf{0}'\\ \mathbf{0}' & \mathbf{1}'_p\end{pmatrix}\\ &\quad\times\left[\operatorname{Diag}\operatorname{acov}\left(\hat{\mathbf{w}}\right)\right]^{-m}\operatorname{acov}\left(\hat{\mathbf{w}}\right)\left[\operatorname{Diag}\operatorname{acov}\left(\hat{\mathbf{w}}\right)\right]^{-m}\\ &\quad\times\begin{pmatrix}\mathbf{1}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{1}_p\end{pmatrix}\left\{\begin{pmatrix}\mathbf{1}'_p & \mathbf{0}'\\ \mathbf{0}' & \mathbf{1}'_p\end{pmatrix}\left[\operatorname{Diag}\operatorname{acov}\left(\hat{\mathbf{w}}\right)\right]^{-m}\begin{pmatrix}\mathbf{1}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{1}_p\end{pmatrix}\right\}^{-1}. \end{aligned} \tag{20}$$

From Equations 11, 15, and 19, the asymptotic covariance matrices of $(\hat{A}_{\mathrm{GLS}}, \hat{B}_{\mathrm{GLS}})$ and $(\hat{A}_{\mathrm{ULS}}, \hat{B}_{\mathrm{ULS}})$, respectively, in relation to $(\hat{A}_{\mathrm{W}m}, \hat{B}_{\mathrm{W}m})$ are

$$\operatorname{acov}\left[\left(\hat{A}_{\mathrm{GLS}}\,\hat{B}_{\mathrm{GLS}}\right)',\quad\left(\hat{A}_{\mathrm{W}m}\,\hat{B}_{\mathrm{W}m}\right)\right] = W^{-1}, \tag{21}$$

and

$$\begin{aligned}\operatorname{acov}\left[\left(\hat{A}_{\mathrm{ULS}}\,\hat{B}_{\mathrm{ULS}}\right)';\quad\left(\hat{A}_{\mathrm{W}m}\,\hat{B}_{\mathrm{W}m}\right)\right] &= \frac{1}{p}\begin{pmatrix}\mathbf{1}'_p & \mathbf{0}'\\ \mathbf{0}' & \mathbf{1}'_p\end{pmatrix}\operatorname{acov}\left(\hat{\mathbf{w}}\right)\left[\operatorname{Diag}\operatorname{acov}\left(\hat{\mathbf{w}}\right)\right]^{-m}\begin{pmatrix}\mathbf{1}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{1}_p\end{pmatrix}\\ &\quad\times\left\{\begin{pmatrix}\mathbf{1}'_p & \mathbf{0}'\\ \mathbf{0}' & \mathbf{1}'_p\end{pmatrix}\left[\operatorname{Diag}\operatorname{acov}\left(\hat{\mathbf{w}}\right)\right]^{-m}\begin{pmatrix}\mathbf{1}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{1}_p\end{pmatrix}\right\}^{-1}. \end{aligned} \tag{22}$$

As described above, the population values of the parameters in Equations 19–22 can be replaced by their estimates.

Comparison of Linking Methods

The M/M, TRF, ULS, WLS, and GLS methods for estimating linking coefficients were compared on four simulated datasets and one real dataset.

Simulated Data

Four datasets—Data 1, Data 2, Data 3, and Data 4—were simulated with true linking coefficient values. Data 1 and 2 assumed the two-parameter logistic model (2PLM); Data 3 and 4 assumed the 3PLM. Because stable parameter estimates (which lead to empirical SEs) were unavailable for the 3PLM, the 2PLM was used to confirm the accuracy of the formulas. Because the 2PLM is a special case of the 3PLM, this was insufficient to confirm the accuracy of the 3PLM formulas. However, it is usual to impose restrictions or use prior distributions for some parameters in practice. In an extreme case with fixed nonzero guessing parameters, the situation becomes essentially equivalent to that of the 2PLM, which gives partial justification to the method used.

Data 1 and 2. Data 1 and 2 had 10 and 15 internal common items, respectively, and the same number of unique items. That is, Tests 1 and 2 each had 20 items in Data 1, and 30 items in Data 2. Using the 2PLM, the discrimination parameter population values were randomly generated from a uniform (.3, 1.3) distribution. Difficulty parameter population values were generated using a N(0, 1) distribution. θs for 1,000 examinees in Test 1 were randomly generated using N(0, 1); those for Test 2 were randomly generated using N(.5, 1.2^2). $A = 1.2$ and $B = .5$ for Data 1 and 2, respectively. The response data were generated by comparing values of random uniform numbers in the range (0, 1) to the population values of correct response probability functions. That is, when a value of the uniform random number was less than the value of the population probability given

by the population item parameters (with the linking coefficients) and a generated θ, the response was specified as correct. When the value of the uniform random number was equal to or greater than the probability, the response was specified as incorrect.

Data 3 and 4. Data 3 had 15 internal common items and 15 unique items; Data 4 had 20 internal common items and 40 unique items. That is, Tests 1 and 2 each had 30 items in Data 3 and 60 items in Data 4. Using the 3PLM, the discrimination parameters were generated using a uniform (.4, 1.4) distribution. Difficulty parameters were generated as in Data 1 and 2, but with different seeds for random numbers. Guessing parameters were generated using a uniform (.1, .2) distribution. All other variables were as Data 1 and Data 2.

Method. From the generated response patterns, item parameters were estimated separately in each test using MML for the common and unique items. Using the two common item parameter estimate sets, the linking coefficients A and B and their asymptotic SEs were estimated using ULS, WLS ($m = .5$ and $m = 1$), and GLS estimation methods. For the 2PLM, item parameter estimation and estimation of A and B were repeated 100 times, resulting in 100 sets of $\hat{A}$ and $\hat{B}$ with corresponding estimated asymptotic SEs. For the 3PLM, the population item parameters were used in place of parameter estimates. However, the asymptotic SEs were estimated using actual response patterns (see below), even in the case of the 3PLM with population item parameters.

The linking coefficient estimates in the M/M method are

$$\hat{A}_{\mathrm{M}} = \sum_{j=1}^{p} \hat{a}_{2j} / \sum_{j=1}^{p} \hat{a}_{1j} \ , \quad \hat{B}_{\mathrm{M}} = \sum_{j=1}^{p} \hat{b}_{1j}/p - \hat{A}_{\mathrm{M}} \sum_{j=1}^{p} \hat{b}_{2j}/p \ . \tag{23}$$

Using the TRF method, the linking coefficient estimates ($\hat{A}_{\mathrm{TRF}}$ and $\hat{B}_{\mathrm{TRF}}$) are those that minimize

$$f_{\mathrm{TRF}} = \int_{-\infty}^{+\infty} \left\{ \sum_{j=1}^{p} \left\{ \hat{c}_{1j} + \frac{1-\hat{c}_{1j}}{1+\exp\left[-D\hat{a}_{1j}(\theta - \hat{b}_{1j})\right]} - \hat{c}_{2j} \right.\right.$$
$$\left.\left. - \frac{1-\hat{c}_{2j}}{1+\exp\left[-D(\hat{a}_{2j}/A_{\mathrm{TRF}})(\theta - A_{\mathrm{TRF}}\hat{b}_{2j} - B_{\mathrm{TRF}})\right]} \right\} \right\}^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right) d\theta \ , \tag{24}$$

where $\hat{c}_{1j} = \hat{c}_{2j} = 0$ $(j = 1, 2, \ldots, p)$ for the 2PLM. The definition of Equation 24 is slightly different from its original (Stocking & Lord, 1983)—integration was used here in place of the summation over examinees.

When a set of population item parameters ($\boldsymbol{\alpha}$) is given, acov($\hat{\boldsymbol{\alpha}}$) can be theoretically obtained considering all possible response patterns and their probabilities (Bock & Lieberman, 1970). Thus, the true asymptotic SEs for $\hat{A}$ and $\hat{B}$ also can be derived, because $\boldsymbol{\alpha}$ is known. However, when there is a moderate number of items in each test, considerable computation is required. In the simulation, acov($\hat{\boldsymbol{\alpha}}$) and, consequently, the SEs of the linking coefficients, were estimated using only observed response patterns as approximations, which gives somewhat different values of the estimated SEs from replication to replication. This also holds for the 3PLM using population item parameters. For item parameter and linking coefficient estimation using the TRF method, 15 quadrature points were used to integrate out θ parameters.

Results. Table 1 shows the mean, standard deviation (SD), and the mean and SD of the estimated asymptotic SEs over 100 replications for Data 1 and 2, and the mean SEs of linked $\hat{\theta}$s. These latter values were obtained for a pair of linking coefficient estimates, $\hat{A}$ and $\hat{B}$, by (Kolen & Brennan, 1995, Ch. 7)

$$\sqrt{\int_{-\infty}^{+\infty} \widehat{\text{avar}}\left(\hat{A}\theta + \hat{B}\right)\left(1/\sqrt{2\pi}\right)\exp\left(-\theta^2/2\right)d\theta} = \sqrt{\widehat{\text{avar}}\left(\hat{A}\right) + \widehat{\text{avar}}\left(\hat{B}\right)}$$

$$= \sqrt{\left[\widehat{SE}\left(\hat{A}\right)\right]^2 + \left[\widehat{SE}\left(\hat{B}\right)\right]^2}. \quad (25)$$

The mean SE summarizes the SE magnitudes for $\hat{A}$ and $\hat{B}$ in a single value. SDs in Table 1 for Equation 25 were obtained from

$$\sqrt{\left[\text{SD}\left(\hat{A}\right)\right]^2 + \left[\text{SD}\left(\hat{B}\right)\right]^2}, \quad (26)$$

which is a function of two empirical SEs (SDs of parameter estimates from the 100 replications). Mean of SE and SD of SE in Table 1 are the mean and SD of the 100 values of the estimated SEs.

For the 2PLM, Table 1 shows that most of the theoretical SEs were close to simulated values; some theoretical values were somewhat underestimated. The SDs of SE were rather small, indicating a stable property of the estimated SEs. The GLS estimates (mean = 1.156 and 1.146 for true A = 1.2, and mean = .481 for true B = .5) were clearly biased, although their theoretical and simulated SEs were the smallest. The ULS estimates were opposite—small bias with large SEs. That is, mean $\hat{A}_{\text{ULS}} = 1.214$ and 1.200 for Data 1 and 2, respectively, and mean SE for $\hat{\theta}_{\text{ULS}}$ were .094 and .083 with SD = .093 and .093—the largest among those given by the least squares methods. The WLS estimates with $m = 1$ (standard WLS) were biased, but less so than the GLS estimates. The WLS estimates with $m = .5$ were fairly successful—the overall biases were reduced without increasing the SEs to a large extent. That is, for Data 1, mean $\hat{A}_{\text{W1}} = 1.177$ and mean $\hat{B}_{\text{W1}} = .489$, and mean $\hat{A}_{\text{W.5}} = 1.194$ and $\hat{B}_{\text{W.5}} = .495$. For Data 2, mean $\hat{B}_{\text{W.5}}$ did not show reduced bias, but mean $\hat{A}_{\text{W.5}}$ = 1.184 showed reduced bias compared with mean $\hat{A}_{\text{W1}} = 1.168$. The values of SD and mean of SE were similar for the two methods.

Table 2 shows the means of the estimated asymptotic correlations among the six coefficients from ULS, GLS, and WLS for Data 1. It also includes correlations calculated from the 100 simulated parameter estimate sets. The estimates given by the three methods were highly correlated within corresponding pairs of coefficients (e.g., A_{ULS} and $A_{\text{W.5}}$).

Table 3 shows the results for the 3PLM. Information about SEs was available, but information about biases was not. The relative magnitudes of SEs among the six methods were similar to those in Table 1, although the absolute differences were more pronounced in Table 3. The SDs of SE in Table 3 are approximately the same as those in Table 1. However, the SDs of SE in Table 3 do not include the variation due to using estimated item parameters. The SEs for the M/M and ULS methods in Data 4 were substantially smaller than those in Data 3. The SEs of $\hat{\theta}$ (and, therefore, the SEs of $\hat{A}$ and $\hat{B}$) for the TRF and GLS methods in Data 4 were similar to those in Data 3.

Real Data

The real data example was taken from Kolen & Brennan's (1995, Appendix B) data. This dataset had two tests, Test X and Test Y, each containing 36 items (12 internal common items and 24 unique items). 1,655 examinees took Test X and 1,638 took Test Y.

Method. The method for this dataset was the same as for the simulated datasets. The linking coefficients were defined in two ways, from Test X to Test Y (X → Y) and from Test Y to Test X

Table 1
Mean (M) and SD of Linking Coefficients (A and B) and Equated Scores ($\hat{\theta}$), and Their SEs, for the 2PLM From Six Equating Methods and Two Datasets

	Data 1				Data 2			
			SE				SE	
Method	M	SD	M	SD	M	SD	M	SD
M/M								
A_{M}	1.203	.062	.062	.003	1.192	.056	.055	.003
B_{M}	.497	.080	.076	.005	.507	.071	.062	.002
$\hat{\theta}_{M}$		.101	.098	.006		.091	.082	.003
TRF								
A_{TRF}	1.202	.061	.061	.003	1.191	.051	.050	.002
B_{TRF}	.499	.065	.062	.002	.509	.068	.058	.002
$\hat{\theta}_{TRF}$		.089	.087	.004		.085	.077	.003
ULS								
A_{ULS}	1.214	.064	.066	.005	1.200	.054	.054	.003
B_{ULS}	.505	.068	.069	.003	.513	.076	.063	.002
$\hat{\theta}_{ULS}$		.093	.094	.006		.093	.083	.003
WLS[a]								
$A_{W.5}$	1.194	.061	.062	.003	1.184	.053	.053	.002
$B_{W.5}$	.495	.065	.063	.002	.507	.071	.060	.002
A_{W1}	1.177	.061	.060	.003	1.168	.051	.052	.002
B_{W1}	.489	.065	.061	.002	.503	.069	.059	.002
$\hat{\theta}_{W.5}$		.089	.088	.004		.088	.080	.003
$\hat{\theta}_{W1}$		.089	.086	.004		.086	.079	.003
GLS								
A_{GLS}	1.156	.061	.059	.003	1.146	.048	.048	.002
B_{GLS}	.481	.064	.060	.002	.489	.066	.056	.002
$\hat{\theta}_{GLS}$		.088	.084	.004		.082	.074	.003

[a]W.5 indicates $m = .5$; W1 indicates $m = 1$.

(Y → X). Because it was difficult to obtain parameter estimates using MML without restrictions on parameters in the 3PLM, BILOG estimates (Kolen & Brennan, 1995, Table 6.5) were used.

Results. Table 4 shows results for both the 2PLM and the 3PLM. The magnitudes of the estimates of $\hat{A}$ and their corresponding SEs decreased for ULS, WLS ($m = .5$ and 1), and GLS estimation methods. The relative relationships among the results of the six methods generally were similar to those in

Table 2
Correlations Between Linking Coefficient Estimates for Data 1 (Lower Triangle), and Theoretical Values (Upper Triangle)

Method	A_{ULS}	B_{ULS}	$A_{W.5}$	$B_{W.5}$	A_{GLS}	B_{GLS}
A_{ULS}	—	.31	.98	.30	.89	.30
B_{ULS}	.37	—	.32	.97	.32	.89
$A_{W.5}$	.98	.35	—	.34	.96	.35
$B_{W.5}$	.34	.97	.34	—	.36	.96
A_{GLS}	.90	.30	.95	.33	—	.39
B_{GLS}	.31	.89	.33	.96	.34	—

Table 3
Mean (M) and SD of SEs of Linking Coefficients (A and B) and Equated Scores ($\hat{\theta}$) for the 3PLM From Six Equating Methods and Two Datasets

	Pop.	Data 3		Data 4	
Method	Value	M	SD	M	SD
M/M					
A_{M}	1.2	.101	.003	.086	.003
B_{M}	.5	.155	.009	.104	.004
$\hat{\theta}_{\mathrm{M}}$		.185	.008	.135	.004
TRF					
A_{TRF}	1.2	.068	.003	.071	.004
B_{TRF}	.5	.065	.001	.065	.001
$\hat{\theta}_{\mathrm{TRF}}$		.094	.003	.096	.004
ULS					
A_{ULS}	1.2	.109	.003	.095	.004
B_{ULS}	.5	.144	.006	.112	.004
$\hat{\theta}_{\mathrm{ULS}}$		.181	.006	.147	.004
WLS[a]					
$A_{\mathrm{W.5}}$	1.2	.097	.003	.086	.003
$B_{\mathrm{W.5}}$	.5	.090	.002	.078	.002
A_{W1}	1.2	.095	.002	.084	.002
B_{W1}	.5	.081	.002	.072	.002
$\hat{\theta}_{\mathrm{W.5}}$		.132	.003	.116	.003
$\hat{\theta}_{\mathrm{W1}}$		.125	.003	.111	.002
GLS					
A_{GLS}	1.2	.081	.002	.077	.002
B_{GLS}	.5	.074	.002	.074	.002
$\hat{\theta}_{\mathrm{GLS}}$		.110	.002	.104	.002

[a]W.5 indicates $m = .5$; W1 indicates $m = 1$.

Tables 1 and 3. Because most of the c parameter estimates in this dataset were substantially different from zero, the SEs for the 2PLM in Table 4 might be different from those when the model is not true. The accurate asymptotic SEs of the parameter estimates given by applying the 2PLM when the 3PLM holds might be somewhat larger than those in Table 4. Therefore, the SEs for the 2PLM in Table 4 should be taken as lower bounds.

Discussion and Conclusions

None of the six methods—M/M, TRF, ULS, WLS ($m = .5$), WLS ($m = 1$), and GLS—were best in both stability and unbiasedness. GLS estimation had the smallest SEs, which was expected from theory. However, GLS also showed the largest biases when information about bias was available. In the 2PLM, the differences in the magnitudes of the SEs among the four best methods—TRF, WLS ($m = .5$), WLS ($m = 1$), and GLS—were rather small in both the simulated and real data (i.e., within .003 for the mean SEs in Tables 1 and 4). On the other hand, GLS and WLS ($m = 1$) had the first and second largest biases, respectively, which were not negligible. Overall, the TRF method seemed to have the best results, and WLS ($m = .5$) was second best. Similar, but more pronounced results, were obtained for stability in the 3PLM.

As indicated above, it is usual to use prior distributions or restrictions for estimating item parameters for the 3PLM. For such cases, the asymptotic SEs of parameter estimates and, consequently,

Table 4
Linking Coefficients Estimates (A and B), Their SEs, and Equated $\hat{\theta}$s for the 2PLM and 3PLM From Six Equating Methods for Kolen & Brennan's (1995) Data

	2PLM				3PLM			
	X → Y		Y → X		X → Y		Y → X	
Method	Est.	SE	Est.	SE	Est.	SE	Est.	SE
M/M								
A_{M}	.956	.041	1.046	.045	1.217	.108	.822	.073
B_{M}	−.355	.078	.371	.079	−.557	.265	.458	.196
$\hat{\theta}_{\mathrm{M}}$		.088		.091		.286		.209
TRF								
A_{TRF}	.944	.040	1.049	.045	1.091	.052	.935	.046
B_{TRF}	−.370	.051	.389	.051	−.496	.061	.449	.045
$\hat{\theta}_{\mathrm{TRF}}$		.065		.067		.080		.064
ULS								
A_{ULS}	.975	.047	1.047	.047	1.180	.114	.876	.106
B_{ULS}	−.378	.058	.392	.055	−.572	.245	.434	.245
$\hat{\theta}_{\mathrm{ULS}}$		.075		.073		.270		.267
WLS[a]								
$A_{\mathrm{W.5}}$	.946	.041	1.035	.045	1.176	.098	.827	.069
$B_{\mathrm{W.5}}$	−.363	.053	.379	.052	−.593	.159	.547	.092
A_{W1}	.922	.040	1.016	.044	1.169	.094	.795	.064
B_{W1}	−.351	.052	.362	.050	−.557	.141	.578	.080
$\hat{\theta}_{\mathrm{W.5}}$		.067		.068		.187		.115
$\hat{\theta}_{\mathrm{W1}}$		.065		.067		.169		.103
GLS								
A_{GLS}	.895	.038	1.004	.042	1.129	.083	.785	.058
B_{GLS}	−.371	.051	.398	.048	−.556	.124	.602	.074
$\hat{\theta}_{\mathrm{GLS}}$		.064		.064		.149		.094

[a]W.5 indicates $m = .5$; W1 indicates $m = 1$.

the SEs of the linking coefficients might be reduced, depending on the strength of the prior distributions or restrictions employed. That is, formulas for the SEs in the 3PLM might be used as upper bounds of the SEs when there are constraints on parameters. The actual SEs of the estimated linking coefficients in the 3PLM with some constraints on parameters are expected to be located between those for the 2PLM and 3PLM without restrictions on parameters. In such intermediate cases, the TRF method would be expected to provide the best results.

One advantage to using least squares estimation is that asymptotic SEs of the estimated linking coefficients are easily derived from the process of linking coefficient estimation (e.g., Equations 11 and 13). However, the method for deriving the asymptotic SEs of the linking coefficient estimates by the TRF method (Ogasawara, 2001) is somewhat involved.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35,* 179–197.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9,* 413–415.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park CA: Sage.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer.

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

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25,* 53–67.

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

Software

The computer program used to estimate the linking coefficients and their asymptotic SEs by the various least squares methods was written in FORTRAN90. The source code can be obtained from the author.

Acknowledgments

The author thanks Michael J. Kolen for access to the real data used in this article.

Author's Address

Send requests for reprints or further information to Haruhiko Ogasawara, Otaru University of Commerce, 3-5-21, Midori, Otaru 047-8501, Japan. Email: hogasa@res.otaru-uc.ac.jp.

Applied Psychological Measurement Inc.

Announces a Program of Grants for Graduate Students in Psychological and Educational Measurement Programs

Purpose

Grants of up to $750 are available to support costs of psychological and educational measurement research that is part of the graduate student's training. Priority will be given to grants that will be used for purchase of computer software, purchase of computer hardware, purchase of specialized books, and other research expenses. The program may also fund travel to meetings or conventions to deliver an accepted paper.

Qualifications

- Applicant must be currently registered in a Ph.D. or M.A. program with a concentration in psychological or educational measurement at an accredited university.
- Applicant's status must be confirmed by a faculty member in the same program at the applicant's institution.
- Applications will be accepted from applicants worldwide.

Terms of Award

- Maximum award is $750 ($500 for travel originating and terminating in the U.S.).
- Priority will be given to requests that include matching funds from the applicant's university.
- Priority will be given to applicants who have not received a prior award from this program.
- Applicants who have received prior awards are ineligible for additional awards for a period of twelve months.
- Applicants are required to submit a report on the expenditure of funds within 60 days of receipt of an award, including copies of receipts for eligible expenditures, and to refund unspent amounts of $25 or more.
- The number of awards made per year is at the discretion of the Awards Committee (approximately 15–20 awards are anticipated per year, depending on investment experience).
- Awards are made on a quarterly basis. Completed applications received by the end of each calendar quarter (March 31, June 30, September 30, and December 31) will be acted upon within 30 days.

Applications and further information are available from djweiss@umn.edu.

Defining Error Rates and Power for Detecting Answer Copying

James A. Wollack, Allan S. Cohen, Ronald C. Serlin
University of Wisconsin

A familywise approach is described for evaluating the significance of copying indices designed to hold the Type I error rate constant for each examinee. The empirical Type I error rate and power of two indices, ω (Wollack, 1997) and g_2 (Frary, Tideman, & Watts, 1977), are examined under a variety of copying situations. Results indicated that the traditional pairwise approach falsely detected examinees almost three times more often than the nominal α level. Familywise Type I error rates were substantially smaller, although they also tended to be somewhat inflated at small α levels as the percentage of items copied increased. Eliminating the indices detecting a source from the copier, in situations where the copier was also detected from the source, helped control the familywise Type I error rates for all $\alpha \geq .001$. Lack of Type I error control meant power could not be evaluated for g_2 under any of the simulated familywise conditions. Familywise power for ω was reasonable when at least 30% of the items were copied. *Index terms: answer copying, cheating, item response theory, planned comparisons, power, Type I error rate.*

In research on the detection of answer copying, Type I error rates and power typically have been determined between all possible pairs or between all examinees for whom copying was possible (e.g., according to a seating chart). The usual approach in such studies (e.g., Bay, 1995; Bellezza & Bellezza, 1989; Chason, 1997; Frary, Tiedeman, & Watts, 1977; Wollack, 1997) has been to calculate the percentage of incorrect detections (i.e., the Type I error rate) among the noncopying pairs and the percentage of correct detections (i.e., power) among the true copier-source pairs at a particular Type I error rate (α). However, this pairwise approach can result in examinees having markedly different probabilities of being detected incorrectly, depending on the number of potential source papers in their vicinity. That is, during any group-administered test, examinees sitting around the perimeter of the room have fewer opportunities to copy than examinees sitting in the middle of the room, simply because of the number of examinees sitting around them. If an examinee is assumed to be able to copy answers from other examinees seated to his/her left, right, front left, front center, and front right, it is clear that the examinee in the front left corner of the room, for example, can copy only from one examinee, whereas an examinee in the middle of the room can copy from up to five examinees. If the pairwise Type I error rate is computed, noncopying examinees sitting in the interior of the class will have five answer-copying indices computed for them, but the examinee sitting in the front left corner will have only one such index computed. Thus, if a pairwise definition of Type I error is used, the examinees sitting in the interior of a testing room will be approximately five times more likely to be incorrectly identified than examinees with only one potential source paper.

Testing each pairwise comparison at α results in examinees having different probabilities of being incorrectly identified as a potential copier, depending on where they are seated in the room. An examinee with multiple source papers who is identified as a potential copier legitimately would

Applied Psychological Measurement, Vol. 25 No. 4, December 2001, 385–404

be able to argue that the statistical analysis was biased. A more appropriate definition of the Type I error rate would produce identical probabilities of false detection for each examinee. Such an approach would be fairer, because it would hold the Type I error rate at α for the set (i.e., "family") of contrasts for a given examinee. This "familywise" approach would have the effect of ensuring that all examinees have the same probability of being incorrectly identified as a potential copier. In addition, because power is defined as a function of the Type I error rate, the definition of power must be consistent with that of Type I error. In this paper, different approaches to determining the Type I error rate and power are examined, and the extent of their differences is clarified through simulation.

Background

Within the answer-copying framework, a family is the set of comparisons in which the same examinee (i.e., the target examinee) is tested as a potential copier (C) against each potential source examinee (S). Each classroom has N families—one for each examinee. The number of comparisons within each family differs depending on the number of potential source papers sitting around the target examinee.

The concept of a family is not new to answer-copying research. Buss & Novick (1980) argued that a multiple-comparisons procedure was necessary in this context to maintain Type I error control. Also, Bay (1995), Bellezza & Bellezza (1989), and Wollack (1996) recognized that testing each pair at α would result in too many Type I errors. To compensate, Bellezza and Bellezza adopted a pseudo-Dunn correction, testing each contrast at α/cs, where c is the maximum number of people from whom any examinee could copy, and s is the number of examinees in the room. Bay defined a Type I error as "wrongly accusing a student of cheating" (p. 3), and recommended holding this error rate at α for each examinee by using Sidak's (1967) approach of testing each contrast at $\alpha^* = 1 - \sqrt[k]{1-\alpha}$, where k is "the number of examinees from whom [the potential copier] could have copied" (Bay, p. 4). However, Wollack (1996) noted that "if a constant alpha level is used, . . . the resultant proportion of falsely accused examinees may well exceed the specified alpha level" (pp. 91–92). He suggested using the step-up Hochberg (1988) approach to test each comparison within the family at a different α level, such that the overall Type I error rate for the family is held at α. Although these researchers acknowledged the need for familywise control, the ways in which the α-splitting algorithms were implemented and the empirical Type I error rate and power determined were more consistent with pairwise—rather than familywise—control.

Bellezza & Bellezza (1989) and Bay (1995) used formulas that allow for the contrasts to be tested at different α levels, depending on the number within each family. Even so, they both chose to treat the size of an examinee's family as a constant. Bellezza and Bellezza used $c = 3$, the maximum number of people from whom an examinee could copy if multiple test forms were used; Bay used $k = 4$, the average number of people from whom any examinee could copy. Consequently, each contrast was tested with the same critical value (rather than adjusting the critical value to reflect the number of comparisons within a family). The result is that examinees sitting around many others still had a higher probability of being incorrectly identified than examinees sitting around few others. Further, despite using planned comparison procedures that were seemingly consistent with the notion of familywise error control, Bellezza and Bellezza, Bay, and Wollack (1996) all chose to compute Type I error rates and power using a pairwise definition.

The distinction between a pairwise definition and a familywise definition lies in whether each individual contrast or the family of contrasts for a given examinee is the unit of analysis. Under a pairwise definition, error control is maintained at the level of individual pairs being inspected

for possible copying. Under a familywise definition, error control is maintained at the level of the examinee.

There is no need to adopt an α-splitting technique (i.e., multiple comparisons) under a pairwise definition, unless there is reason to believe that the individual examinee pairs are not independent of one another. All three of the above studies regarded copying indices as nondirectional (i.e., a significant test statistic implicates the pair of examinees, but does not identify which examinee is C and which is S). This means that the pair—regardless of which examinee was the source—was the unit of analysis. In this situation, the distinct pairs of examinees likely are independent of one another, and no α-splitting procedure should be used for the pairwise definition. Nevertheless, in all three studies, the researchers adopted an α-splitting technique designed to make the probability of committing a Type I error equivalent for all examinees (i.e., a familywise definition). However, they proceeded to determine the Type I error rate according to a pairwise definition. More specifically, Bay (1995) and Wollack (1996) counted the number of Type I errors (found using a multiple-comparisons procedure to increase the critical value), but divided them by the total number of pairs for whom copying did not occur. Using real data in which the true copiers (if any) were unknown, Bellezza & Bellezza (1989) compared the percentage of significant pairs (again using an adjusted critical value) to the percentage of pairs for whom copying was possible.

The confusion between operational and theoretical definitions of the Type I error rate and power is due largely to the difficulty in conceptualizing and operationalizing error rates at the family level. Pairwise analysis is relatively easy to implement, whereas familywise analysis requires using a method of maintaining the familywise Type I error rate at the nominal level. In addition, there are at least two definitions of power that are consistent with the familywise definition of error and are appropriate for investigating answer copying.

Familywise Error Control

Pairwise Versus Familywise Control

Traditionally, researchers have adopted a pairwise definition (Angoff, 1974; Bellezza & Bellezza, 1989; Cody, 1985; Frary et al., 1977; Hanson, Harris, & Brennan, 1987; Wollack, 1997). By doing this, decisions about whether copying occurred are made by treating the pair as the primary unit of analysis. A copying index then is computed between each pair of examinees for whom copying is possible, and each comparison independently contributes to the Type I error rate and power. Type I error rate is computed by dividing the number of incorrectly identified pairs by the number of noncopying pairs. Power is computed by dividing the number of correctly identified pairs by the number of true C-S pairs. Unfortunately, as noted earlier, this approach results in different Type I error rates depending on where the examinee is seated.

In contrast, the definition used here for familywise error is such that the probability of being incorrectly identified as a copier is the same for all examinees in the room. In a familywise approach, the family is the unit of analysis—an index still is computed between all pairs for whom copying is possible, but all contrasts within a family contribute a single value toward calculating the Type I error rate and power. Type I error rate then is computed by dividing the number of incorrectly identified families by the total number of noncopying families, and power is computed by dividing the number of correctly identified families by the total number of simulated copiers.

Type I Error

Pairwise rate. The definition of pairwise Type I error rate is the probability of incorrectly identifying a pair of examinees. This definition requires that C and S be linked correctly; otherwise,

it is a Type I error. Obviously, detecting an examinee not involved in copying is a Type I error. However, under the above definition, correctly detecting a true C with the incorrect S is also a Type I error. Similarly, correctly detecting a true S with the incorrect C is a Type I error. In fact, if it is assumed that copying indices are directional, then detecting a true S as having copied from a true C is also a Type I error. If the indices had been assumed as nondirectional, this would have constituted a correct detection.

Familywise rate. Familywise Type I error rate is defined as the probability of identifying a noncopying examinee as a C. This differs from the pairwise Type I error rate in two important ways. First, a Type I error can be committed only within a family centering around a noncopying examinee. If the target examinee is a true C, however, it is not possible to commit a Type I error within that family. Given that the target examinee is a noncopier, any statistically significant contrast within that family means that a Type I error was committed. Within each family, a Type I error either was or was not committed. Multiple Type I errors in a family count only once as a single error. Therefore, the familywise Type I error rate is computed as the number of families in which at least one Type I error was committed, divided by the total number of noncopying examinees. The index is directional for the familywise definition of Type I error.

Power

Pairwise. The pairwise definition of power is the probability of correctly identifying a true C-S pair. Again, both C and S must be identified correctly. In the event that an examinee has copied from multiple Ss, each pair individually contributes toward the power. This definition is directional and is consistent with the pairwise definition of a Type I error.

Familywise. The familywise definition of power is the probability of correctly identifying a true C. This definition is also directional and consistent with the familywise definition of a Type I error. However, there are two different ways to examine the familywise power that are relevant to answer-copying research:

1. Per-pair power (Einot & Gabriel, 1975; Kirk, 1995; Seaman, Levin, & Serlin, 1991) can be thought of as a type of average power in which each C-S pair individually contributes to power. Per-pair and pairwise power are identical—both require the correct C and S to be identified, and each correctly identified pair counts toward power.
2. Family-level power differs from per-pair power in that it does not require that a true S be identified, provided the target examinee is a true C.

Copying Scenarios

To understand the different ways of computing Type I error rate and power, possible copying scenarios (and patterns of detection) are presented and their effects on the various Type I error rates and power are examined. Table 1 provides ten different answer-copying scenarios and presents the contributions to the Type I error rate and power under pairwise and familywise definitions for each scenario. In Scenarios 1–3, N is the target (noncopying) examinee. Scenarios 4–10 pertain to situations in which copying occurred. Here, C identifies the true C and S the true S(s). Other examinees are identified by number. The arrows are used to show the existence and direction of a statistically significant copying index. The absence of an arrow between a pair of examinees means that the index calculated for that pair is nonsignificant.

In Scenario 2, under "Type I error" for pairwise control in Table 1, 1/5 indicates that one pairwise Type I error was committed—N was erroneously detected as copying from 5—and five such errors were possible. Under "power" for pairwise control, the contribution of each family toward power is shown (the ratio of true C-S pairs to the number of true C-S pairs detected in the family). Recall

Table 1
Type I Error and Power Scenarios for a Noncopying Target Examinee (N), a True Copier (C), and a Source Examinee (S)

			Pairwise Control		Familywise Control		
						Power	
Scenario Number	Copying Situation		Type I Error	Power	Type I Error	Per Pair	Familywise
1	2 3 4 1 N 5		0/5	—[a]	0	—	—
2	2 3 4 1 N → 5		1/5	—	1	—	—
3	2 3 4 1 N → 5		2/5	—	1	—	—
4a	2 3 4 1 C ← S		0/4	0/1	—	0/1	0
4b	3 4 C ← S	6 7	1/5	—	1	—	—
5	2 3 4 1 C S		0/4	0/1	—	0/1	0
6	2 3 4 1 C → S		0/4	1/1	—	1/1	1
7	2 3 S_2 1 C → S_1		0/3	2/2	—	2/2	1
8	2 3 S_2 1 C → S_1		0/3	1/2	—	1/2	1
9	2 3 4 1 C → S		1/4	1/1	—	1/1	1
10	2 3 4 1 C S		1/4	0/1	—	0/1	1

[a]Indicates that result was not applicable.

that there will be no contribution to the power if the target examinee is a noncopier. Similarly, familywise Type I errors are not possible if the target examinee is a C.

Most of the scenarios are straightforward. For example, in Scenario 8, there are two possible S papers for C. Because the target examinee is a C, it is not possible to commit a familywise Type I error. Of the three nonsource papers, none are detected, yielding a pairwise Type I error rate of 0/3. Of the two Ss, only one is detected. Consequently, the pairwise and per-pair power are both 1/2. The familywise power is 1 because a true C was detected from at least one family member (S_1).

Some scenarios are not as straighforward. In Scenario 4a, C copies from S. If the copying indices properly are assumed to be directional, the copying is undetected. A loss in power is recorded, and the pairwise Type I error is 0/4. However, because each examinee is inspected individually as a target examinee, the family for which S is the target examinee (shown in 4b) is also of interest. Because S (a noncopying examinee) is identified as a C, a Type I error has occurred. Previous research in which copying indices have been treated as nondirectional (Bay, 1995; Frary et al., 1977; Wollack, 1997; Wollack & Cohen, 1998) would count identifying S copying from C as power.

In Scenario 10, the target examinee is a C. The true S goes undetected, but C is detected as copying from Examinee 4. Under a pairwise framework, this constitutes a Type I error and a loss of power. Using a familywise approach, it does not count as a Type I error, because the target examinee is a C. The per-pair power is 0, because the true C-S pair was not detected. However,

because a true C was identified without regard to whether the correct S was identified, the familywise power is 1.

Controlling Familywise Type I Error Rates

Three procedures are discussed here for holding the Type I error rate at level α for each family: Dunn's (1961), Sidak's (1967), and Holm's (1979).

Dunn's (1961) Procedure

Dunn's (1961) approach is computationally simple. Dunn found that testing each comparison with $\alpha = \alpha_{FW}/k$ yields a familywise Type I error rate that is less than or equal to α_{FW}, according to the Bonferroni inequality. All contrasts within a family are tested with a common α, but the per-contrast α might be different in other families, depending on k.

Sidak's (1967) Procedure

A multiplicative inequality (Sidak, 1967) indicates that, if a set of k independent contrasts each is tested with α, the probability of committing at least one Type I error among that set is no greater than α_{FW}, where $\alpha_{FW} = 1 - (1 - \alpha)^k$. Consequently, to hold α_{FW} constant for all families, each contrast within its family could be tested at $\alpha = 1 - \sqrt[k]{1 - \alpha_{FW}}$. Within a particular family, each contrast is tested at the same α level, but between families, the contrasts can be tested at different α levels, depending on k.

Holm's (1979) Procedure

Holm (1979) developed a modification of the Dunn procedure in which the p values of the test statistics within a family are rank-ordered from smallest to largest. The contrast with the smallest p value is tested for statistical significance using $\alpha = \alpha_{FW}/k$, the same α level used in the Dunn test. If the smallest is found to be nonsignificant, all other contrasts are considered nonsignificant as well. However, if it is significant, the contrast with the second smallest p value is then tested with $\alpha = \alpha_{FW}/(k-1)$. This process continues by testing the contrast with the next smallest p using $\alpha = \alpha_{FW}/(k-2)$, and so forth, until one contrast is nonsignificant or the kth contrast is tested using α. Holm's sequential, step-down procedure has been shown to hold the Type I error rate at α_{FW} for a set of contrasts. Further, because the α level used for testing a contrast for significance becomes larger as contrasts continue to be statistically significant, the Holm procedure offers more power than either Dunn's or Bonferroni's, once the initial contrast has been found to be significant.

Adjacency Control

One of the reasons that copying indices have been used nondirectionally is that there is reason to believe that the indices computed for Person A copying from Person B and Person B copying from Person A are not independent of one another, particularly if one of the examinees copied from the other. For example, if Person A copied 10 items from Person B, the index for Person A copying from Person B would be expected to be inflated. However, in computing the index for Person B copying from Person A, that index might also be expected to be inflated because they have identical responses on those 10 items. This problem exists only for adjacent pairs (i.e., pairs of examinees sitting beside each other), because the reciprocal index for Person B copying from Person A will not be computed if they are a front-back pair (i.e., Person B is sitting in front of Person A).

To account for the fact that examinees are more likely to be identified as Cs if they are Ss, an adjacency control can be used in which adjacent pairs are held to a more stringent criterion than front-back pairs. The most straightforward approach would be to use an additional Dunn correction on adjacent pairs. Therefore, each contrast between adjacent pairs would be tested for statistical significance at $\alpha^*/2$, where α^* is the α level that would have been used for that contrast (after any other corrections had already been made), were it not an adjacent pair.

Method

The purpose of this study was to investigate the impact of using a pairwise versus a familywise definition of answer copying on the Type I error rate and the power of statistical indices to detect answer copying. These statistical properties were examined as a function of test length, sample size, percent of items copied, and α level.

Data

Nominal response model (Bock, 1972) item parameters were taken from Wollack (1997) for an 80-item, 5-alternative college English placement test and a 40-item, 5-alternative college mathematics placement test. Item responses were generated for the two test lengths and samples of $N =$ 100 and 500 examinees. The computer program GENIRV (Baker, 1986) was used to generate simulated item responses for the nominal response model. The latent trait (θ) was distributed normal (0, 1). Examinees were randomly assigned to seats within the room. A different seating chart was used for each dataset within each condition.

Item Selection

Only random-strings copying was considered here. Hanson et al. (1987) suggested that this is the most frequent type of copying. Answer copying was simulated in strings of 4 consecutive items by replacing C's answers with S's answers. Tests were divided into 10 or 20 strings of 4 consecutive items (for 40- and 80-item tests, respectively), such that each string had an equal probability of being copied.

Within each test length $\times$ sample size condition, four different percentage levels of items copied were simulated, such that C copied 10%, 20%, 30%, or 40% of the items. All Cs in a particular condition were simulated to copy the same percentage of items. As a result, there were 2 (test length) $\times$ 2 (sample size) $\times$ 4 (percentage of items copied) $=$ 16 conditions examined.

Copier Selection

Within each dataset, 5% of the examinees were randomly selected to be Cs. For each examinee selected to be a C, an S examinee was randomly selected among the examinees seated within copying distance of C, subject to the constraint that a C could only copy answers from an S if $\theta_S > \theta_C$. If there were no potential Ss satisfying this constraint, then no copying was simulated for this examinee and another examinee was randomly selected.

α Levels

Type I error rate and power were evaluated at seven different nominal α levels: .00001, .0001, .0005, .001, .005, .01, and .05.

Replications

Within each test length × sample size × percent of copying condition, a total of 2,000 datasets were generated. In the 100-examinee conditions, this yielded a pairwise Type I error rate based on 854,000 pairs; a familywise Type I error rate was based on 190,000 families. Familywise and pairwise power were based on 10,000 Cs. In the 500-examinee condition, the pairwise Type I error rate was based on 4,644,000 pairs, the familywise Type I error rate was based on 950,000 families, and both familywise and pairwise power were based on 50,000 Cs.

Copying Indices

Each dataset was analyzed using two answer-copying detection indices: g_2 (Frary et al., 1977) and ω (Wollack, 1997). g_2 and ω are computed similarly. The generic form of both indices is:

$$\frac{h_{\mathrm{CS}} - \sum_{i=1}^{n} P_{\mathrm{C}}(u_{i\mathrm{S}})}{\sqrt{\sum_{i=1}^{n} [P_{\mathrm{C}}(u_{i\mathrm{S}})]\,[1 - P_{\mathrm{C}}(u_{i\mathrm{S}})]}}, \tag{1}$$

where

C is the examinee being inspected as a possible copier,

S is the examinee being inspected as a possible source,

h_{CS} is the number of items answered identically by C and S, and

$P_{\mathrm{C}}(u_{i\mathrm{S}})$ is the probability of C selecting S's answer to item i.

The difference between ω and g_2 is in how $P_{\mathrm{C}}(u_{i\mathrm{S}})$ is computed. For g_2, Frary et al. (1977) considered the classical test theory item and distractor difficulties, as well as the ratio of C's number-correct score to the mean number-correct score for all examinees. For ω, Wollack (1996, 1997) used the nominal response model (Bock, 1972). Both indices are assumed to have standard normal sampling distributions.

Although g_2 has been used more than any other copying index in the literature (Bay, 1995; Chason, 1997; Chason & Maller, 1996; Hanson et al., 1987; Iwamoto, Nungester, & Luecht, 1996; Iwamoto, Nungester, Watson, & Luecht, 1997; Iwamoto, Watson, Nungester, & Luecht, 1997; Wollack, 1997), recent evidence has shown that it does not hold the nominal Type I error rate when computed using a pairwise definition (Wollack, 1997). ω has been found to hold the pairwise Type I error rate at its nominal level, regardless of whether item parameters are known (Chason, 1997; Wollack, 1997) or estimated from the data (Wollack & Cohen, 1998).

Here, item and θ parameters were estimated separately for each simulated sample within each condition using the computer program MULTILOG (Thissen, 1991).

Detection of Type I Error Rates and Power

In this study, g_2 and ω were treated directionally. That is, the two indices computed for adjacent pairs were regarded as distinct. It is important to note that g_2 and ω typically have been used as nondirectional indices (Frary et al., 1977, and Wollack, 1997, respectively). However, regarding these indices as directional, thereby assuming that the reciprocal indices computed among adjacent pairs are statistically independent of one another, would seem to be more consistent with a familywise definition.

Type I error rate. The significance of g_2 and ω was tested at the one-tailed critical value corresponding to the upper α on the normal curve. Seven different Type I error rates were computed, differing with respect to their definition (pairwise or familywise) and their level of statistical control.

Pairwise Type I error rates. The two pairwise Type I error rates were computed as the number of noncopying pairs that were detected divided by the total number of noncopying pairs. The first pairwise Type I error rate was evaluated using no control of α (PW-NC), by testing every pair with full α. The second Type I error rate was evaluated using the Dunn procedure (PW-D) to test each contrast with α/k (where k is the number of pairs in the family).

Familywise Type I error rates. The five familywise Type I error rates were computed as the number of noncopying families that were detected divided by the total number of noncopying families. The first familywise Type I error rate was evaluated as using no control (FW-NC) and the second used Dunn control (FW-D). A third Type I error rate was evaluated using a Holm (1979) control (FW-H) so that, within a family, if the contrast with the smallest p value (not yet tested) was statistically significant, the contrast with the next smallest p value (not yet tested) was evaluated using a slightly higher p level.

A fourth familywise Type I error rate was evaluated using the Dunn procedure with adjaceny control (FW-DA) and a fifth used the Holm procedure with adjaceny control (FW-HA). Under adjacency control, the p values of all indices computed between adjacent pairs (as opposed to front-back pairs) were doubled prior to applying the Dunn or Holm modifications. For the DW-HA procedure, this was equivalent to testing those adjacent pairs at $\alpha/2k$. For the FW-HA procedure, this meant that adjaceny control was applied before the p values were rank-ordered.

The Type I error rates for FW-D and FW-H (and for FW-DA and FW-HA) will be identical. Because each noncopying family contributes either a 0 or 1 toward the familywise Type I error rate regardless of how many pairs were statistically significant, whether the contrast with the smallest p value is significant is all that has to be considered. Because this contrast will be the first tested, the stepwise nature of the Holm procedure will not have begun, so this contrast will be tested for significance using the same critical value for the Dunn and Holm procedures (and for FW-DA and FW-HA). Consequently, for the purpose of reporting familywise error rates, the Dunn-Holm conditions will be referred to as FW-DH and the Dunn-Holm conditions with adjaceny control will be referred to as FW-DH-A.

Power. Three different types of power were considered: pairwise, familywise per-pair, and family-level. Pairwise power was calculated for both pairwise Type I error rate conditions (PW-NC and PW-D), and both familywise per-pair power and family-level power were computed for all five familywise Type I error rate conditions (FW-NC, FW-D, FW-H, FW-DA, and FW-HA).

For the pairwise case, power was computed as the number of true C-S pairs detected divided by the total number of true C-S pairs. Familywise per-pair power was calculated as the sum over all families of the number of true C-S pairs detected per family divided by the total number of Cs. Note that, although per-pair power might differ conceptually from pairwise power, they are identical. Family-level power was computed as the total number of true Cs who were detected (even if from an incorrect S) divided by the total number of true Cs.

The family-level power for FW-D and FW-H (and FW-DA and FW-HA) will be identical. This is because with a family-level definition, the power issue becomes whether C was identified from at least one S (regardless of whether it is the correct S). Therefore, the only contrast of interest (to the family-level power) is the one with the smallest p value. If it is significant, then the Dunn and Holm conditions record a "hit." If it is nonsignificant, the Dunn and Holm conditions record a "miss." Identifying a significant second contrast (where Holm offers more power) is important for per-pair power, but not for family-level power.

Analysis

A regression technique identified the variables that had a significant impact on the results across all conditions. Type I error rate is known to be affected by α and by the type of planned comparison approach used, but the impact of sample size, test length, and percentage of items copied is less clear. Therefore, a two-stage hierarchical regression was performed. First, the dependent variable (either Type I error rate or power) was regressed on α and the type of planned comparison method. Next, sample size, test length, and percentage of items copied were entered into the regression equation. Data were collapsed across all nonsignificant variables to help identify meaningful patterns in the results and to eliminate noise due to having estimated Type I error rate and power under many different conditions. Separate hierarchical regressions were performed for pairwise and familywise definitions, and for ω and g_2.

Results

Empirical Type I Error Rates

Empirical Type I error rates were considered acceptable if they were less than or equal to Cochran's (1952) criterion for robustness. Specifically, power data were provided if the Type I error rate was:

less than .000015 for $\alpha = .00001$,
less than .00015 for $\alpha = .0001$,
less than .0006 for $\alpha = .0005$,
less than .0015 for $\alpha = .001$,
less than .006 for $\alpha = .005$,
less than .015 for $\alpha = .01$, and
less than .06 for $\alpha = .05$.

The average pairwise and familywise empirical Type I error rates are shown in Tables 2 and 3, respectively, for ω, and are combined in Table 4 for g_2. The regression analyses indicated that the percentage of items copied, α, and type of multiple comparison procedure affected the Type I error rate for ω. The percentage of items copied, α, and sample size were found to affect the Type I error rate for g_2.

In general, Tables 2 and 3 show that ω controlled the Type I error rate for PW-NC, FW-D, FW-H, FW-DA, and FW-HA when 10% or 20% of the items were copied. Type I error rates were inflated for 30% and 40% copied items. Pairwise and familywise Type I error rates for these procedures for g_2 (Table 4) were inflated in all conditions and at all α levels.

Pairwise Type I error rates. In evaluating the pairwise Type I error rates, the PW-NC rows in Table 2 are of most interest. Under a pairwise definition, all pairs are independent; therefore, no α control should be needed. When only 10% of the items were copied, ω demonstrated good pairwise error control at all α levels in the PW-NC condition (Table 2). However, as the percentage of items copied increased, the PW-NC Type I error rates of ω became inflated at smaller α levels. For example, for $\alpha = .0001$, the PW-NC Type I error rate increased from .00005 in the 10% condition to .00023 in the 20% condition, .00088 in the 30% condition, and .00201 in the 40% condition. When 20% of the items were copied, the PW-NC Type I error rate was inflated for $\alpha \leq .0005$. When 30% of the items were copied, the PW-NC Type I error rate was inflated for $\alpha \leq .001$. When 40% of the items were copied, the PW-NC Type I error rate was inflated for $\alpha \leq .005$. The PW-NC Type I error rate for g_2 was inflated in all conditions and at all α levels (Table 4).

Familywise Type I error rates. Type I error rates for the FW-NC conditions were expected to yield inflated Type I error rates, because most families involved computing multiple contrasts. The

Table 2
Type I Error Rates of ω for Pairwise Control
Averaged Across N and Test Length Conditions

% Copied	α Level						
and Method	.00001	.0001	.0005	.001	.005	.01	.05
10% Copied							
PW-NC	.00000	.00005	.00025	.00052	.00277	.00578	.03339
PW-D	.00000	.00001	.00006	.00011	.00058	.00119	.00646
20% Copied							
PW-NC	.00006	.00023	.00061	.00100	.00371	.00695	.03513
PW-D	.00003	.00010	.00026	.00038	.00112	.00191	.00774
30% Copied							
PW-NC	.00039	.00088	.00159	.00212	.00506	.00831	.03613
PW-D	.00023	.00054	.00094	.00119	.00227	.00316	.00907
40% Copied							
PW-NC	.00130	.00201	.00280	.00332	.00607	.00916	.03648
PW-D	.00080	.00133	.00186	.00217	.00324	.00409	.00967

Dunn and Holm conditions, however, with and without adjacency control, were expected to provide better control. The FW-NC Type I error rate for ω (Table 3) was inflated in all conditions and at all α levels. Across all conditions, the FW-NC Type I error rate was a median of 4.59 times larger than the corresponding pairwise error rate.

Table 3
Type I Error Rates of ω for Familywise Control
Averaged Across N and Test Length Conditions

% Copied	α Level						
and Method	.00001	.0001	.0005	.001	.005	.01	.05
10% Copied							
FW-NC	.00002	.00023	.00113	.00232	.01238	.02560	.13835
FW-DH	.00001	.00005	.00026	.00051	.00259	.00534	.02862
FW-DH-A	.00001	.00004	.00019	.00037	.00195	.00400	.02176
20% Copied							
FW-NC	.00026	.00100	.00274	.00449	.01652	.03072	.14515
FW-DH	.00012	.00045	.00112	.00170	.00501	.00856	.03420
FW-DH-A	.00009	.00031	.00082	.00125	.00370	.00645	.02639
30% Copied							
FW-NC	.00176	.00401	.00720	.00959	.02261	.03677	.14898
FW-DH	.00102	.00246	.00430	.00543	.01026	.01424	.04006
FW-DH-A	.00077	.00193	.00343	.00441	.00834	.01172	.03227
40% Copied							
FW-NC	.00596	.00915	.01267	.01500	.02710	.04051	.15025
FW-DH	.00450	.00705	.00943	.01081	.01553	.01923	.04357
FW-DH-A	.00388	.00621	.00836	.00955	.01375	.01687	.03640

The results of the familywise Type I error analysis were much like the results of the pairwise analysis. When only 10% of the items were copied, ω held the familywise Type I error rate at α in all conditions in which FW-D, FW-H, FW-DA, and FW-HA corrections were used. As the percentage of items copied increased, the control of the familywise Type I error rate decreased. Overall, adjacency control did have an important impact on the Type I error control. The median Type I error rates under adjacency control were 23% lower than their counterparts not using adjacency

Table 4
Type I Error Rates for g_2 Averaged Across Test Length and Percent of Items Copied

No. Examinees and Method	α Level						
	.00001	.0001	.0005	.001	.005	.01	.05
100 Examinees							
PW-NC	.00122	.00396	.09260	.01338	.03221	.04744	.12212
PW-D	.00064	.00187	.00427	.00611	.01436	.02083	.05105
FW-NC	.00531	.01638	.03586	.05017	.11026	.15461	.34215
FW-DH	.00281	.00770	.01614	.02208	.04560	.06169	.12191
FW-DH-A	.00230	.00632	.01354	.01860	.03838	.05192	.10102
500 Examinees							
PW-NC	.01226	.02583	.04261	.05276	.08745	.10965	.19617
PW-D	.00720	.01570	.02625	.03255	.05357	.06646	.11152
FW-NC	.04778	.09096	.13749	.16314	.24310	.29050	.46426
FW-DH	.02647	.05083	.07584	.08882	.12546	.14452	.19821
FW-DH-A	.02351	.04544	.06788	.07955	.11183	.12827	.17315

control. However, only in the 20% copying condition with $\alpha = .001$ did it result in controlling the nominal level α Type I error rate, which was inflated without adjacency control.

When 20% of the items were copied, the FW-DH Type I error rates were inflated for $\alpha \leq .001$ and for $\alpha \leq .0005$ when FW-DH-A were used. When 30% of the items were copied, Type I error rates were inflated for $\alpha \leq .005$ with and without adjacency control. When 40% of the items were copied, the Type I error rate was inflated at all α levels except .05 with and without adjacency control. For g_2 (Table 4), the familywise Type I error rate was inflated at all α levels in all conditions, regardless of the type of control or sample size.

Empirical Power

The pairwise and familywise empirical power rates for ω are shown in Table 5 for the pairwise condition, in Table 6 for the familywise per-pair condition, and in Table 7 for the familywise family-level condition. As expected, power was influenced by the percentage of items copied and test length, as well as α and the type of control. Consequently, the power data provided were collapsed only across sample size. In addition, power data are provided only for those conditions where the empirical Type I error rates provided in Tables 2–4 met Cochran's (1952) criterion for robustness (Seaman et al., 1991).

The PW-NC power of ω (Table 5) was fair when at least 20% of the items were copied on the 80-item test and at least 30% of the items were copied on the 40-item test. The familywise power of ω (Tables 6 and 7) was generally low unless at least 30% of the items were copied in the 80-item condition and 40% of the items were copied in the 40-item condition. There was little difference between the two definitions of familywise power (per-pair and family-level), and between the Dunn and Holm methods used to control the Type I error rate. Adjacency control did result in a slight reduction of power. Power data are not provided for g_2, because it failed to demonstrate the requisite control of the Type I error rate in all conditions and all α levels, except for PW-D control with $N = 100$ and $\alpha = .05$.

Pairwise power. The pairwise power of ω (Table 5) increased as a function of two variables: the percentage of items copied and test length. The PW-NC power of ω was low when 10% of the items were copied. In the 40-item conditions, the PW-NC was less than .02523, even for α

Table 5
Power of ω for Pairwise Control Averaged Across N Conditions for 40- and 80-Item Tests (Results Are Shown if the Type I Error Rates in Table 2 Satisfied Cochran's Criterion for Robustness)

% Copied	α Level						
and Method	.00001	.0001	.0005	.001	.005	.01	.05
40 Items							
10% Copied							
PW-NC	.00001	.00047	.00199	.00360	.01406	.02523	.10909
PW-D	.00000	.00013	.00053	.00101	.00429	.00761	.03002
20% Copied							
PW-NC				.02658	.07829	.12278	.31812
PW-D		.00223	.00640	.01005	.03009	.04847	.13503
30% Copied							
PW-NC					.27563	.35199	.58796
PW-D				.08433	.15860	.20672	.36882
40% Copied							
PW-NC						.54878	.74294
PW-D					.33954	.39836	.56219
80 Items							
10% Copied							
PW-NC	.00030	.00121	.00446	.00709	.02821	.04872	.17902
PW-D	.00018	.00041	.00141	.00241	.00824	.01493	.05580
20% Copied							
PW-NC				.10901	.22207	.29853	.55056
PW-D		.01320	.03971	.05628	.11635	.15970	.31532
30% Copied							
PW-NC					.49013	.57233	.77954
PW-D				.23544	.34038	.40541	.58912
40% Copied							
PW-NC						.78112	.90356
PW-D					.59651	.65433	.79045

levels as high as .01. When $\alpha = .05$, the PW-NC power of ω was only .10909. The PW-NC power in the 80-item conditions was not substantially higher; it was less than .04872 for $\alpha \leq .01$ and averaged .17902 when $\alpha = .05$. When 20% of the items were copied, the PW-NC power increased substantially, particularly in the 80-item condition. In the 40-item condition, power was still low; for $\alpha = .01$, the average PW-NC power was .12278. In the 80-item condition, however, the average power was .29853 for $\alpha = .01$, more than six times higher than the power in the 10% condition.

Interpretation of the power results when 30% or 40% of the items were copied is difficult—the pairwise Type I error rate was inflated for $\alpha < .005$ in the 30% condition and $\alpha < .01$ in the 40% condition. However, when 30% of the items were copied, the average PW-NC power at those α levels where Type I error control was maintained increased substantially in the 40- and 80-item conditions. The average PW-NC power at $\alpha = .01$ was .35199 in the 40-item condition, a 187% increase over the 20% condition. In the 80-item condition, the PW-NC power was .57233, a 92% increase over the 20% condition. When 40% of the items were copied, the PW-NC power of ω at $\alpha = .01$ increased 56% in the 40-item condition to .54878 and 36% in the 80-item condition to .78112.

Familywise power. Familywise Type I error rates were inflated at several α levels. Results of the familywise power analysis were similar to the results from the pairwise analysis, except that familywise power of ω was generally low unless the extent of copying was at least 30% of the

Table 6
Power of ω for Familywise Per-Pair Control Averaged Across N Conditions for 40- and 80-Item Tests (Results Are Shown if the Type I Error Rates in Table 3 Satisfied Cochran's Criterion for Robustness)

% Copied	α Level						
and Method	.00001	.0001	.0005	.001	.005	.01	.05
40 Items							
10% Copied							
FW-D	.00000	.00013	.00053	.00101	.00429	.00760	.03002
FW-H	.00000	.00013	.00053	.00101	.00429	.00760	.03038
FW-DA	.00000	.00007	.00047	.00087	.00301	.00592	.02340
FW-HA	.00000	.00007	.00047	.00087	.00301	.00592	.02340
20% Copied							
FW-D					.03009	.04847	.13503
FW-H					.03018	.04856	.13563
FW-DA				.00838	.02494	.04034	.11340
FW-HA				.00838	.02494	.04034	.11340
30% Copied							
FW-D						.20672	.36882
FW-H						.20687	.36990
FW-DA						.18604	.33460
FW-HA						.18604	.33460
40% Copied							
FW-D							.56219
FW-H							.56294
FW-DA							.53040
FW-HA							.53040
80 Items							
10% Copied							
FW-D	.00018	.00041	.00141	.00241	.00824	.01493	.05580
FW-H	.00018	.00041	.00141	.00241	.00824	.01494	.05607
FW-DA	.00017	.00038	.00117	.00184	.00660	.01212	.04575
FW-HA	.00017	.00038	.00117	.00184	.00660	.01212	.04575
20% Copied							
FW-D					.11635	.15970	.31532
FW-H					.11640	.15986	.31636
FW-DA				.04892	.10219	.13971	.28221
FW-HA				.04897	.10219	.13971	.28221
30% Copied							
FW-D						.42522	.58912
FW-H						.42550	.58998
FW-DA						.37807	.55341
FW-HA						.37822	.55380
40% Copied							
FW-D							.79045
FW-H							.79105
FW-DA							.76514
FW-HA							.76554

Table 7
Power of ω for Familywise Family-Level Control Averaged Across N Conditions for 40- and 80-Item Tests (Results Are Shown if the Type I Error Rates in Table 3 Satisfied Cochran's Criterion for Robustness)

% Copied	α Level						
and Method	.00001	.0001	.0005	.001	.005	.01	.05
40 Items							
10% Copied							
FW-D	.00001	.00021	.00076	.00139	.00628	.01177	.05308
FW-H	.00001	.00021	.00076	.00139	.00628	.01177	.05308
FW-DA	.00000	.00010	.00067	.00117	.00450	.00928	.04070
FW-HA	.00000	.00010	.00067	.00117	.00450	.00928	.04070
20% Copied							
FW-D					.03231	.05309	.15603
FW-H					.03231	.05309	.15603
FW-DA				.00887	.02653	.04358	.12923
FW-HA				.00887	.02653	.04358	.12923
30% Copied							
FW-D						.21266	.38718
FW-H						.21266	.38718
FW-DA						.19067	.34954
FW-HA						.19067	.34954
40% Copied							
FW-D							.57524
FW-H							.57524
FW-DA							.54134
FW-HA							.54134
80 Items							
10% Copied							
FW-D	.00023	.00050	.00161	.00284	.01053	.01895	.07551
FW-H	.00023	.00050	.00161	.00284	.01053	.01895	.07551
FW-DA	.00017	.00047	.00136	.00220	.00852	.01548	.06056
FW-HA	.00017	.00047	.00136	.00220	.00852	.01548	.06056
20% Copied							
FW-D					.12055	.16623	.33417
FW-H					.12055	.16623	.33417
FW-DA				.05051	.10576	.14515	.29734
FW-HA				.05051	.10576	.14515	.29734
30% Copied							
FW-D						.43158	.60138
FW-H						.43158	.60138
FW-DA						.38349	.56411
FW-HA						.38349	.56411
40% Copied							
FW-D							.79666
FW-H							.79666
FW-DA							.77076
FW-HA							.77076

questions on the 80-item test and 40% of the questions on the 40-item test. Per-pair power (Table 6) and family-level power (Table 7) were very similar in all situations, particularly for $\alpha < .05$. The per-pair FW-D and FW-H conditions yielded essentially the same power results, with differences between them usually only noticeable in the fourth decimal place. As mentioned previously, the FW-D and FW-H conditions yielded identical power in the family-level conditions. Because of these

similarities, only the results of the FW-H condition will be discussed. Using adjacency control resulted in lower power, though the differences were rarely meaningful.

FW-H power was very poor in the 10% copying conditions. When 40 items were used, per-pair and family-level FW-H power were both barely higher than the corresponding familywise Type I error rate. Power was less than .05308 for all conditions and at all α levels. The 80-item condition produced slightly higher FW-H power, although it was still barely higher than α.

FW-H power in the 20% condition remained low in the 40-item tests. Per-pair FW-H and FW-HA power averaged .04856 and .04034, respectively, when $\alpha = .01$. The corresponding family-level FW-H and FW-HA power averages were .05309 and .04358, respectively. FW-H power, although still low, was substantially higher in the 80-item condition. Inflated Type I error rates made it impossible to evaluate FW-H power at small α levels. At $\alpha = .01$, the per-pair FW-H and FW-HA power averaged .15986 and .13971, respectively. Average family-level FW-H and FW-HA power were .16623 and .14515, respectively.

FW-H power increased considerably in the 30% conditions, but the α levels at which power was evaluated were reduced further. In the 40-item conditions, the average per-pair FW-H power was .20687 for $\alpha = .01$, and .18604 for FW-HA. The corresponding average family-level power was .21266 for FW-H and .19067 for FW-HA. When 80 items were used, the average per-pair FW-H power was .42550 and average power for FW-HA was .37822. Average family-level FW-H and FW-HA power were .43158 and .38349, respectively.

FW-H power in the 40% condition was good for both test lengths, but the power data were only meaningful for high α levels. In fact, FW-H and FW-HA power data could only be reported for the $\alpha = .05$ condition. With the 40-item test, the familywise power data at $\alpha = .05$ were very similar to the familywise power data for the 80-item test with $\alpha = .05$ when 30% of the items were copied. The average per-pair FW-H and FW-HA power at $\alpha = .05$ was .56294 and .53040, respectively. Average family-level FW-H and FW-HA power was .57524 and .54134, respectively. The 80-item conditions offered very good power to detect Cs at the $\alpha = .05$ level. Average per-pair power was .79105 and .76554 for the FW-H and FW-HA conditions, respectively. The corresponding family-level powers were .79666 and .77076. Although power was good for the 40% copying condition, it is unlikely that a copying analysis would be performed with an α level as high as .05 in practice.

Reconsidering the Definition of a Type I Error

The Type I error rate of ω, even with controls on α, was inflated in many situations when more than 20% of the items were copied. This is in contrast with previous research (Chason, 1997; Wollack, 1997; Wollack & Cohen, 1998), which found good control of the pairwise Type I error rate. This result is not surprising. The difference between previous studies and this study is that in this study, copying indices were assumed to be directional. In previous research, a significant index between adjacent pairs did not identify which examinee was C and which was S. Here, an attempt was made to distinguish the two such that a significant index from a noncopying S to a true C was considered a Type I error. Although it is true that these reciprocal indices are technically Type I errors (when the indices are used directionally), they were detected at a rate far exceeding α. This is because, by copying many answers from S, C also implicated S. This effect is most pronounced at small α levels where it takes only a few significant reciprocal indices to produce inflated Type I error rates.

To investigate the severity of this problem, the data were re-analyzed so that indices identifying S as copying from a true C were not considered Type I errors, provided that the index computed for C copying from S also was statistically significant. Under this approach, it was assumed that, should the two indices between an adjacent C-S pair both be significant, an investigation would be

initiated that would lead to the correct identification of C. After such an investigation, S would no longer be considered a Type I error.

Re-scoring the data in this manner had a dramatic impact on the pairwise and familywise Type I error rates of ω. As an example of the severity of this problem, in the 500-examinee, 80-item, 40% copying condition, of the 50,000 true C-S pairs, 40% of which (i.e., 20,000) were expected to be between adjacent examinees, it was found that in 19,425 cases, the indices for the C-S and S-C pairs were both statistically significant at $\alpha = .05$ when no multiple comparison procedure was used. Even when $\alpha = .00001$, this condition provided 6,298 cases where both indices were significant.

Table 8 shows that the adjusted Type I error rates were considerably more aligned with α, particularly as the percentage of items copied increased. For the pairwise definition, the nominal Type I error rates were controlled for $\alpha \geq .0005$. For the FW-DH and FW-DH-A definitions, the nominal Type I error rates were controlled for $\alpha \geq .001$. The adjusted Type I error rates no longer differed as a function of the percentage of items copied. (Power data for those cells for which adjusted Type I error rates were controlled are not provided here due to space limitations, but are available from the authors.)

Table 8
Type I Error Rates of ω With Significant Source-Copier Pairs Removed
Averaged Across N, Test Length, and Percent Copied Conditions

% Copied	α Level						
and Method	.00001	.0001	.0005	.001	.005	.01	.05
Pairwise							
PW-NC	.00017	.00030	.00058	.00086	.00308	.00596	.03286
PW-D	.00010	.00016	.00025	.00032	.00080	.00174	.00638
Familywise							
FW-NC	.00072	.00130	.00253	.00378	.01346	.02597	.13434
FW-DH	.00042	.00068	.00103	.00134	.00342	.00844	.02772
FW-DH-A	.00037	.00061	.00091	.00115	.00281	.00461	.02178

The adjusted Type I error rates were necessarily lower for g_2 as well, but the adjustment did not increase the number of conditions for which the Type I error rate was controlled. After the adjustment, the Type I error rates for g_2 were inflated in every situation except PW-D (a condition that is difficult to interpret) for 100 examinees at $\alpha = .05$. The general pattern of controlled versus uncontrolled error rates was identical to the pattern without the adjustment, shown in Table 4.

Discussion

The traditional pairwise definition of Type I error rate in identifying copying behavior resulted in identifying too many examinees as Cs. The FW-NC data for ω and g_2 (Tables 3 and 4) provided an indication of the percentage of examinees who would be identified as Cs when each contrast was tested for significance at α. In all conditions and at all α levels, the FW-NC error rate was greater than α. Even in the conditions in which the pairwise Type I error rate was controlled, the ω index yielded FW-NC Type I error rates that were an average of 3.04 times larger than α. For example, in the 10% copying condition (from Table 2), the PW-NC Type I error rate for ω at $\alpha = .05$ was .03339, indicating that only over 3% of the noncopying pairs were detected. However, the FW-NC Type I error rate was .13835 (Table 3), indicating that almost 14% of the noncopying examinees were detected. Using the adjusted Type I error rates (Table 8), the FW-NC Type I error rate was .13434. With g_2, the PW-NC Type I error rate with 100 examinees at $\alpha = .05$ was .12212 (Table 4),

but over 34% of noncopying examinees were identified. For the ω index, FW-NC Type I error rates (at $\alpha = .05$) ranged from .13835 (in the 10% copying condition) to .15025 (in the 40% copying condition). For g_2, they ranged from .34215 ($N = 100$) to .46426 ($N = 500$). Clearly, because the uncontrolled familywise Type I error rates were so much larger than α, an appropriate multiple comparison approach should be considered to attempt to keep the familywise error rate at a more acceptable level.

Pairwise and familywise Type I error rates for ω were inflated in several situations when α was small and the percentage of items copied was large. This effect was observed because the two indices computed for adjacent pairs were highly correlated. Although ω is computed conditional on the examinee's θ (as estimated from his/her item responses), it is also conditional on the number of items in common between two examinees. If the number of answer matches is high enough for one index to reach statistical significance, it is likely that the reciprocal index will also be significant, unless the difference between the two examinees' θ estimates is substantial. When one examinee actually copies from another, this effect is intensified. Copying not only increases the number of answers that are common to both examinees, but it also brings the two θ estimates closer together than they might otherwise be.

The adjusted definition of Type I error used for adjacent S-C pairs is, of course, only possible in a simulation study. Making this assumption resulted in much better control of the nominal Type I error rates—pairwise Type I error was controlled for $\alpha \geq .0005$, and familywise Type I error was controlled for $\alpha \geq .001$, regardless of the number of examinees, test length, or percentage of items copied. It is useful to compare the results for the adjusted definition with the unadjusted result to see what the improvement could be when some additional knowledge is available regarding the true C.

Sample size had a statistically significant effect on the Type I error rates for g_2—rates became larger as sample size increased. Under all conditions studied, Type I error rates were inflated. Adjusting the Type I error rates as described above did not improve this result. Neither the percentage of items copied nor the test length affected the Type I error rates of g_2.

ω had fair-to-good pairwise power when 30% of the items were copied on a 40-item test and 20% of the items were copied on an 80-item test. ω had good power when copying 40% of the items on the 40-item test and 30% of the items on the 80-item test. The familywise power of ω, however, was quite low for 30% or fewer of the items copied on a 40-item test and 20% or fewer of the items copied on an 80-item test. The familywise power of ω was fair to good, however, when copying 40% of the items on the 40-item test and 30% of the items on the 80-item test.

As expected, the FW-H procedure offered more per-pair power to detect copiers than did the FW-D procedure, although the differences were in the fourth decimal place in most cases. This is probably a function of the way in which the simulations were implemented. Had simulated copiers been allowed to use multiple Ss, the FW-H procedure would likely have provided a more noticeable improvement.

Also as expected, adjacency control resulted in a decrease in power. However, this decrease was much smaller than anticipated. As an example, across the conditions where power was reported, adjacency control lowered power by no more than .03764. Adjacency control did help lower the Type I error rates, and in one case helped align it with expectation. Therefore, when a seating chart is available, it is recommended that the adjacency correction be considered to control for the lack of independence among adjacent pairs, particularly at small α levels.

Family-level power was negligibly higher than per-pair power. Again, this was probably a function of the way in which the simulations were implemented. Had copiers been allowed to use multiple sources, the family-level power likely would have been higher. Instead, the only cases

identified using the family-level definition that were not identified using the per-pair definition were pairwise Type I errors, which occurred at a very low rate.

The inflated familywise Type I error rates (without making the adjacency adjustment), as the percentage of items copied increased, suggest that additional work is needed to control the effects of pairs comprised of a noncopying source and a true copier. However, the results reported here for the 30% and 40% conditions represent worst-case scenarios. In these simulations, all copiers were copying 30% or 40% of the items. A more likely scenario would be that examinees copying 30% or 40% would constitute a small percentage of the total copiers. The majority of the copiers would likely copy 10% or 20% of the items, conditions for which the Type I error rate is better controlled. Nevertheless, Hanson et al. (1987) suggested using empirical rather than theoretical sampling distributions, whenever possible, to improve Type I error control for copying indices. Unfortunately, empirical sampling distributions constructed using only examinees who could not have copied from one another is a possibility only if the same form of the test is administered to many examinees in multiple testing locations.

Conclusions

This study focused exclusively on a particular scenario in which multiple comparisons are needed. Indeed, there are many situations other than detection of answer copying that require the use of multiple-comparisons procedures. In fact, any time multiple correlated tests are performed (as was done here by computing two copying indices for adjacent pairs), a multiple-comparisons procedure is necessary to maintain the probability of committing a Type I error at its nominal level. However, multiple-comparisons procedures are not limited to this situation alone. The impetus for this study was based on the fact that if there are multiple families, each possessing a different number of contrasts, and a common probability of committing a Type I error for all families is desired, it is necessary to apply a multiple-comparisons procedure within each family.

In spite of the loss of Type I error control under some conditions, a familywise definition of Type I error rate is preferable to a pairwise definition, because it does help to appreciably reduce the probability of falsely identifying an examinee as copying. Also, if it is assumed that those source examinees detected with their copiers are successfully identified during a follow-up investigation, the familywise Type I error control is very good.

Regardless, caution should be applied when using any copying index. It has been argued that statistical indices alone should not be used to question the validity of the examinee's score without accompanying evidence of wrongdoing, such as direct observation by a proctor (Buss & Novick, 1980; Haney, 1998). Wollack (1997) suggested using copying indices to identify which examinees to monitor closely during subsequent exams. In this case, ω could be used with a fairly high familywise α (such as $\alpha = .01$) to ensure adequate power. Examinees detected on two independent exams would be suspected of copying with $p < .0001$. Under ideal circumstances, a true copier would not only be detected twice with ω, but actually observed to be copying on at least one occasion. Also, ω does not need to be used as a general screening tool at all—ω could still be used to provide corroborating evidence against an examinee observed copying during an exam.

References

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association, 69,* 44–49.

Baker, F. B. (1986). *GENIRV: Computer program for generating item responses* [Computer program]. Madison WI: University of Wisconsin, Department of Educational Psychology, Laboratory of Experimental Design.

Bay, L. (1995, April). *Detection of cheating on multiple-choice examinations.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco CA.

Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple choice tests by using error-similarity analysis. *Teaching of Psychology, 16,* 151–155.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 46,* 443–459.

Buss, W. G., & Novick, M. G. (1980). The detection of cheating on standardized tests: Statistical and legal analysis. *Journal of Law and Education, 9,* 1–64.

Chason, W. M. (1997, March). *A comparison of several classical and IRT-based methods to detect aberrant response patterns.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Chason, W. M., & Maller, S. (1996, April). *Utility of the Rasch person-fit statistic in detecting answer copying: A comparison.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Cochran, W. G. (1952). The chi-square test of goodness of fit. *Annals of Mathematical Statistics, 23,* 315–345.

Cody, R. P. (1985). Statistical analysis of examinations to detect cheating. *Journal of Medical Education, 60,* 136–137.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56,* 52–64.

Einot, I., & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association, 70,* 574–583.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 2,* 235–256.

Haney, W. M. (1998, April). *The value of supplementary evidence in evaluating unusual answer concordance.* Paper presented at the annual meeting of the American Educational Research Association, San Diego CA.

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying* (Research Rep. Series No. 87-15). Iowa City IA: American College Testing.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75,* 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

Iwamoto, C., Nungester, R. J., & Luecht, R. M. (1996, April). *Power of similarity methods and person-fit analysis to detect copying behavior.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Iwamoto, C. K., Nungester, R. J., Watson, S. A., & Luecht, R. M. (1997, March). *Effect of test length and comparison group selection on the power of similarity methods to identify copiers.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Iwamoto, C. K., Watson, S. A., Nungester, R. J., & Luecht, R. M. (1997, March). *The use of response similarity and person between fit measures to detect irregular behaviors.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences (3rd ed.).* Pacific Grove CA: Brooks/Cole.

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110,* 577–586.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62,* 626–633.

Wollack, J. A. (1996). Detection of answer copying using item response theory. Unpublished doctoral dissertation, University of Wisconsin, Madison. *Dissertation Abstracts International, 57/05,* 2015.

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement, 21,* 307–320.

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22,* 144–152.

Acknowledgments

Portions of this paper were presented at the annual meeting of the American Educational Research Association, San Diego CA, April, 1998. The authors thank the editor and two anonymous reviewers for their insightful comments and suggestions.

Authors' Addresses

Send requests for reprints or further information to James A. Wollack, Allan S. Cohen, or Ronald C. Serlin, University of Wisconsin, 1025 W. Johnson Street, Madison WI 53706, U.S.A. Email: jwollack@facstaff.wisc.edu, ascohen@facstaff.wisc.edu, rcserlin@facstaff.wisc.edu.

Book Review

Item Response Theory for Psychologists

Susan E. Embretson and Steven P. Reise
Mahwah NJ: Erlbaum, 2000, 371 pp., approx. $39.95 (soft cover)

Item response theory (IRT) allows the user to specify a mathematical function to model the relationship between a latent trait, θ, and the probability that an examinee with a given θ will correctly answer a test item. Until the 1980s, IRT research focused largely on the estimation of model parameters, the assessment of model-data fit, and the application of these models to a range of testing problems using dichotomously scored multiple-choice items. Earlier IRT textbooks clearly reflect this emphasis (e.g., Baker, 1985; Hambleton & Swaminathan, 1985; Hulin, Drasgow, & Parsons, 1983; Lord, 1980; Wright & Stone, 1979).

During the 1980s, research on IRT models and their applications increased dramatically as a much broader research agenda was adopted and pursued. For example, research on performance assessments, polytomous response formats, and multidimensional traits began in earnest, as did work on computerized adaptive testing. An outcome of this expanded focus was a host of new IRT models that allowed researchers to tackle complex problems, not only in achievement testing, but also in areas such as attitude, personality, cognitive, and developmental assessment. It also created the need—only now beginning to be met—for computer software to operationalize these models.

Currently, few introductory textbooks cover these important advances from the last two decades. Instead, these developments are summarized only in more advanced books (e.g., Fischer & Molenaar, 1995; van der Linden & Hambleton, 1997) and in journals that sometimes focus on these special issues (e.g., Ackerman, 1996; Drasgow, 1995; Hambleton, 2000; Meijer & Nering, 1999). Clearly, an introductory text was necessary for bridging the gap between existing knowledge and conventional wisdom on IRT with important new developments that now characterize the field. *Item Response Theory for Psychologists* bridges this gap. It is up to date and comprehensive, and it serves as a guide to extending the use of IRT-based methods.

Overview

The book consists of four parts.

Part I. Chapter 1 is the only chapter in Part I. In it, current IRT applications are described, a brief history of IRT is presented, and the content of the book is surveyed.

Part II. Chapters 2 and 3 are devoted to IRT principles. These chapters are designed to compare and contrast basic principles and practices in IRT and classical test theory (CTT). They are intended to provide "an intuitive understanding of IRT principles" (p. 8). However, readers should have a firm grasp of the concepts in IRT and CTT before reading these chapters.

Part III. Chapters 4–9 are devoted to topics in binary IRT models, polytomous IRT models, scaling, θ estimation, item calibration, and model-data fit. These chapters are noteworthy for their instructional emphasis and for their broad coverage of IRT fundamentals. Basic concepts (e.g.,

Applied Psychological Measurement, Vol. 25 No. 4, December 2001, 405–408

polytomous IRT, item and θ parameter estimation, person fit) are carefully described and illustrated. Also included in Part III are a broad range of topics illustrating how the field has expanded in the last two decades—diverse unidimensional, polytomous, and multidimensional models are described. Moreover, the importance of assessing dimensionality using recent concepts (e.g., essential undimensionality as operationalized with Stout's DIMTEST) and sophisticated procedures (e.g., nonlinear factor analysis) are highlighted.

Part IV. Applications of IRT models are featured in Chapters 10–13. Part IV best emphasizes the authors' interest in psychological applications of IRT. Chapter 10 is devoted to topics familiar to most measurement specialists: linking, differential item functioning (DIF), computerized adaptive testing, and scale analysis. Chapters 11 and 12, on the other hand, are devoted to topics that might be unfamiliar to some measurement specialists, ranging from cognitive and developmental assessment to personality and attitude assessment. Also discussed is how IRT principles are used to solve substantive problems in these areas. Chapter 13 is an instructional chapter on calibrating selected IRT models using seven popular computer programs.

A Psychometrican's Point-of-View

This book has many strengths. It is comprehensive and instructional. As a result, it serves as a strong graduate-level textbook that will appeal to the introductory and advanced reader. Unidimensional, polytomous, and multidimensional models and their applications in educational and psychological settings are surveyed, ranging from the Rasch model to more complex IRT models that include discrimination and guessing parameters. Considerations and controversies related to model selection and application are discussed.

Four chapters—Chapters 5, 7, 8, and 13—are worth noting. Chapter 5 provides an excellent description of five popular polytomous IRT models, each of which is presented mathematically, illustrated with a real-data example, and operationalized with a computer program. The link between model and computer program is invaluable for readers who intend to conduct IRT analyses with polytomous data. This link is reinforced in Chapter 13, which discusses computer programs used when conducting IRT parameter estimation. Chapters 7 and 8, on θ and item parameter estimation, provide an excellent review of the technical issues surrounding these topics using dichotomous data. The chapters are easy to read and well illustrated.They will provide readers with a strong conceptual understanding of many parameter estimation techniques, their applications, and interpretations.

The authors present sage advice on topics and issues that reflect the current views of many researchers and practitioners. For example, on the topic of model-data fit, they recommend implementing newer procedures (e.g., TESTFACT, POLYFACT, NOHARM, and LISCOMP) and not depending on heuristic indices such as "variance accounted for by the first factor" or "ratio of first to second eigenvalue." With the presentation of current information, the authors also confront some controversies and unresolved issues. This confrontation strengthens the book by emphasizing important research opportunities. In fact, research problems are identified throughout the text on such topics as construct validation (p. 180), dimensionality assessment (p. 231), applications of person-fit indices (p. 242), model comparison strategies (pp. 243–244), multidimensionality and differential item functioning (p. 263), DIF in the context of personality assessment (p. 319), cognition and assessment (p. 290), attitude assessment (p. 315), and computerized personality assessment (p. 317).

Item Response Theory for Psychologists places an emphasis on diverse models and computer software to operationalize these models. This link is essential for readers conducting IRT analyses. Although the connection is explicit for some models (e.g., Chapters 5 and 13), it is implicit for

others. For example, the linear logistic trait model, hyperbolic cosine IRT model, parallelogram IRT model, nonparametric regression with kernel smoothing, multicomponent latent trait model, general latent trait model, multidimensional Rasch model for learning and change, and SALTUS model all are presented in detail. Also described are procedures such as tree-based regression and rule-space analysis. Some of these models are illustrated further in Part IV, demonstrating their usefulness when addressing substantive problems in the areas of cognitive, developmental, attitude, and personality assessment. Unfortunately, there are few references to the software needed for calibrating the parameters of these particular models and, consequently, no obvious means for operationalizing the models. The reader intending to use IRT would benefit if each model presented in the text were linked with an appropriate computer program. Models and procedures without readily available calibration software should be noted.

A Research Psychologist's Point-of-View

This book is both stimulating and somewhat frustrating. It is intended for psychologists who are familiar with neither IRT nor other relatively recent innovations in test theory and development, who could benefit from the use of IRT methods in their own research, and who have reasonable knowledge of classical statistics and measurement issues. With chapters devoted explicitly to comparisons of CTT, IRT, and applications of IRT in cognition, development, personality, and attitude assessment, the authors clearly want their message to be not only accessible but persuasive to researchers naive in IRT.

However, to be successful in this goal, they must (1) convince readers that the principles and assumptions of IRT are inherently more reasonable and useful than those of CTT, at least under some conditions; and (2) illustrate how IRT can be used to study mainstream topics in psychological research. The first of these requirements is addressed in Chapter 2, in which ten "rules" of CTT are compared with their counterparts in IRT. This approach is direct and appealing, but the result is not entirely compelling. In many cases, differences in CTT and IRT are described, but why these differences are important is not mentioned. For example, the difference between CTT and IRT is clearly described regarding the standard error of estimate and scale properties, but little or no effort is made to explain why or under what conditions the IRT versions might be considered superior. In other cases (e.g., use of IRT to measure change), the description is too cursory to provide much insight.

Too often, assertion—rather than detailed reasoning—is used. Comparing CTT and IRT principles directly has considerable didactic value, but a single, short chapter might not be sufficient for the job. Many readers not already familiar with IRT might abandon the task after finding this shortcoming. However, readers naive in IRT will appreciate the efforts made throughout the book to highlight difficulties with traditional measurement approaches and to promote reflection on commonly used practices.

The second requirement mentioned above, illustrating the use of IRT on mainstream topics, is addressed indirectly in several technical chapters and directly in three chapters devoted specifically to applications in research. In Chapter 10, the use of IRT methods in real-world testing, including computerized adaptive testing, is described. This chapter is likely to be interesting and informative to most readers, and it successfully conveys the appealing properties of IRT methods for testing research and practice. In Chapter 11, the focus is on applications in cognitive and developmental psychology, and on personality and attitudinal research in Chapter 12. These chapters are likely to receive mixed reviews from the intended audience. Readers certainly will be impressed with the range of topics to which IRT can be applied, and also with the range of methods available. The

advantages of methods, for example, yielding person response functions and enabling identification of qualitative differences between groups will be obvious immediately.

Another flaw is the inconsistency in the descriptions of various applications. In the best cases, how a method can be applied and the types of inferences that can be made are mentioned, as well as how these inferences can differ from those that might arise from traditional methods. This approach highlights the value added by using IRT. Chapter 12 is particularly useful in this respect, containing a very helpful section devoted to the advantages and disadvantages of IRT. In too many cases, however, the authors merely show how data can be analyzed with IRT methods, but provide few clues as to how these methods can lead to insights that would not otherwise be transparent.

It is not always clear how IRT research would fit into the current study of mainstream issues. For example, cognitive researchers often use a combination of accuracy, latency, and self-report measures to determine how individuals solve problems. Because so many different kinds of data are used, and because analyses are quite intensive and not entirely quantitative in nature, sample sizes tend to be relatively small. In contrast, the IRT methods described in the book are used to analyze accuracy data based on paper-and-pencil tests, and the sample sizes tend to be an order of magnitude larger than what is common in cognitive research. To optimize the appeal to cognitive researchers, it would be helpful to show not only that IRT methods can be applied in cognitive research, but also how they can complement and build on the insights gained with traditional methods.

Many research psychologists are relatively unaware of IRT, its principles, and its potential advantages over CTT. In their mainstream journals, research psychologists see very few applications of IRT. Most are well aware of shortcomings inherent in traditional measurement and statistical techniques, but they continue to operate with one hand manipulating the old tools and the other grasping for new ones. Clearly, the conditions are rife for new approaches. *Item Response Theory for Psychologists* makes a very strong and enthusiastic case for the potential of IRT methods in general, but it might not be sufficiently compelling to encourage many researchers to take the leap from CTT to IRT. It will attract the attention of some, however, and might lead others to look for collaborators who can help them develop expertise in IRT methods. We anticipate that this book will contribute significantly to the momentum for change.

Mark J. Gierl and Jeffrey Bisanz
University of Alberta

References

Ackerman, T. (1996). Developments in multidimensional item response theory. *Applied Psychological Measurement, 20(4).* [Special issue.]

Baker, F. B. (1985). *The basics of item response theory.* Portsmouth NH: Heinemann.

Drasgow, F. (1995). Polytomous item response theory. *Applied Psychological Measurement, 19(1).* [Special issue.]

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications.* New York: Springer.

Hambleton, R. K. (2000). Advances in performance assessment methodology. *Applied Psychological Measurement, 24(4).* [Special issue.]

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood IL: Dow Jones-Irwin.

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

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing. *Applied Psychological Measurement, 23(3).* [Special issue.]

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.

Book Review

Computerized Adaptive Testing: Theory and Practice

Wim J. van der Linden and Cees A. W. Glas (Eds.)

Dordrecht, The Netherlands: Kluwer, 2000, 323 pp., approx. $89.95 (hard cover)

Computerized Adaptive Testing: Theory and Practice contains fifteen chapters of cutting-edge research in computerized adaptive testing (CAT). The book contrasts sharply with Wainer et al.'s (2000) CAT primer (see Reise, 2001). Wainer et al.'s book was written for beginning and intermediate audiences and focuses on basic research and implementation issues surrounding CAT. In van der Linden and Glas' book, however, the reader is presented with a series of research articles written specifically for advanced psychometricians. The book is divided into five parts: item selection and examinee scoring in CAT, examples of CAT applications, item banks, determining model fit, and using testlets in CAT.

Part 1

Part 1 contains three chapters that focus on recent research in item selection and trait (θ) estimation in CAT. In Chapter 1, van der Linden and Pashley review classic approaches to item selection (e.g., maximum information) in CAT and introduce the mathematics underlying several modern approaches. These approaches include item selection based on Kullback-Leibler (in contrast to Fisher) information, a likelihood-weighted information approach, and a Bayesian approach.

In Chapter 2, van der Linden introduces the notion of "shadow tests" and demonstrates how they can be used to create constrained (e.g., item exposure controlled) optimized CATs. Shadow tests are created by a computer optimization algorithm to meet statistical criteria (e.g., maximizing information), as well as a specific set of nonstatistical constraints (e.g., item content constraints). These shadow tests are created prior to administering each item in CAT. A variety of illustrative applications also are provided, along with citations for relevant software.

In Chapter 3, Segall elaborates on a specific Bayesian approach for implementing multidimensional CAT. The aim of this model ". . . is to increase the efficiency of adaptive item selection and scoring algorithms by extending unidimensional methods to the simultaneous measurement of multiple dimensions" (p. 54). The results are quite promising in terms of efficiency, especially if the multiple domains are highly correlated. The specific benefits to be accrued by multidimensional CAT in a wide range of specific testing contexts remains a question for future research using a diverse collection of real item banks.

Part 2

Part 2 contains three chapters on CAT applications. In Chapter 4, Mills and Steffen describe major implementation issues involved with the CAT Graduate Record Examination. Included are discussions of how item banks are developed and maintained and how missing responses are scored.

Applied Psychological Measurement, Vol. 25 No. 4, December 2001, 409–411

In Chapter 5, Verschoor and Straetmans describe the MATHCAT system. This system, which is used in Dutch colleges for adult students, is particularly interesting because it contains two components: (1) a placement component used to assign people to courses, and (2) an achievement component used to compare examinees in terms of proficiency.

In Chapter 6, Luecht and Nungester describe the implementation of a computerized adaptive sequential testing (CAST) algorithm. The CAST algorithm forms the basis of the computerized version of the United States Medical Licensing Examination. What makes this approach particularly interesting is that "CAST sacrifices adaptive flexibility for control" (p. 126) by making heavy use of automated test assembly procedures. The pros and cons of various formulations of the CAST model are discussed.

Part 3

Part 3 focuses on item bank development and maintenance. Chapter 7, by Parshall, Davey, and Pashley, describes innovative item types of potential value in CAT. These are defined as items that include sound, graphics, animation, or video. The authors describe five dimensions on which innovative item types might vary: (1) item format, (2) how the examinee responds to the item, (3) the use of media, (4) level of interactivity, and (5) how the item is scored. Examples of innovative item types and relevant citations are provided. However, as the authors point out, although computerized administration offers a wider range of item possibilities relative to paper-and-pencil tests, more research is needed to determine whether these new item types will lead to better measurement of traditional constructs.

In Chapter 8, Veldkamp and van der Linden describe the creation of item banks for CAT when working from a highly structured blueprint. This chapter provides an excellent discussion of shadow tests and optimization algorithms, given a test with many constraints.

Chapter 9, by Stocking and Lewis, provides further elaboration on the topic of item exposure control. A definitive review of existing item exposure control algorithms is provided. Also, a new algorithm is presented based on conditioning on estimated θ, rather than true θ.

Part 4

The fourth section covers model fit and invariance topics (i.e., whether parameters remain the same under different testing conditions or for different groups of examinees). In Chapter 10, Glas considers the phenomenon of item parameter drift. This occurs when an item's psychometric properties change over time, perhaps due to exposure or other reasons (e.g., item parameters might change when going from a paper-and-pencil calibration sample to implementation in CAT). Two methods of identifying item parameter drift are described. The first uses a Lagrange multiplier statistic, and the second is a cumulative sum statistic designed specifically to evaluate whether an item is becoming easier over time. Glas's chapter also includes an exceptionally clear discussion of marginal maximum likelihood estimation.

In Chapter 11, van Krimpen-Stoop and Meijer address the difficult problem of detecting response aberrance (i.e., person misfit) in a CAT context. This is a particularly challenging issue, because CATs are typically designed to administer items to match an examinee's θ level. Thus, items with a relatively small item difficulty range are selected. However, person misfit is optimally identified when examinees are administered items of wide-ranging difficulty. The authors propose a series of new cumulative sum person-fit statistics that are based on statistical process control theory. The chapter reports basic exploratory monte carlo research and demonstrates how each of a series of new person-fit statistics is designed to assess a different form of response aberrance.

Zwick's Chapter 12 describes basic approaches to identifying differential item functioning in the CAT context. This topic also presents a challenge, for a variety of reasons. For example, some items might receive a lot of exposure in CAT, while others receive very little. Most interesting, an empirical Bayesian approach to computing the Mantel-Haenszel statistic is presented, with promising results.

Part 5

The last part of the book contains three chapters on testlets. A testlet is a set of items that are administered together and scored as a group. The creation of testlets traditionally has been proposed as a potential solution to violations of local independence that might arise, for example, when items are all linked to a common reading passage. In Chapter 13, Wainer, Bradlow, and Du lay out the logic and rationale for "testlet response theory." They develop a Bayesian random effects model for testlets and demonstrate how the Gibbs sampler can be used to estimate parameters. The beauty of the approach is that it allows the effect of item dependence to be parameterized, which in turn allows researchers to better document the effect of violating local independence. The testlet model and traditional approach are compared using Graduate Record Examination and Scholastic Aptitude Test datasets.

Chapter 14, by Glas, Wainer, and Bradlow, is essentially a follow-up to the previous chapter, but it uses a more traditional marginal maximum likelihood approach to item parameter estimation. In Chapter 15, Vos and Glas extend testlet response theory to the mastery testing domain.

Summary

This edited volume should be useful to psychometricians conducting CAT research or considering applying CAT to existing measures. The chapters are integrated nicely and cohesive, with frequent references to other chapters in the volume. The full presentation of equations and methods provided by most chapters will be appreciated. This book especially is relevant to researchers working on large-scale, high-stakes programs where issues such as item content balance and examinee cheating are particularly salient. There are two basic themes—one specific and one broad—that occur repeatedly across the various chapters: optimization algorithms for CAT and Bayesian methodologies. Many of the chapters can be considered definitive of the current state of research on these topics. The complexity and sophistication of CAT, along with interesting and complex research questions that still need to be addressed, are clear after reading *Computerized Adaptive Testing: Theory and Practice.*

Steven P. Reise
University of California, Los Angeles

References

Reise, S. P. (2001). Book review. Computerized adaptive testing: A primer (2nd ed.). *Applied Psychological Measurement, 25,* 307–309.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer (2nd ed.).* Mahwah NJ: Erlbaum.

Volume 25 Author and Subject Index