ETS®

Listening. Learning. Leading.®

Research Spotlight

Issue 2 – February 2009

Acknowledgments

The second issue of ETS Research Spotlight would not have been possible without the contributions of the following ETS staff members:

Article reviews	Dan Eignor
Editing and layout	Jeff Johnson
Cover photography	Bill Petzinger
Cover design	Marita Gray
Technical assistance	Charlie Betz

ETS Research Spotlight
Issue 2

Copies can be downloaded from:
www.ets.org/research

February 2009
Research & Development Division
Educational Testing Service

To comment on any of the articles in this issue, write to ETS Research Communications via e-mail at: R&DWeb@ets.org

Table of Contents

Foreword

In 2008, ETS researchers and scientists published 70 reports in the ETS Research Report Series, more than 40 articles in refereed journals, 19 book chapters, and edited or coedited three books. In addition, our researchers gave hundreds of presentations at conferences and professional meetings around the world.

All of this activity has been in support of ETS's mission as a nonprofit organization: To advance quality and equity in education by providing fair and valid assessments, research, and related services for all people worldwide.

In four articles adapted from the ETS Research Report Series, Issue 2 of ETS Research Spotlight provides a small taste of the range of assessment-related research capabilities of the ETS Research & Development Division. Those articles cover assessment-related research aimed at developing models of student learning, applying standard-setting methodology, advancing equating methodology, and assessing information literacy.

In the last section of this issue, we include the abstracts from all 2008 contributions to the ETS Research Report Series. For a look at older reports in this series—dating back to 1948—interested readers can visit the ETS ReSEARCHER searchable database on the Web (http://search.ets.org/custres/).

If you have any questions about any of the articles in our research portfolio, please contact us. You can send your inquiry via e-mail to R&DWeb@ets.org.

Ida Lawrence
Senior Vice President
ETS Research & Development

A Developmental Writing Scale

Yigal Attali and Donald Powers

Editor's note: *Currently, there is no satisfactory way to chart the development of children's writing skills. In order to improve the science of assessment, ETS supports research to better the field's understanding of the underlying constructs tests are supposed to measure. In this study, ETS researchers created a developmental writing scale based on objective and automatically computed measures of writing skill (word choice, grammatical conventions, and fluency). The scale was constructed through a large-scale data collection effort that involved a national sample of over 12,000 4th through 12th grade students. By allowing greater comparability of scores within and across grade levels, the developmental writing scale offers some advantages for improving the current practice of writing assessment.*

Writing is a complex literacy skill that develops slowly over time. Composing a text requires coordinating low-level skills, such as reading and handwriting/typing, with high-level skills, such as problem-solving related to content and rhetoric issues. Currently, there is no entirely satisfactory way to chart the development of children's writing skills. Creating a developmental scale of writing depends upon good measurement, and writing is difficult to assess. Indirect measures of writing can be reliable, but tend to assess only isolated writing skills (e.g., sentence composing, editing, etc.).

Direct measures of writing that require students to compose extended text (e.g., essays) elicit a fuller range of writing skills. However, scoring essays depends on extensive training of raters to develop a shared interpretation of a particular scoring standard. This means that essay scores from one assessment cannot be compared as such to those of another assessment or teacher, and comparability of scores across grade levels or time is limited. Moreover, even with extensive training, it is difficult for raters to agree on a score. On a 6-point scale, two raters would typically assign the same score to a particular essay only half of the time. In spite of these scoring difficulties, the writing assessment field clearly prefers direct assessment over indirect assessment (Eliott, 2005), even while it continues to wrestle with the validity of essay tests (e.g., Huot, 1996).

The subjectivity-related problems of human scoring present certain opportunities for automated essay scoring. Scores from the automated essay scoring system e-rater® have been shown to strongly predict human holistic scores, correlating with a human rater more strongly than a second human rater, and also exhibit greater reliability over time than scores awarded by human raters (Attali & Burstein, 2006; Attali, 2007).

E-rater also allows more consistent scores, since the same set of writing measures is used in scoring essays across the developmental spectrum. This consistency should make it possible to evaluate a student's development from year to year, as well as make comparisons across grades. In addition, the scoring of specific dimensions or traits of the essay allows for a finer analysis and control of the construct (Bennett & Bejar, 1998).

Data Sources

Data for the development of the scale was gathered from a national sample of 170 schools, representing over 500 classes from 4th, 6th, 8th, 10th, and 12th grade, and over 12,000 students. The students wrote (in 30-minute sessions) up to four essays (in two modes of writing, descriptive and persuasive) in a web-based system that provided an immediate score and feedback report to the student after an essay was submitted. The feedback report was similar to the report provided in Criterion℠, an online writing environment designed to give students feedback on writing, developed by Educational Testing Service.

Students wrote on topics selected from a pool of 20 topics. In order to allow greater comparability across grade levels, topics were allocated to classes in up to three grade levels (e.g., a topic was presented to 4th, 6th, and 8th grade classes).

The study took place over two consecutive years during the spring of the school year. Not all students completed all four essays assigned to them; on average students completed three essays. A small subsample of students repeated the study in both years.

Table 1: **Features Used in the Present Study**

Feature	*Description*
Grammar	Based on rates of errors such as fragments, run-on sentences, garbled sentences, subject-verb agreement errors, ill-formed verbs, pronoun errors, missing possessives, and wrong or missing words
Usage	Based on rates of errors such as wrong or missing articles, confused words, wrong form of words, faulty comparisons, and preposition errors
Mechanics	Based on rates of spelling, capitalization, and punctuation errors
Style	Based on rates of cases such as overly repetitious words, inappropriate use of words and phrases, sentences beginning with coordinated conjunctions, very long and short sentences, and passive voice sentences
Essay Length	Based on number of words in the essay
Vocabulary	Based on frequencies of essay words in a large corpus of text
Word Length	Average word length

Methods

Initial analyses of the data were designed to create scale scores on a single developmental scale, based on seven e-rater measures computed for each essay (see Table 1 for a description).

First, the representativeness of the school and student sample with respect to the national population was assessed and parameters for the correction of biases in the sample were developed. Biases were assessed with respect to the number of students in each grade level, school type (public or private), school locality (city, urban, or rural), and percent of minority students. Within each subgroup, the discrepancies in the number of students between study sample and population were assessed and a weighting factor for students from each subgroup was developed (that is, students in oversampled subgroups were given a lower weight, and vice versa). Overall, the discrepancies between the study sample and the population were not large, and their effect on comparisons of group performance was negligible.

Following these analyses, scale scores were created as a weighted linear combination of the seven essay measures. The weights were determined on the basis of the standardized scoring coefficients of a factor analysis of the measures with one factor. The scale scores were standardized across the entire sample of essays, taking into account the biases in the sample that were quantified in the first phase.

Figure 1 shows a histogram of the scaled scores by grade, together with the fitted normal distribution. The overall mean and standard deviation of the scaled scores are 0 and 1, respectively. The increase in average performance is .64 between grade 4 and 6, and drops to .21 between grades 10 and 12. The histogram also shows a slightly lower variability in grade 4 compared to other grades.

Following the creation of scale scores, several studies and analyses were performed to evaluate the feasibility of the developmental scale. The purpose of the first set of analyses was to estimate the degree to which different topics are associated with different score levels. An important assumption of the developmental scale is that different topics are interchangeable. That is, if scores from different topics are to be used interchangeably to estimate student developmental progress, student mean scores across different topics should be similar.

The purpose of a second study was to compare human raters to the automated scale in terms of their sensitivity to student developmental progress across grade levels. The scale may, for example, under-estimate the progress of students because it

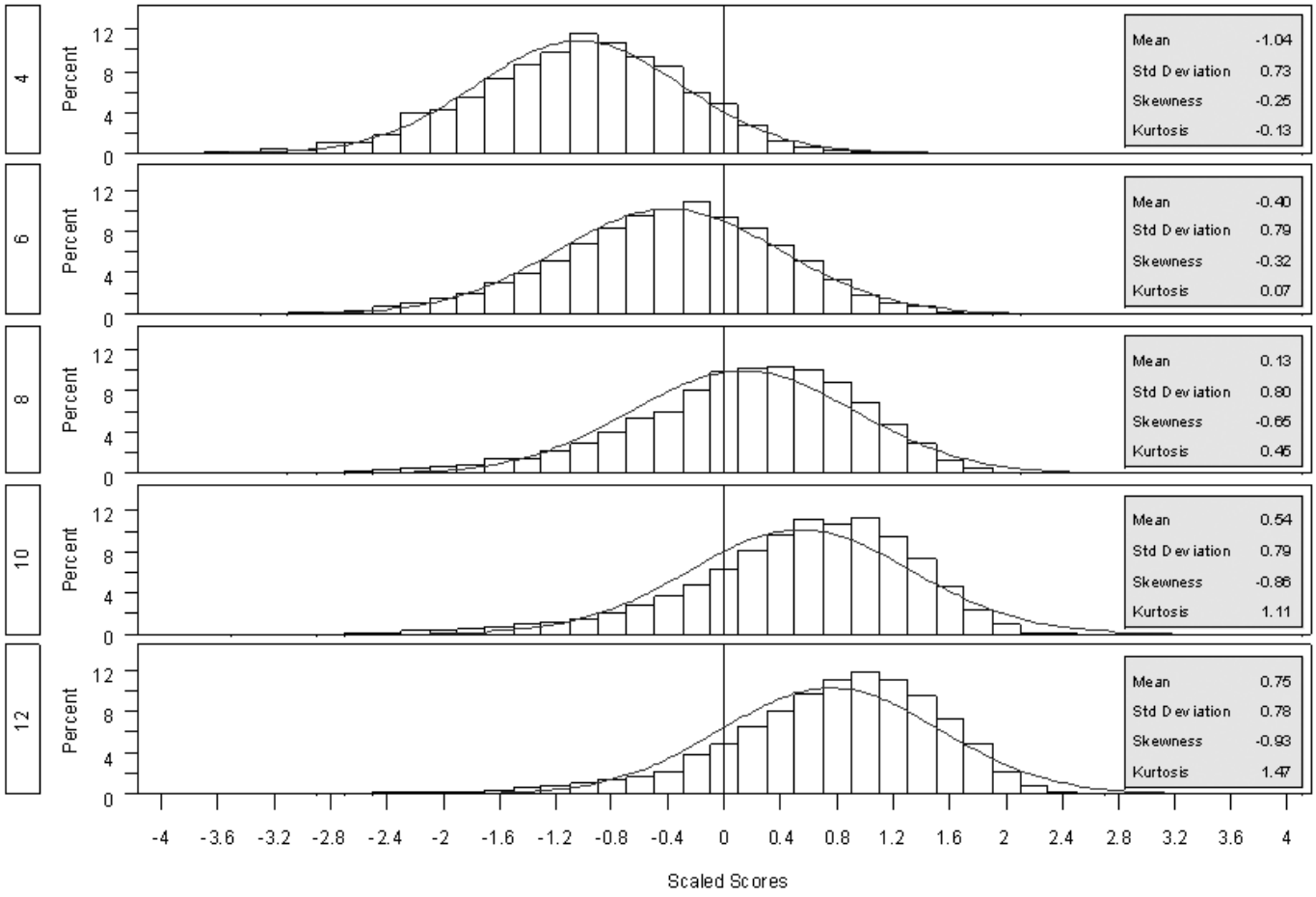

Figure 1. **Histogram of scale scores by grade level**

may not be sensitive to particular aspects of writing that human raters attend to. To perform this comparison, experienced raters scored the essays of a sub-group of study participants from 6th, 8th, and 10th grade who wrote essays on a single pair of topics (one descriptive and one persuasive). However, half of the student essays were presented as 6th grade essays and half as 10th grade essays, although both groups of essays were written by students from all three grade levels. The main research question was whether an interaction between (true) grade level (6, 8, and 10) and score mode (human or machine) would be revealed.

The goal of the third study was to validate the cross sectional predictions of grade level writing performance in a direct longitudinal dataset, available for the students who participated in both years of the study. Performance of around 400 students from 4th, 6th, and 10th grade was available in the subsequent year, when they attended 5th, 7th, or 11th grade. The expected gains of these students across a one-year interval were compared with their actual gains over this time period to confirm the predictions of the developmental scale.

The purpose of the fourth study was to investigate whether the underlying structure of writing performance as it is measured by the developmental scale is similar across grade levels. This is important in order to verify that the meaning of scores across grade levels is similar. To this end, a multiple-group confirmatory factor analysis across the five grade levels was performed to evaluate whether a common factor structure could be supported across grade levels. Several candidate factor structures were examined, and the best supported structure was also tested with respect to its invariance (in factor loadings, error variances, and factor correlations) across grade levels.

Results

A brief summary of results of the four studies follows.

Topic effect. The estimation of differences in topic difficulty was performed using a hierarchical analyses where the lower-level essay scores are cross-classified by two higher-level factors, students and topics, and these two factors are treated as random effects. Additionally, topic and student characteristics

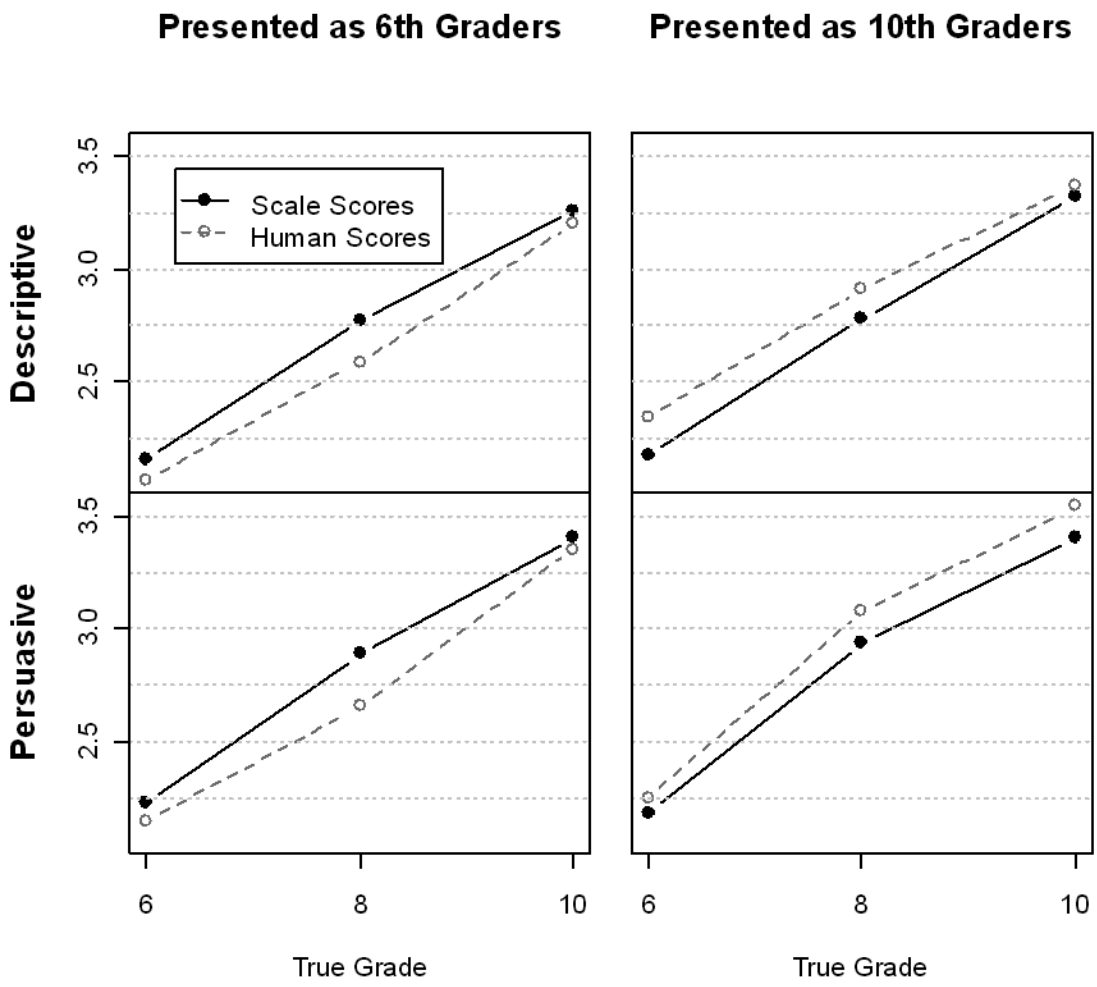

Figure 2. **Profiles of scores for true grade, writing mode, and grade presentation**

were added to the model as predictors of essay scores. Specifically, student grade level explained 30% of student variance, and the expected average increase in scores between two adjacent grade levels was .22. After taking grade level into account, only 2% of score variance could be attributed to differences in topic difficulty. Furthermore, topic mode was also a significant predictor of essay scores (persuasive essays were associated with lower scores than descriptive essays, by .14 on average). This predictor explained 50% of the variance in topics. Therefore, after taking into account student grade level and topic mode, only less than 1% of essay score variance could be attributed to differences in topic difficulty. These results support the use of topics interchangeably in the context of a developmental scale.

Human scoring study. As was explained above, in this study human raters scored 6th, 8th, and 10th grade essays believing that they were written by either 6th graders or 10th graders. Automated scale scores were compared to the human scores of these essays. The main research hypothesis was that there would not be an interaction between the (true) grade level of students and type of scoring (human or automated). To test this hypothesis, a repeated-measures ANOVA was conducted with the mode of writing (descriptive and persuasive) and type of scoring (human and e-rater) as within-subjects independent measures and with true grade level and apparent grade level (6th or 10th grade) as between-subjects independent measures. As expected, the interaction was not significant, $F(2, 884) = .81$, $p = .44$. Moreover, none of the three-way interactions with mode of writing or apparent grade level was significant (and neither the four-way interaction). Figure 2 shows that in all cases the trends of mean human and scale scores are almost parallel. These results support the premise that automated scale scores and human scores are equally sensitive to performance differences across grade levels.

Longitudinal study. In order to compare the *expected* gains of students across a one-year interval with their *actual* gains over this time period, scale scores were converted to grade-standardized scores at every grade level[1].

Grade standardization allows a natural comparison of scores across grade levels; the meaning of equal grade-standardized scores across two years is that expected gains were confirmed by actual gains. For example, Table 2 shows that the average performance of 4th-grade repeaters in their first year was .38 *SD* higher than the average 4th-grade student. The average performance of the same students the following year was .42 *SD* higher than the average *5th grader*. A repeated-measures ANOVA was conducted to evaluate the hypothesis that grade-standardized scores did not change from the first to the second year. As expected, this effect was not significant in each of the three grade levels of repeating students.

Factor analyses. Previous exploratory factor analysis (Attali, 2007) suggested three possible structures for the developmental data: a single factor solution; a two factor solution with essay length and style as a fluency factor and all other features as a second factor; and a three-factor solution with the fluency factor, a grammatical conventions factor (with the grammar, usage, and mechanics features) and word usage factor (with vocabulary and word length features). A confirmatory factor analysis revealed that the three-factor solution best fit the data. Table 3 presents the factor loadings, factor correlations, and error variances across grade levels. Invariances in factor loadings and error variances could not be supported across the five grade levels, but some support for invariance in factor correlations across grade levels was found.

Importance of the Study

By allowing greater comparability of scores within and across grade levels, the developmental writing scale offers some advantages for improving the current practice of writing assessment. Based on the results of the factor analysis, it seems possible from a psychometric perspective to provide scores for three components of writing (word choice, conventions, and fluency), and thus to further enhance the assessment of writing. These improvements in assessment may in turn allow a better understanding of the development of writing proficiency as it is manifested in essay writing. For example, the factor

***Table 2*: Mean (and SD) of Grade-Adjusted Repeater Scores**

Grade (Year 1)	N	Year 1	Year 2
4	125	.38 (.83)	.42 (.78)
6	221	.33 (.77)	.36 (.77)
10	55	.50 (.58)	.35 (.59)

***Table 3:* Three-Factor Model: Fluency (F), Conventions (C), and Word Choice (W)**

	G4	G6	G8	G10	G12
Factor loadings					
Essay Length (F)	.99	1.01	1.06	.93	.92
Style (F)	.51	.63	.58	.58	.50
Grammar (C)	.70	.77	.84	.83	.84
Usage (C)	.63	.62	.61	.61	.59
Mechanics (C)	.36	.38	.43	.45	.42
Vocabulary (W)	.59	.49	.71	1.08	1.30
Word Length (W)	.60	.68	.79	.81	.85
Factor correlations					
F ↔ C	.86	.85	.75	.74	.74
C ↔ W	-.12	.07	.13	.20	.13
F ↔ W	-.16	.03	.05	.15	.12
Error variances					
Essay Length	.09	.01	.02	.07	.07
Style	1.34	.84	.66	.47	.42
Grammar	.32	.37	.31	.37	.33
Usage	.29	.48	.69	.75	.71
Mechanics	.82	.79	.79	.86	.87
Vocabulary	.52	.48	.35	.02	.30
Word Length	.43	.19	.31	.52	.66

1 Since norming data was not available for odd grade levels of the second year (5th, 7th, and 11th grade), grade standardization was based on interpolation of the trajectories of mean and standard-deviations of performance across the even grade levels.

analysis suggests that fluency and conventions of writing are not fully distinguished in lower grades, and that fluency is more dominant in these lower grades, whereas word choice becomes dominant in higher grades. Thus, we believe that our efforts may enable greater diagnosis as well as a more accurate assessment of progress in the development of writing skills.

References

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL® essays* (ETS Research Rep. No. RR-07-21). Princeton, NJ: Educational Testing Service.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Available from http://www.jtla.org.

Bennett, R. E., & Bejar, I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Eliott, N. (2005). *On a scale: A social history of writing assessment in America*. New York: Peter Lang Publishing.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*, 549-566.

Using Standard-Setting Methodology for Linking Assessment Scores to Proficiency Scales: TOEFL iBT™ and TOEIC® Assessment Exemplars

Richard J. Tannenbaum and E. Caroline Wylie

Editor's note: *There are instances when it is necessary to link scores on an assessment to some kind of proficiency scale even though consideration of the proficiency scale was not part of the original assessment development process. If another test exists that places students on that proficiency scale, then a study can be conducted to relate performances on the two tests, thereby empirically linking scores on the new assessment to the proficiency scale. However in certain circumstances there is no existing test that has scores based on the proficiency scale. In such cases, a judgmental approach is required to create the linking. This study represents such an occasion.*

The Common European Framework Reference for Languages: Learning, Teaching, Assessment (CEFR) is intended to overcome barriers to communication among language instructors, educators, curriculum designers, and agencies working in the field of language development by providing a common basis for describing and discussing stages of language development and the skills needed to reach different levels of language proficiency (The Common European Framework of Reference). The CEFR describes language proficiency in reading, writing, speaking, and listening on a six-level scale, clustered in three bands: A1–A2 (Basic User), B1–B2 (Independent User), and C1–C2 (Proficient User).

The CEFR scales are becoming accepted in Europe as one means of reporting the practical meaning of test scores in ways that have a socially constructed meaning for teachers and other test-score users. That is to say, if a test score can be mapped (linked) to one of the levels of the CEFR, it becomes clearer what that score means—what candidates with at least that score are likely able to do. While the CEFR is not without its detractors (see, for example, Weir, 2005), the CEFR is widely accepted as the benchmark against which language tests used across Europe should be compared.

The purpose of this study was to identify minimum scores (cutscores) on two English-language tests (the TOEFL iBT™ and TOEIC® assessments) that correspond to the A1 through C2 proficiency levels of the CEFR. Minimum scores were to be identified separately for the Speaking, Writing, Listening, and Reading sections of the two assessments.

By mapping test scores onto the CEFR, an operational bridge is built between the descriptive levels of the CEFR and psychometrically sound, standardized assessments of English-language competencies, facilitating meaningful classification of test takers in terms of CEFR-based communicative competence as well as tracking progress of test takers in English-language development. The study was not intended or designed, however, to establish a concordance between scores on the two English-language assessments, such that scores on one assessment could be used to identify comparable scores on the other assessment. Scores from each assessment were independently mapped to the CEFR levels; no attempt was made to link scores or score distributions across the assessments.

Linkages were determined through expert judgment, following standard-setting procedures: a modified Angoff method for Listening and Reading (selected-response) sections (Brandon, 2004; Cizek & Bunch, 2007) and a performance-sample approach—a hybrid of judgmental policy capturing (Jaeger, 1995) and dominant profile approaches (Plake, Hamilton, & Jaeger, 1997)—was implemented for Speaking and Writing (constructed-response) sections. Recent reviews of research on standard-setting approaches also reinforce a number of core principles for best practice: careful selection of panel members and a sufficient number of panel members to represent varying perspectives; sufficient time devoted to ensure development of a common understanding of the domain under consideration; use of an appropriate standard-setting methodology that allows for adequate training of judges; development of a description of each performance level; multiple rounds of judgments; and the inclusion of empirical data where appropriate to inform judgments (Brandon, 2004; Hambleton & Pitoniak, 2006). The approaches used in this study adhere to all of these guidelines.

Methodology

Two panels of English language instructors, administrators or directors of language programs, and language testing experts from various European countries participated. Panel 1 focused on the TOEFL iBT assessment and consisted of 23 experts from 16 countries. All members of Panel 1 were familiar with the TOEFL iBT assessment and the test takers who typically took that assessment. Panel 2 focused on the TOEIC assessment and consisted of 22 experts from 10 countries. All members of Panel 2 were familiar with the TOEIC assessment and the test takers who typically took that assessment. Five experts, familiar with both assessments, served both on Panel 1 and Panel 2.

Before the studies, experts on both panels were given an assignment to review selected tables from the CEFR for each language modality and to note key characteristics or indicators from the tables that described an English language learner (candidate) with *just enough skills* to perform at each of the CEFR levels. The tables were selected to provide panelists with a broad understanding of what learners are expected to be able to do for each of the language modalities. As they completed this pre-study assignment, they were asked to consider what distinguishes a candidate with just enough skills to be considered performing at a CEFR level from a candidate with not quite enough skills to be performing at that level. For example, they were asked to consider what the least able C2 speaker can do that the highest performing C1 speaker cannot do, what the least able C1 speaker can do that the highest performing B2 speaker cannot do, and so on. The assignment was intended as part of a calibration of the panelists to a shared understanding of the minimum requirements for each CEFR level.

During the study, the panelists defined the minimum skills needed to reach each level of the CEFR. The panelists worked in three small groups, with each group defining the skills of the least able candidate for the A2, B2, and C2 levels; this was done separately for Writing, Speaking, Listening, and Reading. Panelists referred to their pre-study assignments and to the CEFR tables for each modality. A whole-panel discussion occurred for each level and a final definition for each level was established. Definitions of the least able candidate for A1, B1, and C1 levels were accomplished through whole-panel discussion, using the A2, B2, and C2 descriptions as "boundary markers." As before, the panelists also referred to their pre-study assignment and the relevant CEFR tables. These definitions served as the frame of reference for standard setting judgments; that is, panelists were asked to consider the test items in relation to these definitions.

A modified Angoff approach—consistent with the standard setting process outlined in the *Manual for Relating Language Examinations to the CEF*[*R*] (2003, Council of Europe)—was implemented for Reading and Listening modalities measured using selected-response items. Panelists were trained in the process and then given the opportunity to practice making their judgments. At this point, panelists were asked to sign a training evaluation form confirming their understanding and readiness to proceed, which all panelists did. Then they went through three rounds of operational judgments, with feedback and discussion between rounds. For each item, panelists were asked to consider the agreed upon definition of just-qualified (least able) candidates (for A2, B2, C2) and to judge the probability that a just-qualified (least able) candidate within the level would have the skills needed to answer the item correctly.

In order to facilitate setting six cutscores on each modality, panelists initially focused on A2, B2, and C2 levels; once established, these cutscores formed the boundaries for the A1, B1, and C1 cutscores. For example, the lower boundary for B1 was the cutscore for A2 and the upper boundary was the cutscore for B2. The task for the panelists was determining where within that range to locate the B1 cutscore.

Panelists were asked to use the following judgment scale (expressed as probabilities): 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 100. The higher the probability, the more likely the item would be answered correctly by just-qualified A2, B2, and C2 candidates. Panelists were instructed to focus only on the alignment between the skill demanded by the item and the skill possessed by a just-qualified candidate, and not to factor guessing into their judgments.

The sum of each panelist's cross-item judgments represents his or her recommended cutscore. Each panelist's recommended cutscore was provided to the panelists. The panel's average (panel's recommended cutscore), and the highest and lowest cutscores (unidentified) were compiled and presented to the panel to foster discussion. Panelists were then asked to share their judgment rationales. As part of the feedback and discussion, item performance information and P+ values (proportion of test takers answering each item correctly) were shared. In addition, P+ values were calculated for candidates scoring at or above the 75th percentile on that particular section (i.e., the top 25% of candidates) and for candidates at or below the 25th percentile (i.e., the bottom 25% of candidates). Examining item difficulty for the top 25% of candidates and the bottom 25% of candidates was intended to give panelists a better understanding of the relationship between overall language ability for that modality (total section score) and each of the items. The partitioning, for example, enabled panelists

to see any instances where an item was not discriminating, or where an item was found to be particularly challenging or easy for test takers at the different ability levels.

Before making their round two judgments, panelists were asked to consider their peers' rationales and the normative information. For round two, judgments were made, not at the item level, but at the overall level of the modality (section); that is, panelists were asked to consider if they wanted to recommend a different section-level score for A2, B2, and C2. The transition to the section (modality) level introduced a shift from discrete items to the overall construct of interest. This holistic approach seemed more relevant and appropriate to the language construct of interest than did deconstructing the construct through another series of item-level judgments. Panelists had no difficulty with the holistic approach; this approach had also been used in a previous CEFR linking study (Tannenbaum & Wylie, 2005).

After making their second round of judgments, similar feedback was provided, but in addition, the percentage of candidates who were classified into each of the three levels was presented. The round-two average judgments for A2, B2, C2, were applied to existing test score distributions for the modality of focus and the percentages of candidates classified into each level was presented and discussed. Following this level of feedback, the panelists had a final opportunity to change their section-level recommended cutscores. These final judgments were compiled and shared with the panelists; they were then asked to locate the A1, B1, and C1 levels. Specifically, they were asked to review the A1, B1, and C1 descriptions of just-qualified candidates and to identify the minimum section-level scores for candidate just performing at these levels. Their judgments were constrained by the now-established A2, B2, and C2 cutscores. Panelists had an opportunity to discuss whether they considered any of the threshold proficiency levels to be located closer to one boundary than another. Once there had been a wide-ranging discussion, panelists then made their final individual judgments as to the minimum score associated with A1, B1, and C1 levels.

A performance-sample approach was implemented for the Speaking and Writing modalities measured using constructed-response items. As with the modified Angoff approach, three rounds of judgments took place, with feedback and discussion, informed by data (average item scores—instead of P+ values—partitioned as described above, and classification information). Panelists were asked to review the scoring rubrics and then to review (listen to or read) samples of candidate performance across items (i.e., profiles of performance) at various points along the raw point scale for that modality. They were then asked to identify the score for that modality that would be expected of just-qualified candidates. Once the three rounds for A2, B2, and C2 were completed, the panelists were, as before, asked to locate the cutscores for the A1, B1, and C1 levels.

Panelists had the option of writing N/A (not applicable) for a cutscore if they deemed that the test section was not challenging enough to reach the upper levels of the CEFR, or if the test section was too challenging for candidates at the lower CEFR levels. In order for a cutscore to be reported, at least 67% of the panel had to make a cutscore recommendation. All cutscore decisions and subsequent discussions were based on raw scores, or the number of points expected to be earned by a just-qualified candidate on the form of the test reviewed.

Results

Tables 1 and 2 present the recommended cutscores for each modality for the tests reviewed. The cutscore represents the minimum score judged necessary to enter each CEFR proficiency level. As can be seen in Table 1 for the TOEFL iBT assessment, Panel 1 did not believe that the Writing, Listening, and Reading sections of the test were accessible to just-qualified candidates at the A1 level. At this level candidates may only be expected to recognize familiar words or to understand very short simple texts, one phrase at a time. Panelists believed that these limited skills were exceeded by the TOEFL iBT assessment. The panelists stated that these sections were too demanding for such candidates. The panel also believed that Listening and Reading sections were too demanding for just-qualified candidates at the A2 level. Conversely, the panel believed that the Writing, Speaking, and Listening sections were not challenging enough to recommend cutscores for just-qualified candidates at the C2 level. Overall, these results suggest that TOEFL iBT discriminates at the B1 through C1 levels of the CEFR.

The results for the TOEIC assessment are summarized in Table 2. Panel 2 believed that Writing, Speaking, Listening, and Reading sections were not challenging enough to recommend cutscores for just-qualified candidates at the C2 level for similar reasons to those expressed by Panel 1. The panel held the same view for Reading at the C1 level.

At the conclusion of standard setting for each test, panelists were asked to complete an evaluation form. This form served the purpose of collecting information about the perceived quality of the standard setting process. Panelists were asked to rate the clarity with which various aspects of the study were presented, and were asked to indicate overall their level of comfort with the full set of recommended cutscores. For both tests, the majority

Table 1: **Scaled Score Cutscore Results for the TOEFL iBT™ Assessment**

	Writing (maximum 30 points)	Speaking (maximum 30 points)	Listening (maximum 30 points)	Reading (maximum 30 points)
A1	-	8	-	-
A2	11	13	-	-
B1	17	19	13	8
B2	21	23	21	22
C1	28	28	26	28
C2	-	-	-	29

Table 2: **Scaled Score Cutscore Results for the TOEIC® Assessment**

	Writing (max. 200 points)	Speaking (max. 200 points)	Listening (max. 495 points)	Reading (max. 495 points)
A1	30	50	60	60
A2	70	90	110	115
B1	120	120	275	275
B2	150	160	400	385
C1	200	200	490	-
C2	-	-	-	-

of panelists indicated that the homework assignment was useful preparation, that the purpose of the study and instructions were clear, training was sufficient, and the feedback/discussion process was helpful. Additional prompts asked panelists about what was most influential in their decision making process. For both tests, the definition of the *just-qualified candidate* and the panelists' own professional experience were the two most influential factors. Finally, each panelist was asked to indicate their level of comfort with the final results. For both tests, the modal response was *very comfortable*.

Discussion

In accordance with the goals of this study, linkages were established between each section of the TOEFL iBT assessment and levels B1, B2, and C1 of the CEFR, and between each section of the TOEIC assessment and levels A1 through C1 of the CEFR, with the exception of Reading at the C1 level.

The difficulty of linking test scores to the CEFR should not be underestimated. The CEFR, according to Weir (2005), does not provide sufficient information about how contextual factors affect performance across the levels or adequately delineates how language develops across the levels in terms of cognitive processing. This may lead to difficulties in interpreting differences across the CEFR levels. Some of this was evident during the panelist discussions of the CEFR when developing the just-qualified descriptions. Panelists noted that the descriptive language of the CEFR was not consistently applied across the levels, making it more difficult for them to differentiate among the levels. The difficulty, however, also is a function of the tests. It is more likely that tests developed specifically to map to the CEFR would pose less of a linking challenge than tests relying only on a post hoc approach, as was the present case. Although the tests considered in this study measured the basic communicative modalities, all covered by the CEFR, the items on the tests were not specifically developed to operationalize these modalities necessarily as depicted by the CEFR. Although this did not preclude setting cutscores for some of the levels, it most likely was the reason why not all intended CEFR levels were mapped. The value of using level descriptors to inform test development, thereby increasing alignment and the potential meaningfulness of cutscores, was recently noted by Bejar, Braun, and Tannenbaum (2007) in the context of No Child Left Behind testing.

Although not all targeted CEFR levels were mapped, there was positive evidence of the quality of the standard setting process. The majority of panelists for each test reported that they were adequately trained and prepared to conduct their standard-setting judgments, and that the standard-setting process was

easy to follow. Panelists reported that the definition of the just-qualified candidate most influenced their judgments and that they were able to use their professional experience to inform their judgments. Furthermore, the majority of panelists reported that they were comfortable with the recommended cutscores. Procedural validity is an important criterion against which to evaluate the quality of the standard-setting process (Hambleton & Pitoniak, 2006; Kane, 2001).

External validity evidence is also desirable and most often takes the form of convergence with other sources of information (Hambleton & Pitoniak, 2006; Kane, 2001). In the present case, for example, convergent evidence could be obtained from teacher ratings of their students' English-language proficiency in terms of the CEFR (Council of Europe, 2003). Although a convergence of evidence would lend further support of the reasonableness of the panel-based cutscores, the meaning of a divergence of evidence is less clear, given that there is no true cutscore. "Differences in results from two different procedures would not be an indication that one was right and the other wrong; even if two methods did produce the same or similar cutscores, we could only be sure of precision, not accuracy" (Cizek & Bunch, 2007, p. 63). With this in mind, the cutscores from this study should be considered recommendations only; they are not absolutes. Potential users of these cutscores are advised to consider their specific needs and circumstances, and other relevant information that may be germane to determinations of the English-language proficiency of their test takers that was not part of this set of studies. It is reasonable for users to adjust these recommended cutscores to better accommodate their needs.

This set of standard-setting studies, we believe, represents a significant step forward in the evolution of research concerned with linking test scores to the CEFR. The use of a performance-sample approach for constructed-response items, which enabled panelists to consider profiles of responses; the inclusion of item-data partitioned by test-taker ability levels; the shift from item-level judgments in the first round for the selected-response items to a more holistic judgment for the subsequent rounds; and the locating of the of the Level 1 cutscores in relation to the Level 2 cutscores all reflect innovative and creative design elements in research studies whose primary objective is relating test scores to the CEFR. Continued advances in this area of applied research would seem warranted, given the increasing emphasis (and hence importance) of being able to interpret the meaning of test scores in terms of the proficiency levels of the CEFR.

References

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1-30). Maple Grove, MN: Journal of Applied Metrics Press.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17, 59-88.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.

Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEF)* (Manual: Preliminary pilot version. DGIV/EDU/LANG, 2003, 10). Strasbourg, France: Council of Europe, Language Policy Division.

Hambleton, R. K. & Pitoniak, M. J. (2006). Setting performance standards. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, *8*, 15-40.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Plake, B. S., Hamilton, R. K., & Jaeger, R. M. (1997). A new standard setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement*, *57*, 400-411.

Tannenbaum R.J. & Wylie, E.C. (2005). *Mapping English language proficiency test scores onto the Common European Framework* (ETS Research Rep. No. RR-05-18). Princeton, NJ: Educational Testing Service.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*, 281–300.

Equating of Mixed-Format Tests in Large-Scale Assessments

Sooyeon Kim, Michael E. Walker, and Frederick McHale

Editor's note: *As a way of balancing the measurement validity of constructed-response items with the cost-effectiveness of multiple choice items, tests often include both types of items. Like all assessments, such mixed-format tests are more meaningful when their results are more reliable. In this study, ETS psychometricians have examined ways to link different forms of tests that contain both types of items.*

Many large-scale testing programs increasingly make use of constructed response (CR) items in their assessments, often in conjunction with multiple-choice (MC) items. MC items are economically practical and ensure objective and reliable scoring, whereas CR items tend to be difficult to score objectively and reliably. However, proponents argue that CR items tend to resemble more closely the real-world tasks associated with the construct to be measured. Because both MC and CR items display strengths as well as weaknesses, many assessments tend to be of mixed format, including both MC and CR items. Mixed format tests pose some challenges in the area of equating. This study examined several procedures for equating mixed-format tests to evaluate the most effective procedures.

Perhaps the most commonly used equating design is the non-equivalent groups with anchor test (NEAT) design. An anchor composed of items common to the two forms being equated is used to adjust for differences in the ability of the groups taking each form. A major difficulty when trying to equate tests with a CR component is the difficulty of identifying a satisfactory anchor test. CR items are typically not reused across different test forms because of ease of memorization. Using an all-MC anchor will lead to biased equating results (Kim & Kolen, 2006; Li, Lissitz, & Yang, 1999); possibly because MC and CR items measure somewhat different constructs (Bennett, Rock, & Wang, 1991; Sykes, Hou, Hanson, & Wang, 2002). Even if CR items are reused, raters could change their scoring standards from one test administration to the next. In this case, the CR anchor items would confound differences in rater severity with true group ability differences.

One possible equating design that avoids the use of an anchor is an equivalent groups (EG) design, where the two test forms (reference and new) are spiraled in a single administration. When feasible, this equating design is preferable to a NEAT design because there is no need to adjust for group ability differences in the equating. Use of this procedure would be based on the assumption, however, that the previously administered reference form to which the new test form would be equated behaved identically in the current administration as in the previous one. Changes in scoring severity for the CR items would make this assumption untenable.

Trend Scoring Method

Tate (1999; 2000) articulated a solution to the problem of subjective or changing scoring standards in the context of the NEAT design. He suggested a preliminary linking study in which any across-year changes in rater severity could be isolated, so that across-group ability differences could be accurately assessed and the tests could be properly equated. The linking study involves rescoring responses to the CR anchor items obtained from the reference population. A representative sample of anchor item papers for examinees from Year 1 (the reference year) is inserted into the rating process for Year 2 (the new year). The responses, obtained from the reference group of examinees, are rescored by the raters scoring responses for the same items for the new group of examinees. Thus, these *trend papers* have two sets of scores associated with them: one from the old set of raters and one from the new rater group. Simply, trend scoring is rescoring the same examinee papers across scoring sessions. Such a rescoring eliminates any effect of group (the same group is scored at both sessions) and allows for the detection of any scoring shift across sessions.

Purpose

The purpose of this study was to examine systematically four procedures to place the mixed-format new form on scale with the mixed-format reference form. The four procedures make use of different equating designs and anchor compositions. The first three procedures operated in the context of a NEAT design, and the fourth followed an EG design. Four linking designs were examined: (a) an anchor with only MC items; (b) a mixed-format anchor test containing both MC and CR

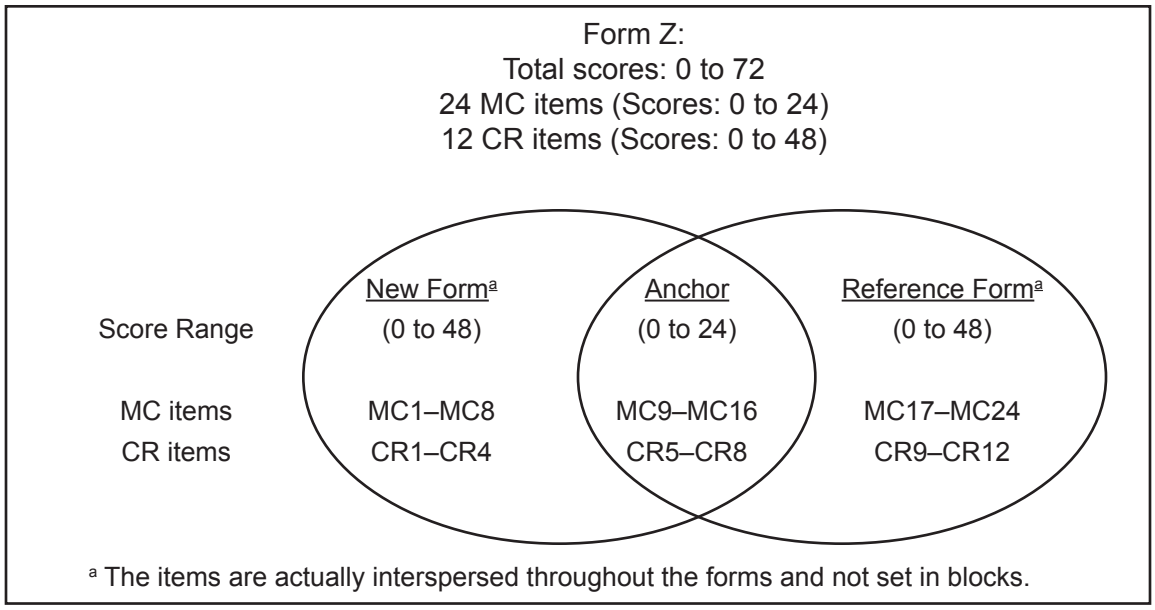

Figure 1. **Schematic of two parallel forms, new form and reference form.**

items; (c) a mixed-format anchor test incorporating CR anchor item rescoring (i.e., trend scoring); and (d) an EG design with rescoring for all CR items, thereby avoiding the need for an anchor test. Two major questions were of interest: (1) Which equating design is the most effective for linking tests with CR items? (2) What anchor test composition (a mix of MC and CR, or MC-only) works best in the NEAT design?

Method

Data

The data were taken from two administrations of a subject test, comprising 24 MC and 12 CR items (called Form Z), of a large-scale testing program. For each examinee, each CR item was scored independently by a single rater on a 0-to-2 scale weighted by 2 (such that each CR item score could range from 0 to 4). For one administration, the 12 CR items for 417 examinees were scored by Rater Group A. These 417 examinees constituted the reference group in this study. In another administration, the same 12 CR items for those 417 examinees were independently scored by Rater Group B. These same raters (Group B) also scored the 12 CR items for a separate group of examinees ($N = 3{,}126$). These 3,126 examinees constituted the new group in this study. Note that two independent sets of scores for all CR items were available for the 417 reference examinees, but only a single set of CR scores was available for the 3,126 new examinees.

Simulated Forms

Two parallel forms (designated new form and reference form) were created from the original test form (Form *Z*). Figure 1 shows the basic layout for the two parallel forms. The new and reference forms each consisted of 16 MC and 8 CR items. Those forms had 8 MC and 4 CR items in common, which were used as the anchor in a NEAT design. The construction of two forms from a test given at a single administration allowed us to mimic the typical equating of alternate forms while having the advantage of yielding data from a single group of examinees that took all of the items on both forms.

Criterion

For the purposes of the study, the reference form as scored by Rater Group A (reference form/Rater A) served as the reference form and the new form as scored by Rater Group B (new form/Rater B) served as the new form. The criterion represented the true linking of the new form/Rater B combination to the reference form/Rater A combination. This linking was estimated using a single group design making use of the 417 examinees whose CR items were scored by both sets of raters. The schematic of this design is presented in the upper section of Figure 2. Because the differences between the linear and nonlinear functions were negligible for almost all raw score points, the linear function (i.e., setting means and standard deviations equal) was used as the criterion.

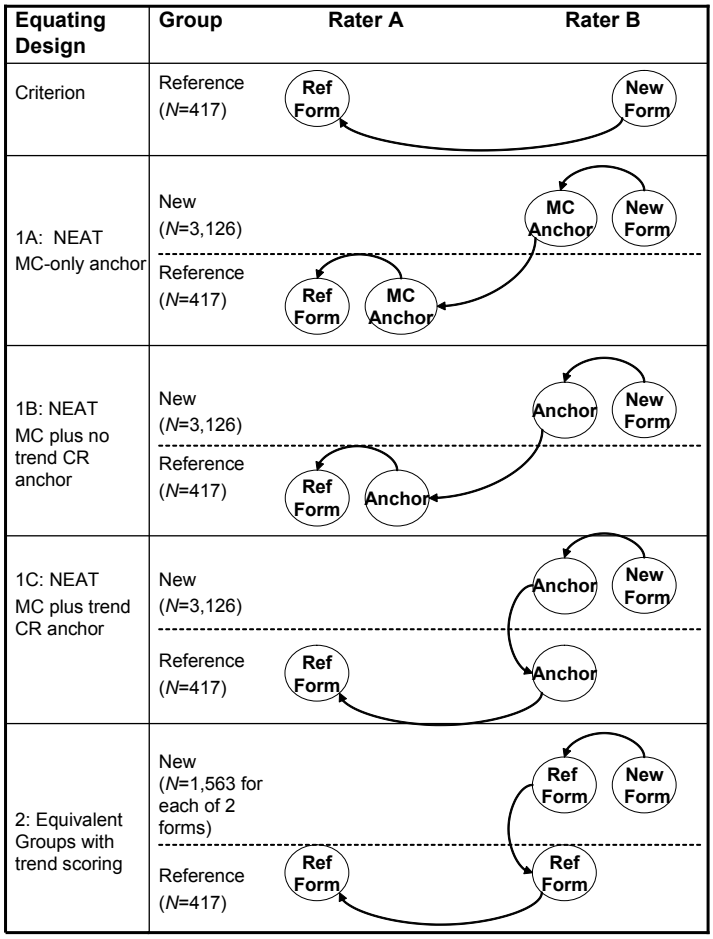

Figure 2. **Schematic of the criterion and equating designs.**

Equating Designs

Two equating designs, (1) a NEAT design and (2) an EG design, were considered in this study. In both designs, the new form/Rater B was equated to the reference form/Rater A. The 417 examinees served as the reference population, and the 3,126 examinees served as the new form population. The lower panels of Figure 2 present the schematics of these equating designs. In this figure, the curved arrows indicate the chain of linking among the four circled scores. Generally, the chain consists of equating the scores on the new form to scores on the anchor and then equating scores on the anchor to scores on the reference form. This chain formed by these two equatings links the scores on the new form to scores on the reference form.

In the first design, the NEAT design, three different anchor compositions were examined: (A) only MC items; (B) MC and CR items, where the CR items were not adjusted for rater severity; and (C) MC and CR items, where the CR items were adjusted for rater severity via trend scoring. Figure 2 shows how, in Design 1A, the new form was equated to the reference form via the MC anchor score. In Design 1B, CR items were included in the anchor, but no rater adjustment was made. The four common CR items were scored by different sets of raters for the two forms, by Rater Group B in the new form and by Rater Group A in the reference form.

In Design 1C, by comparison, the four common CR items were scored by the same raters (Rater Group B) in both the reference and new form groups. Operationally, this was accomplished by trend scoring the CR anchor items for the reference form examinees. These rescored CR items were combined with the MC anchor items and used as the anchor score in the NEAT equating. In this way, the anchor scores were made equivalent across the new and reference form groups.

The second design, an EG design with trend scoring, represented an alternative to the NEAT design. In this case,

***Table 1*: Summary of Deviance Measures**

		Bootstrapped statistics		
Equating design	RMSD	Bias	Equating error	RMSE
NEAT design				
1A: MC only anchor	1.490	1.496	0.420	1.554
1B: MC plus no-trend CR anchor	1.593	1.603	0.238	1.620
1C: MC plus trend CR anchor	0.414	0.415	0.360	0.549
EG design with trend scoring	0.129	0.084	0.401	0.410

the new sample ($N = 3{,}126$) was randomly split to spiral the new form with the reference form. Then the new form scores for the 1,563 examinees were linked to the reference form scores for the other 1,563 examinees in an EG design. For the reference form group ($N = 417$), all CR items were rescored by Rater Group B, who also scored the new form group (i.e., the CR items were trend scored for the reference form group). In Figure 2, the reference form is listed on the right-hand side of the figure for both the new sample ($N = 1{,}563$) and the reference sample ($N = 417$). These scores are directly comparable, because they represent the same test form scored by the same raters (Rater group B). For the 417 examinees, Figure 2 also shows a reference form on the left side of the diagram. This represents the reference form as originally scored by Rater Group A. The scores on the reference form scored by Rater Group B were linked to the scores on the reference form scored by Rater Group A in a single-group design. In this way, Design 2 linked the new form/Rater B to the reference form/Rater B (via the EG design in the new group), to the reference form/Rater A (via the single group design in the reference group), without the need for an anchor test.

The Measure of Accuracy

The chained linear equating method (Livingston, 2004) was used for all equatings. The new-form equated raw scores obtained in each equating design were compared with the criterion. The differences among the conversions were squared and then weighted by the relative proportion of the new form examinees at each score point. The square root of the sum of the differences defined the Root Mean Squared Difference (RMSD) index.

Furthermore, a total of 500 bootstrap samples were obtained in each equating design using a resampling technique to estimate equating error and bias. In each replication, examinees were randomly drawn *with replacement* from each reference and new form group until bootstrap samples consisted of the same number of examinees as in the actual reference and new form groups. Then the new form scores were equated to the reference form for each of those 500 samples in each equating design using the chained linear method. In this case, equating bias was defined as the mean difference between chained linear equating and the criterion equating over 500 replications. The standard deviation of these differences over 500 replications was used as a measure of the standard error of equating (SEE). The square root of the sum of squared bias and squared SEE defined the Root Mean Squared Error (RMSE) index.

Results

The second column in Table 1 presents the difference between the chained linear function for each design and the criterion, using the RMSD measure. Figure 3 plots the conditional equated raw score difference between chained linear equating and the criterion in each equating design across the raw score region where most examinees' scores were observed.

The EG design incorporating trend scoring yielded the smallest RMSD of the four designs. The EG design was more effective than the NEAT design in enhancing the accuracy of equating. With the anchor test design, the use of trend CR items in the anchors greatly improved equating. The RMSD value was much smaller in this mixed anchor case than in either the MC-only anchor or MC plus no-trend CR anchor cases. The MC plus no-trend CR anchor case yielded the largest RMSE of the four cases. This result clearly indicated that incorporating no-trend CR anchor information into the anchor would appear to be problematic unless CR scoring standards are well maintained and implemented consistently over time by human raters. Such a requirement is extremely difficult to meet in practice except for the most objective of scoring rubrics (see

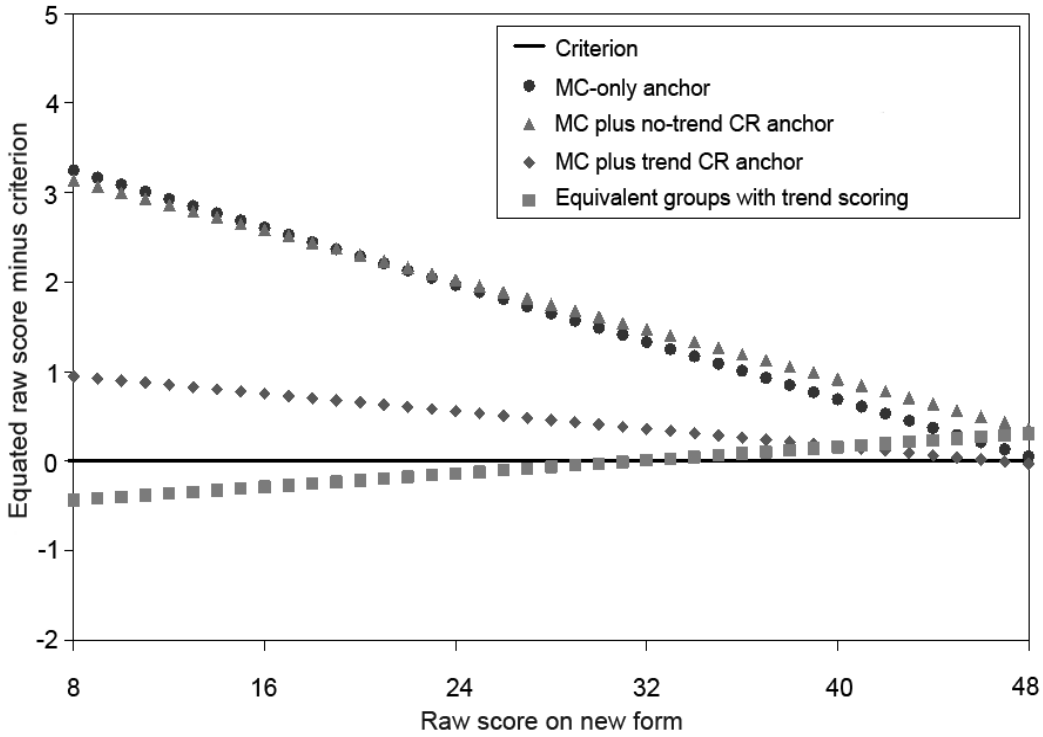

Figure 3. **Differences between chained linear equating and the criterion in the four equating designs.**

Fitzpatrick, Ercikan, Yen, & Ferrara, 1998).

The last three columns in Table 1 present the summary of the weighted average root mean squared bias, equating error, and RMSE for each equating design. Although equating error was fairly comparable for the four designs, the magnitude of bias was substantially larger in both the MC-only and MC plus no-trend CR anchor cases than in either the MC plus trend CR case or in the EG design. In general, the EG design fared well with respect to bias and equating error, leading to the smallest RMSE. The EG design exhibited near-zero bias, as one would expect because it did not rely on an anchor for group adjustments. The MC plus trend CR anchor yielded the next smallest bias and RMSE. Because the correlation between the MC anchor and total scores (r = .55 – .57) was not substantially high in this case, the use of MC items alone resulted in a large bias.

Conclusions

This study showed that equating bias caused by a scoring shift could be controlled by using a trend scoring method. The trend scoring method has statistical strengths in detecting a CR scoring shift. The trend CR anchor displayed much better performance than did the no-trend CR anchor in recovering the criterion equating function, primarily through a reduction in bias. The use of the mixed anchor might be harmful when no-trend CR items are incorporated as an anchor in the presence of a change in CR scoring standards.

Using MC items alone as anchors to control for differences among test forms containing CR items may be inappropriate when the correlation between the CR and MC components of the test are not high (as here: r = .44 – .45) due to the possible multidimensionality of mixed format tests. The MC only anchor design produced large RMSD, bias, equating error, and RMSE, compared to the mixed (MC plus trend CR) anchor and EG designs. This result is consistent with previous findings (Kim & Kolen, 2006; Li, Lissitz, & Yang, 1999).

Among the four designs, the EG design seems to be the best model psychometrically in adjusting for changes in the scoring standards for the CR common items. However, the differences observed in performance between the MC plus trend CR anchor design and the EG design are not great. There are tradeoffs between the two designs that may make one design preferable to the other. For example, some items have to be common in both test forms to use the MC plus trend CR anchor design, but this requirement is not necessary for the EG design. On the other hand, only common items need to be rescored in the mixed anchor design, but all CR items need to be trend-scored in the EG design. Only the new test form needs to be administered in the mixed anchor design, but both test forms

need to be spiraled in each administration if the EG design is used. Finally, in principle an EG design requires a substantially larger number of examinees than a NEAT design to achieve the same level of equating error. Given the limitations listed above, practitioners may choose one or the other of the NEAT or EG designs, depending upon the situation.

The present study is meaningful for two reasons. First, this study examined the effectiveness of equating designs incorporating trend scoring using non-IRT equating methods and actual data from an operational test. Second, the results of this study draw attention to an important issue, often overlooked in operational settings. In many cases, a test form containing CR items may be reused, and the original test score conversion, obtained when the test was first equated, is applied in subsequent administrations. This research demonstrates that the use of the original test score conversion may be inappropriate unless the CR scoring standards are well maintained over time. Trend scoring should be implemented for reprints so that differences in rater severity can be statistically removed through the process of equating. The findings of the present study are promising, and thus many practitioners may consider these findings when selecting an equating design when CR items are involved.

References

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28*, 77-92.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education*, *11*, 195-208.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*, 357-381.

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999, April). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.

Sykes, R. C., Hou, L., Hanson, B., & Wang, Z. (2002, April). *Multidimensionality and the equating of a mixed-format math examination*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, *36*, 336-346.

Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*, 329-346.

Multiple Methods of Assessing Information Literacy: A Case Study

Irvin R. Katz, Norbert Elliot[1], Yigal Attali, Davida Scharf[1], Donald Powers, Heather Huey[1], Kamal Joshi[1], and Vladimir Briller[2]

Editor's note: *Virtually every aspect of higher education is now touched by technology. Today's college and university students have grown up with technology, but how effectively can they use the information that they access through technological environments? In this study, ETS research scientists collaborated with staff and faculty from the New Jersey Institute of Technology to examine two measures of information literacy: The ETS iSkills™ assessment and the NJIT Information Literacy Scale.*

In the fall of 2004, librarians and faculty at New Jersey Institute of Technology (NJIT) began a formal investigation of the information literacy skills of undergraduate students. Working with specialists in research and information literacy at the university's Robert Van Houten Library, instructors in the department of humanities worked to design an information literacy model based on standards derived from the Association of College & Research Libraries (ACRL). In that the faculty had been assessing the writing skills of students enrolled in general undergraduate requirements (GUR) in humanities since 1996, a traditional portfolio assessment system had emerged that allowed reliable and valid programmatic information to be gained about student writing (Elliot, Briller, & Joshi, 2007). A new portfolio assessment system launched in spring 2005—termed the NJIT Information Literacy Scale (ILS)—shifted the assessment focus from traditional writing to information literacy assessment (Scharf et al., 2007). While allowing similarly strong validity evidence to be warranted as the original portfolio system, the information literacy scores were lower than anticipated. Instructional and library faculty were interested in learning more about the information literacy skills of their students.

In fall 2005, NJIT and ETS undertook a collaborative research agreement to investigate more fully—by means of multiple approaches—the variables of information literacy as they were evidenced within student performance at a public comprehensive technological university. The collaboration would bring together the portfolio-based assessment approach of NJIT with the performance-based, automatically scored iSkills™ assessment, which was designed to measure information literacy skills as they appear in technological environments (Katz, 2005). The collaboration was designed to provide insight into the following questions:

- What kinds of validity evidence could be warranted based on the relationship of the two measures of information literacy to other variables? We hypothesized that the ETS iSkills assessment and the NJIT ILS posited associations that were congruent yet distinct. The relationships between these measures and general academic measures (course grade, grade point average [GPA], and scores on the College Board's SAT® Math [SAT-M] and SAT Verbal [SAT-V] assessments) refine our understanding of the discrimination between the cross-disciplinary concept of information literacy that underlies iSkills and NJIT's concept of information literacy in the humanities.
- Based on the consequences of the release of the scores and the consequences of the collaboration itself, what kinds of evidence could be warranted to describe the impact of the ETS and NJIT collaboration upon the NJIT community? While the impact of various kinds of portfolio scoring at NJIT had been demonstrated—both internally to the institution (e.g., New Jersey Institute of Technology, 2007, pp. 53–54) and externally to a national community (e.g., Coppola & Elliot, 2007)—the impact of an assessment system using a nationally developed assessment of information literacy and a locally developed assessment of that construct is unknown.

The American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999) provide a meaningful heuristic to the process of validation in *Standards for*

[1] New Jersey Institute of Technology
[2] Pratt Institute

Educational and Psychological Testing. By reflecting on the construct at hand, examining the relationship of the construct to other variables, and documenting the consequences of the assessment activity, we offer the following case study as a heuristic by which the concept of information literacy may be more fully understood.

Concepts and Measures

ETS iSkills™ Assessment

The ETS iSkills assessment targets the skillful use of information within technological environments. This scoping of information literacy was motivated by the information challenges posed by technology: Researching and communicating of information is often mediated by technology, and the wealth of information available via information and communication technology (ICT) challenges students' ability to locate relevant information efficiently, manage overwhelming information skillfully, and communicate effectively and ethically. The specification of this type of information literacy—termed ICT literacy—derives from the American Library Association (1989) definition of information literacy, the conclusions of an international panel formed to investigate literacy issues with regards to technology (International ICT Literacy Panel, 2002), and the results from designing the iSkills assessment in collaboration with representatives of seven U.S. college and university systems. The definition of ICT literacy adopted reflects a comprehensive view of information literacy that is not tied to any specific discipline:

> ICT literacy is the ability to appropriately use digital technology, communication tools, and/or networks to solve information problems in order to function in an information society. This includes having the ability to use information as a tool to research, organize, and communicate information and having a fundamental understanding of the ethical/legal issues surrounding accessing and using information.
> (Katz, 2005, p. 45)

The iSkills assessment embodies this form of information literacy as an Internet-delivered, automatically scored, performance-based assessment. Assessment administration takes approximately 75 minutes, divided into two sections lasting 35 and 40 minutes, respectively. During this time, students respond to 15 interactive tasks, each comprising a real-world scenario, such as a class or work assignment, that frames the information task. Students solve the tasks in the context of a simulation (e.g., e-mail, Web browser, or library database) having the look and feel of typical applications. Katz (2007) provides further details on the assessment, including its development and field testing.

The NJIT Information Literacy Scale (ILS)

It is the social and cognitive aspects of written communication, reflecting critical thinking and problem-solving ability, that are under investigation in the student portfolios required by the NJIT Department of Humanities (Bazerman, 2008; Flower, 1994). Indeed, these portfolios serve as the vehicles that capture the information literacy skills of NJIT undergraduate students as that ability is exhibited through critical reflection and problem exposition within courses.

The NJIT definition of information literacy was and remains based on the definition of information literacy offered by the Middle States Commission on Higher Education (MSCHE; Middle States Commission on Higher Education, 2006):

> Within the Department of Humanities at NJIT, information literacy is the ability to demonstrate that a coherent, planned intellectual framework has been used to identify, find, understand, and use information in drafting, revising, and finalizing researched, persuasive writing (Scharf et al., 2006).

The NJIT effort is designed to allow all instructors teaching undergraduate courses to come together each semester and, within less than 3 hours, reliably evaluate the work of a representative number of students. During the assessment period, trained instructors evaluate four key characteristics (described below) of student portfolios, artifacts capturing the work completed in a 15-week semester. Within the portfolio are contained a variety of documents, depending on the cohort and instructor: annotated planning bibliographies, proposals for research projects, drafts of various writing tasks, evidence of collaborative work, and researched final documents. With the courses in technical writing, the documents may be contained in a student-designed Web site, the site itself designed according to audience-based usability principles. Common to all aspects of the undergraduate curriculum is the emphasis on persuasion.

Differences in the iSkills and the ILS Assessments

While the NJIT ILS was informed by a literature review similar to that which informed the ETS ICT literacy framework, the purpose of the NJIT assessment differs in two ways from the iSkills assessment. First, within the specific institutional site, humanities instructors were interested in the variables of information literacy as they were articulated within the undergraduate curriculum. As such, the NJIT ILS focused on written products—the researched, persuasive documents

contained in portfolios. Unlike the iSkills assessment, the NJIT assessment does not therefore account for the process by which students completed their classroom writing assignments. Second, because students take humanities courses from first through senior years, instructors realized that they were capable of investigating differences within grade levels—but not across grade levels—of the offered curriculum. The NJIT assessment focused on the context in which the portfolios emerged; hence the portfolio scores were not designed to follow grade level but, instead, were designed to reveal performance of students in the humanities classes in which they were enrolled.

The present case study of NJIT undergraduate students allowed investigation of the way that students defined, accessed, evaluated, managed, integrated, created, and communicated information in the broad context of information literacy, a task that is often (though not always) executed within a set time frame reflective of the 75-minute assessment. As well, the case study allowed investigation into the effectiveness with which students cited sources, launched independent research, employed appropriate sources, and integrated their ideas with the ideas of others, a task that is often (though not always) executed within a set time frame reflective of the duration of a semester course. Overall, we expected this case study to help identify valuable discriminant evidence on the different aspects of information literacy captured by the two measures.

Method

Participants and Procedure

A simple random sample of upper-division students was created across each section of two representative humanities writing courses: cultural history and technical writing. These students, along with students from the senior seminar who were selected as described below, were identified for portfolio submission. The senior seminar students consisted of a census of all whose transcripts revealed that they had never taken any course outside of NJIT; hence, these students, while small in number, represented a meaningful population of NJIT students. Overall, students were found in cultural history (n = 95), as well as in technical writing (n = 48) and the senior seminars (n = 33). The sample resulting from the sampling plan closely matched the NJIT student population. Students were tested in a proctored computer lab on the iSkills assessment in late March 2006, and in May 2006, portfolios of targeted students were evaluated according to the NJIT ILS.

Two readers independently read and evaluated each portfolio using the NJT ILS scoring rubric. Analyses of weighted kappa suggested that the independent ratings are in moderate to substantial agreement (0.48-0.73) for the cultural history and senior seminar portfolios, and in fair agreement (0.29-0.39) for the technical writing portfolios.

Analyses

The goal of the analysis was to investigate the similarities and differences in measurement provided by the iSkills assessment and NJIT portfolio rubric. The variables included in the analyses are:

- *iSkills scores.* Scores on the iSkills assessment range from 400 to 700. Reliability (Cronbach alpha) is approximately .80.
- *Component ILS.* The ILS includes four component scales (citation, application, evidence, and integration), which are each rated from 1 to 6 (*very strongly disagree* to *very strongly agree* with the rubric's analytic statements). The score for each component is the sum of the ratings, ranging from 2 to 12, from two judges. To simplify discussion, analyses used the mean of the four component scores for each student. Cronbach alpha for the four-item scale is .87. Note that the technical writing students were scored on a modification of the ILS that included only the citation and application scales. For this group, Cronbach alpha for the two-item scale is .77.
- *Course grade.* This variable is a numerical translation of the letter grade each student received in the humanities class from which he or she was recruited. The variable ranges from 0 (F) to 4 (A). Course withdrawals and incompletes were interpreted as missing data.
- *GPA*. This variable is the undergraduate GPA of each student, including all courses up to the semester in which the study occurred. As with course grade, values range from 0 to 4.
- *SAT-M and SAT-V.* These SAT-M and SAT-V scores were obtained from NJIT student records. Each score ranges from 200 to 800. Cronbach alpha for SAT-M and SAT-V have been reported as .92 and.93, respectively (Ewing, Huff, Andrews, & King, 2005).

Results

Table 1 provides the means and standard deviations of the information literacy variables for the three cohorts. The cultural history students' iSkills scores were comparable with those of students enrolled in technical writing. In that both cultural history and technical writing have the same prerequisite first-year writing course, the nearly identical iSkills scores of both

***Table 1*: Means (Standard Deviations)**

	Cultural history (*n* = 95)	Technical writing (*n* = 48)	Senior seminar (*n* = 33)	*F*(2,143)	*p*	Partial η^2
iSkills scores	548.5_a (36.9)	547.2_a (39.5)	568.3_b (28.2)	3.9	<.05	0.05
Component ILS	7.4_a (2.1)	6.2_b (2.0)	7.4_a (1.7)	4.90	< .01	0.06
Course grade	3.2 (0.8)	3.4 (0.8)	3.5 (0.5)	1.90	ns	
GPA	2.9_a (0.5)	2.9_a (0.6)	3.1_a (0.4)	3.30	< .05	0.04
SAT-M	585.4 (73.6)	571.5 (88.1)	610.3 (67.5)	2.20	ns	
SAT-V	511.5 (81.8)	513.0 (98.7)	528.4 (40.1)	0.55	ns	

Note: Different subscripts within a row represent means different at the 0.05 level by Tukey's Honestly Significant Difference test. Partial η^2 is an effect size measure representing the proportion of total variance attributed to the effect.

groups suggest that additional information literacy instruction may not be forthcoming from other coursework outside of humanities courses. As expected, the students in the senior seminar demonstrated significantly higher iSkills scores than students enrolled in cultural history and technical writing.

Distressing to the NJIT instructional staff was the absence of evidence that students in technical writing had gained proficiency in the areas of citation and evidence of independent research, the two variables examined in this cohort of student portfolios. The component ILS score for these two combined variables is significantly lower than the scores for either the cultural history or the senior seminar students. The technical writing students' component ILS (composed of only citation and evidence of independent research; M = 6.2, SD = 2.0) is also lower than the mean of the two corresponding scores for cultural history (M = 7.8, SD = 2.1) and senior seminar students (M = 7.7, SD = 1.6). As a score of 7 is considered the lowest acceptable by NJIT instructional faculty, the performance of technical writing students does not meet expectations.

Less distressing, however, were the scores of the senior seminar students. Scharf et al. (2007), investigating a similar group of students at NJIT (*n* = 100), found ILS scores below the cut score of 7 on each variable that makes up the component ILS score used in the current study. One year later, in the current study, the ILS scores (M = 7.4, SD = 1.7) of a similar cohort were higher, although just barely meeting expectations.

The correlation analyses (Table 2) show moderate relationships among most measures. Both the iSkills and ILS portfolio scores are correlated with course grades and GPA, the latter at a similar level to other research on the iSkills assessment (Katz & Smith-Macklin, 2007). Scores on iSkills correlate well with SAT scores, in particular with SAT-V scores, as befits a measure of information handling skills. Finally, moderate correlations exist between iSkills and ILS scores (r = .21). However, these correlations may be mediated by students' general academic skills as measured by SAT scores. Partial correlations controlling for SAT-M and SAT-V are lower (ρ = .15). The ETS and NJIT assessments may be more distinct than related.

Discussion

The results reflect two appropriately different definitions and measures of information literacy. The process of critical reflections is the vehicle by which the concept of information literacy is operationalized within the NJIT humanities framework. The content domain of information literacy, intermixed with highly demanding reading and persuasive writing tasks, is executed in a different time frame and with an approach distinct from that used by the ETS iSkills assessment. The tasks and the constructs both assessments embody may be related, yet they are nevertheless distinct. Of course, information literacy as mediated by a humanities-oriented framework for writing may itself be distinct from the goal-directed writing of specific disciplines. Indeed, recent work suggests a relationship between iSkills assessment scores and grades in a business writing course (Katz, Haras, & Blaszczynski, 2008). As recent theory suggests, the validation of information literacy is a process in which validity arguments emerge and are warranted over time (Brennan, 2006; Mislevy, 2007).

Without question, NJIT nevertheless values the constructs reflected in both the ETS and NJIT measures. The humanities tasks assessed by the ILS fit within the mission of that academic unit, just as the iSkills tasks apply across the curriculum. Students need to write and reflect during extended periods, just as they need to evaluate rapidly much of the information they encounter daily. Indeed, as shown in the full report of this work, the correlations between the iSkills and ILS scores are higher in

***Table 2*: Intercorrelations**

	iSkills	Component ILS	Course grade	GPA	SAT-M	SAT-V
iSkills	-	0.21**	0.21**	0.27**	0.38**	0.49**
Component ILS		-	0.37**	0.25**	0.08	0.17*
Course grade			-	0.54**	0.20*	0.37**
GPA				-	0.32**	0.41**
SAT-M					-	0.52**
SAT-V						-

*p < .05. **p < .01.

sophomore, junior, and senior students, as compared with first-year students (Katz et al., 2008). Perhaps students gain both sets of cognitive complex abilities—those demanded by the iSkills and ILS tasks—as they progress through the curriculum. If so, then that congruence of domain and task would be an ideal academic outcome: an integration of discriminate skills, related yet distinct, required for all graduates.

Consequences of the Collaboration for NJIT

Evaluation of the consequences of assessment must be part of all program evaluation (Kane, 2006; Messick, 1994). Instead of considering consequences a factor apart from the investigation of construct and concurrent relationships, the consequences involved with the assessment of information literacy should be warranted as equally important to the success of the assessment. Along with the gains realized through a highly articulated model of information literacy and empirical assessment of student ability, NJIT has realized a more fully articulated sense of information literacy as administrators and instructors have begun, in committee and classroom, to address the information literacy skills of students. Even the rater agreement measures noted in this case study may be understood as evidence of the capability of faculty and librarians to unite in pursuit of a common assessment goal involving a new, yet critical, literacy that is as important to student success as academic writing ability—higher education's so-called composition emphasis—was to students at the turn of the 20th century.

In preparing *The Future's Edge* (New Jersey Institute of Technology, 2007), a periodic accreditation review report prepared for the MSCHE, the office of the president featured both the traditional writing assessment of portfolios conducted by the department of humanities (pp. 46–47) and the new collaborative information literacy assessment with ETS (pp. 53–54). Demonstration of efforts to assess information literacy within the undergraduate population was clearly important to the NJIT administration, and research with ETS allowed NJIT to follow Category 8 *Characteristics of Programs of Information Literacy That Illustrate Best Practices: A Guideline* in its suggestion that multiple methods for program evaluation are needed for effective outcomes assessment (Association of College & Research Libraries, 2003). Furthermore, librarians at NJIT use these results to promote their university-wide efforts on information literacy education. Informed by the results of the present study, the librarians would be the first to counsel that students who are selected to take the iSkills assessment must be identified from various departments within the academic institution; information literacy must not be seen to reside solely within the domain of the humanities. At present, the information literacy initiative is central to the department of humanities' newly created second-semester first-year composition course, which focuses on researched writing. As well, a planned university-wide initiative gives special attention to assessment and accountability. Viewed as an emerging curricular construct, information literacy efforts continue across the university curriculum.

Conclusions

Perhaps best understood as related yet distinct measures, the NJIT ILS and the ETS iSkills assessment together provide a fuller construct representation of information literacy for the university than either measure separately. While it is clear from this case study that there is much room for improvement regarding the information literacy abilities of NJIT undergraduate students, the across-the-curriculum orientation gained by employing both measures has resulted in an instructional emphasis that is already demonstrating gains in awareness of the important and diverse nature of information

literacy skills within a technological research university. Indeed, in that information literacy may be understood as an emerging construct—a point made by Tyler (2005) in associating information literacy with emerging global competitiveness—it is heartening to see the content of both the NJIT and ETS assessments so readily confirmed by instructors and librarians over the past 3 years. Significantly, at NJIT the definition of literacy as both an individual and communal good is strengthened by the use of both assessment systems, a consequence that has obviated the value dualisms often associated with literacy (Brandt, 2004). The collaborative effort described in this report has provided a combined assessment for New Jersey's only comprehensive technological university. While future studies are planned, they will be possible only because of the unique collaborative research model—one that recognizes the harmony that can and should exist between discriminant measures of information literacy—described in this report.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Library Association. (1989). *Presidential committee on information literacy: Final report.* Chicago: American Library Association.

Association of College & Research Libraries. (2003). *Characteristics of programs of information literacy that illustrate best practices: A guideline*. Retrieved December 30, 2008, from http://www.ala.org/ala/mgrps/divs/acrl/standards/characteristics.cfm

Bazerman, C. (Ed.). (2008). *Handbook of research on writing: History, society, school, individual, text*. New York: Erlbaum.

Brandt, D. (2004). Drafting U.S. literacy. *College English, 66*, 485-502.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1-16). Westport, CT: Praeger.

Coppola, N., & Elliot, N. (2007). A technology transfer model for program assessment in technical communication. *Technical Communication*, *54*, 459-474.

Elliot, N., Briller, V., & Joshi, K. (2007). Portfolio assessment: Quantification and community. *Journal of Writing Assessment, 3*(1), 5-30.

Ewing, M., Huff, K., Andrews, M., & King, K. (2005). *Assessing the reliability of skills measured by the SAT* (College Board Research Notes No. RN-24). Retrieved December 30, 2008, from http://www.collegeboard.com/research/pdf/RN-24.pdf

Flower, L. (1994). *The construction of negotiated meaning: A social cognitive theory of writing*. Carbondale: Southern Illinois University Press.

International ICT Literacy Panel. (2002). Digital transformation: A framework for ICT literacy. Princeton, N.J.: Educational Testing Service. Retrieved January 12, 2009, from http://www.ets.org/Media/Tests/Information_and_Communication_Technology_Literacy/ictreport.pdf

Kane, M. T. (2006) Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Katz, I. R. (2005). Beyond technical competence: Literacy in information and communication technology. *Educational Technology Magazine*, *45*(6), 44-47.

Katz, I. R. (2007). Testing information literacy in digital environments: ETS's iSkills™ assessment. *Information Technology and Libraries, 26*(3), 3-12.

Katz, I. R., Elliot, N., Attali, Y., Scharf, D., Powers, D., Huey, H., Joshi, K., & Briller, V. (2008). *The assessment of information literacy: A case study* (ETS Research Report No. RR-08-03). Princeton, NJ: Educational Testing Service.

Katz, I. R., Haras, C. M., & Blaszczynski, C. (2008). *Instruction and assessment of business students' information literacy.* Manuscript in preparation.

Katz, I. R., & Smith-Macklin, A. (2007). Information and communication technology (ICT) literacy: Integration and assessment in higher education. *Journal of Systemics, Cybernetics, and Informatics, 5*(4), 50-55. Retrieved December 30, 2008, from http://www.iiisci.org/Journal/SCI/Abstract.asp?var=&id=P890541

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Middle States Commission on Higher Education. (2006). *Characteristics of excellence in higher education: Eligibility requirements and standards for higher education*. Philadelphia, PA: Author. Retrieved December 30, 2008, from http://msche.org/publications/CHX06_Aug08080728132708.pdf

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*, 463-469.

New Jersey Institute of Technology. (2007). *The future's edge: NJIT periodic review report prepared for the Middle States Commission on Higher Education*. Newark, NJ: Author. Retrieved December 30, 2008, from http://www.njit.edu/irp/reports/2007/Formatted_PRR_06-11-2007_FINAL.pdf

Scharf, D., Elliot, N., Katz, I., Attali, Y., Powers, D., Huey, H., et al. (2006, July). *Information literacy at NJIT: Toward validity*. Invited presentation to the ETS iSkills National Advisory Committee, Los Angeles, CA.

Scharf, D., Elliot, N., Huey, H., Briller, V., & Joshi, K. (2007). Direct assessment of information literacy using writing portfolios. *Journal of Academic Librarianship*, *33*, 462-477.

Tyler, L. (2005). *ICT literacy: Equipping students to succeed in an information-rich, technology-based society* (ETS Issue Paper). Princeton, NJ: Educational Testing Service. Retrieved December 30, 2008, from http://www.ets.org/Media/Tests/ICT_Literacy/pdf/ICT_Equipping_Students_to_Succeed.pdf

2008 Abstracts from the ETS Research Report Series

Editor's note: *The ETS Research Report Series provides limited dissemination of ETS research, usually prior to formal publication. In this issue of Research Spotlight, we include abstracts of all ETS Research Reports released in 2008. The reports span a range of topics within the educational measurement research field from the theoretical to the practical. Electronic versions of the reports, in PDF format, are free. Some URLs are included below; others may be requested, for individual use only, by writing to R&DWeb@ets.org. Include the report title and number in your message.*

Analytic Scoring of TOEFL® CBT Essays: Scores From Humans and E-rater®

Report Number: RR-08-01, TOEFL-RR-81

Author(s): Y.-W. Lee, C. Gentile, & R. Kantor

Abstract: The main purpose of the study was to investigate the distinctness and reliability of analytic (or multitrait) rating dimensions and their relationships to holistic scores and e-rater® essay feature variables in the context of the TOEFL® computer-based test (CBT) writing assessment. Data analyzed in the study were analytic and holistic essay scores provided by human raters and essay feature variable scores computed by e rater (version 2.0) for two TOEFL CBT writing prompts. It was found that (a) all of the six analytic scores were not only correlated among themselves but also correlated with the holistic scores, (b) high correlations obtained among holistic and analytic scores were largely attributable to the impact of essay length on both analytic and holistic scoring, (c) there may be some potential for profile scoring based on analytic scores, and (d) some strong associations were confirmed between several e rater variables and analytic ratings. Implications are discussed for improving the analytic scoring of essays, validating automated scores, and refining e-rater essay feature variables.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-01.pdf

Investigating the Criterion-Related Validity of the TOEFL® Speaking Scores for ITA Screening and Setting Standards for ITAs

Report Number: RR-08-02, TOEFLiBT-03

Author(s): X. Xi

Abstract: Although the primary use of the speaking section of the Test of English as a Foreign Language™ Internet-based test (TOEFL® iBT Speaking test) is to inform admissions decisions at English medium universities, it may also be useful as an initial screening measure for international teaching assistants (ITAs). This study provides criterion-related validity evidence for the use of TOEFL iBT Speaking for ITA screening and evaluates the effectiveness of using the scores for teaching assistantship (TA) assignment classification. Four universities participated in this study. Local ITA-screening tests or instructor recommendations were used as the criterion measures. Relationships between the TOEFL Speaking test and the local ITA tests were explored through observed and disattenuated correlations. These relationships were moderately strong, supporting the use of the TOEFL Speaking test for ITA screening. However, the strengths of the relationship between the TOEFL Speaking test and the local ITA tests were found to be somewhat different across universities depending on the extent to which the local test engaged and evaluated nonlanguage abilities. Implications of these findings are discussed. Binary and ordinal logistic regressions were used to investigate how effective TOEFL Speaking scores were in separating students into distinct TA assignment categories. At all four universities, TOEFL Speaking scores were significant predictors of students' TA assignments and were fairly accurate in classifying students for TA assignments. ROC curves were used to determine TOEFL Speaking cut scores for TA assignments at each university that would minimize false positives (i.e., true nonpasses classified as passes). The results have considerable potential value in providing guidance on using the TOEFL iBT Speaking scores for ITA screening.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-02.pdf

An Initial Field Trial of an Instrument for Measuring Learning Strategies of Middle School Students

Report Number: RR-08-03

Author(s): O. L. Liu, T. Jackson, & G. Ling

Abstract: Learning strategies have been increasingly recognized as a useful tool to promote effective learning. In response to the lack of available learning strategies measures for middle school students, this study designed an instrument for these students, assessing behavioral, cognitive, and metacognitive strategies. This instrument, the Middle School Learning Strategies (MSLS) scale, is examined in terms of factorial structure, reliability, and correlates. Three factors emerge from the analyses: effective strategies, help seeking, and bad habits. The subscales displayed a reasonable reliability, ranging from .70 to .87. Student grades in language arts, social studies, math, and science were collected as criterion variables. As expected, grades in these four subjects correlated positively with both effective strategies and help seeking, yet negatively with bad habits. As a pilot measure, this instrument has demonstrated promising features as a useful tool for students to evaluate and enhance their learning strategies.

To order a copy of this report, write to R&DWeb@ets.org.

Asymptotic Limits of Item Parameters in Joint Maximum-Likelihood Estimation for the Rasch Model

Report Number: RR-08-04

Author(s): S. J. Haberman.

Abstract: Techniques are developed for approximation and exact computation of the asymptotic limit of the item parameter estimates obtained by application of joint maximum-likelihood estimation to the Rasch model.

To order a copy of this report, write to R&DWeb@ets.org.

Continuous Exponential Families: An Equating Tool

Report Number: RR-08-05

Author(s): S. J. Haberman

Abstract: Continuous exponential families may be employed to find continuous distributions with the same initial moments as the discrete distributions encountered in typical applications of classical equating. These continuous distributions provide distribution functions and quantile functions that may be employed in equating. To illustrate, an application is considered for a randomly equivalent groups design.

To order a copy of this report, write to R&DWeb@ets.org. Specify the title and report number in your request.

Predicting Grades in Different Types of College Courses

Report Number: RR-08-06, CBR-2008-01

Author(s): B. Bridgeman, J. Pollack, & N. Burton

Abstract: The ability of high school grades (high school GPA) and SAT® scores to predict cumulative grades in different types of college courses was

evaluated in a sample of 26 colleges. Each college contributed data from three cohorts of entering freshmen, and each cohort was followed for at least four years. Colleges were separated into four levels by average SAT scores. Grade point averages for four categories of courses (English; science, math, and engineering [S/M/E]; social science; and education) were computed, and analyses were run separately for gender within race/ethnicity classifications. Correlations of the combined predictors with course grades over four or more years, corrected for range restriction, ranged from .45 for education courses to 0.64 for S/M/E courses. The SAT increment, that is, the increase in the multiple correlations when SAT scores are added to high school grades, ranged from 0.03 in education courses to 0.08 in S/M/E courses. Because these seemingly small numbers are frequently misinterpreted, an additional analysis showed how the percentage of students succeeding at a high level (cumulative GPA of 3.5 or higher) increases as SAT scores increase for students with similar high school grades. For example, for students with a high school GPA of 3.7 or higher in colleges where the mean combined SAT score is below 1200, only 2 percent of the students at the lowest SAT level (800 or lower combined score) were highly successful in social science courses. At the highest SAT level (1410–1600), 77 percent were highly successful.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-06.pdf

Approaches to the Design of Diagnostic Item Models

Report Number: RR-08-07

Author(s): E. A. Graf

Abstract: Quantitative item models are item structures that may be expressed in terms of mathematical variables and constraints. An item model may be developed as a computer program from which large numbers of items are automatically generated. Item models can be used to produce large numbers of items for use in traditional, large-scale assessments. But they have potential for use in other areas as well, including diagnostic assessment. In this report, I first review research on diagnostic assessment and then discuss how approaches to diagnostic assessment can inform the design of diagnostic item models.

To order a copy of this report, write to R&DWeb@ets.org.

Linking for the General Diagnostic Model

Report Number: RR-08-08

Author(s): X. Xu & M. von Davier

Abstract: Three strategies for linking two consecutive assessments are investigated and compared by analyzing reading data for the National Assessment of Educational Progress (NAEP) using the general diagnostic model. These strategies are compared in terms of marginal and joint expectations of skills, joint probabilities of skill patterns, and item parameter estimates. The results indicate that fixing item parameter values at their previously calibrated values is sufficient to establish a comparable scale for the subsequent year.

To order a copy of this report, write to R&DWeb@ets.org.

Factor Structure of the TOEFL® Internet-Based Test (iBT): Exploration in a Field Trial Sample

Report Number: RR-08-09, TOEFLiBT-04

Author(s): Y. Sawaki, L. Stricker, & A. Oranje

Abstract: The present study investigated the factor structure of a field trial sample of the Test of English as a Foreign Language™ Internet-based test (TOEFL® iBT). An item-level confirmatory factor analysis (CFA) was conducted for a polychoric correlation matrix of items on a test form completed by 2,720 participants in the 2003–2004 TOEFL iBT Field Study. CFA-based multitrait-multimethod (MTMM) analyses for the Reading and Listening sections showed that the language abilities assessed in each section were essentially unidimensional, while the factor structure of the entire test was best represented by a higher-order factor model with a general factor (English as a second language/English as a foreign language ability) and four group factors for reading, listening, speaking, and writing. The integrated Speaking and Writing tasks, which require language processing in multiple modalities, well defined the target modalities (speaking and writing). These results broadly support the current reporting of four scores corresponding to the modalities and a total score, as well as the test design where the integrated tasks contribute only to the scores for the target modalities.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-09.pdf

Impossible Scores Resulting in Zero Frequencies in the Anchor Test: Impact on Smoothing and Equating

Report Number: RR-08-10

Author(s): G. Puhan, A. A.von Davier, & S. Gupta

Abstract: Equating under the external anchor design is frequently conducted using scaled scores on the anchor test. However, scaled scores often lead to the unique problem of creating zero frequencies in the score distribution because there may not always be a one-to-one correspondence between raw and scaled scores. For example, raw scores of 17 and 18 may correspond to scaled scores of 150 and 153, thereby creating zero frequencies for scaled scores of 151 and 152. These gaps in the frequency distribution may adversely impact smoothing and equating. This study examines the effect of these zero frequencies on log-linear smoothing (Holland & Thayer, 1987) of score distributions and final equating results. Results suggest that although smoothing is significantly affected by the presence of these zero frequencies, as indicated by the likelihood-ratio chi-square, Akaike information criterion (Akaike, 1977), and Freeman-Tukey deviates, the impact on the actual equating results is minimal.

To order a copy of this report, write to R&DWeb@ets.org.

An Alternative Data Collection Design for Equating With Very Small Samples

Report Number: RR-08-11

Author(s): G. Puhan, T. Moses, M. Grant, & F. McHale

Abstract: A single group (SG) equating design with nearly equivalent test forms (SiGNET) design was developed by Grant (2006) to equate small volume tests. The basis of this design is that examinees take two largely overlapping test forms within a single administration. The scored items for the operational form are divided into mini-tests called testlets. An additional testlet is created but not scored for the first form. If the scored testlets are Testlets 1–6 and the unscored testlet is Testlet 7, then the first form is composed of Testlets 1–6, the second form is composed of Testlets 2–7, and Testlets 2–6 are common to both test forms. They are administered as a single administered form, and when a sufficient number of examinees have taken the administered form for an SG equating, the second form (Testlets 2–7) is equated to the first form (Testlets 1–6) using SG equating. As evident, there are at least two merits of the SiGNET design over the nonequivalent groups with anchor test (NEAT) design. First, it facilitates the use of an SG equating design, which has the least random equating error, and second, it allows for the accumulation of sufficient data to equate the second form. Since the examinees scores are based on only the first form (i.e., the operational form), the two forms can be administered until sufficient data are collected to equate the second form. This study compared equatings under the SiGNET and NEAT designs and found reduced bias and error for the SiGNET design in very small sample size situations (e.g., N = 10 or 15). Implications for practice using the SiGNET design are also discussed.

To order a copy of this report, write to R&DWeb@ets.org.

Comparing Alternative Kernels for the Kernel Method of Test Equating: Gaussian, Logistic, and Uniform Kernels

Report Number: RR-08-12

Author(s): Y.-H. Lee & A. A. von Davier

Abstract: The kernel equating method (von Davier, Holland, & Thayer, 2004) is based on a flexible family of equipercentile-like equating functions that use a Gaussian kernel to continuize the discrete score distributions. While the classical equipercentile, or percentile-rank, equating method carries out the

continuization step by linear interpolation, in principle the kernel equating methods could use various kernel smoothings to replace the discrete score distributions. This paper expands the work of von Davier et al. (2004) in investigating alternative kernels for equating practice. To examine the influence of different kernel functions on the equating results, this paper focuses on two types of kernel functions: the logistic kernel and the continuous uniform distribution (known to be the same as the linear interpolation). The Gaussian kernel is used for reference. By employing an equivalent-groups design, the results of the study indicate that the tail properties of kernel functions have great impact on the continuized score distributions. However, the equated scores based on different kernel functions do not vary much, except for extreme scores. The results presented in this paper not only support the previous findings on the efficiency and accuracy of the existing continuization methods, but also enrich the information on observed-score equating models.

To order a copy of this report, write to R&DWeb@ets.org.

Comparing Different Approaches of Bias Correction for Ability Estimation in IRT Models

Report Number: RR-08-13

Author(s): Y.-H. Lee & J. Zhang

Abstract: The method of maximum-likelihood is typically applied to item response theory (IRT) models when the ability parameter is estimated while conditioning on the true item parameters. In practice, the item parameters are unknown and need to be estimated first from a calibration sample. Lewis (1985) and Zhang and Lu (2007) proposed the expected response functions (ERFs) and the corrected weighted-likelihood estimator (CWLE), respectively, to take into account the uncertainty regarding item parameters for purposes of ability estimation. In this paper, we investigate the performance of ERFs and of the CWLE in different situations, such as various test lengths and levels of measurement error in item parameter estimation. Our empirical results indicate that ERFs can cause the bias in ability estimation to fall within [$-0.2,0.2$] for all conditions, whereas the CWLE can effectively reduce the bias in ability estimation provided that it has a good foundation to start from.

To order a copy of this report, write to R&DWeb@ets.org.

Examining an Alternative to Score Equating: A Randomly Equivalent Forms Approach

Report Number: RR-08-14

Author(s): C.-W. Liao & S. A. Livingston

Abstract: Randomly equivalent forms (REF) of tests in listening and reading for nonnative speakers of English were created by stratified random assignment of items to forms, stratifying on item content and predicted difficulty. The study included 50 replications of the procedure for each test. Each replication generated 2 REFs. The equivalence of those 2 forms was evaluated by comparing the raw-score distributions focusing on the greatest difference in the cumulative distributions. For listening, 10 replications produced cumulative distributions that differed at some point by more than 0.10, and 4 replications produced differences greater than 0.15. For reading, only 3 replications produced differences greater than 0.10. The difference between the results for listening and reading reflects the greater variation, within strata, in the difficulty of the listening items. The REF procedure may become more effective if item difficulty can be predicted more accurately.

To order a copy of this report, write to R&DWeb@ets.org.

One Approach to Detecting the Invariance of Proficiency Standards Over Time

Report Number: RR-08-15

Author(s): J. Qian

Abstract: This study explores the use of a mapping technique to test the invariance of proficiency standards over time for state performance tests. First, the state proficiency standards are mapped onto the National Assessment of Educational Progress (NAEP) scale. Then, rather than looking at whether there is a deviation in proficiency standards directly, the invariance of their NAEP equivalents is tested over time. The basis of the mapping technique is an enhanced method that was originally designed for comparing performance standards for public school students set by different states when the state tests are comparable. This approach can also be used to detect score inflation over time for state tests.

To order a copy of this report, write to R&DWeb@ets.org.

Effect of Immediate Feedback and Revision on Psychometric Properties of Open-Ended Sentence-Completion Items

Report Number: RR-08-16, GREB-03-15

Author(s): Y. Attali, D. Powers, & J. Hawthorn

Abstract: Registered examinees for the GRE® General Test answered open-ended sentence-completion items. For half of the items, participants received immediate feedback on the correctness of their answers and up to two opportunities to revise their answers. A significant feedback-and-revision effect was found. Participants were able to correct many of their initial incorrect answers, resulting in higher revised scores. In addition, the reliability of the revised scores and their correlation with GRE verbal scores were higher. The possibility of using revision scores as a basis for measuring potential future learning is discussed.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-16.pdf

Model-Based Weighting and Comparisons

Report Number: RR-08-17

Author(s): J. Qian

Abstract: In survey research, sometimes the formation of groupings, or aggregations of cases on which to make an inference, are of importance. Of particular interest are the situations where the cases aggregated carry useful information that has been transferred from a sample employed in a previous study. For example, a school to be included in the sample of the High School Effectiveness (HSES) study must contain one or more cases transferred from the National Educational Longitudinal Study of 1988 (NELS:88). To calculate the aggregation inclusion probabilities, this study investigated three statistical models and, based on these models, derived the school weights for the HSES study. This study also assessed the effects of weighting by comparing the statistics yielded from different sets of weights: (a) those from an empirical population database and (b) those from data generated from simulation based on the principles of a superpopulation. Both categorical data and continuous variables were analyzed in the comparison.

To order a copy of this report, write to R&DWeb@ets.org.

Evaluating the Effectiveness of a Full-Population Estimation Method

Report Number: RR-08-18

Author(s): H. Braun, J. Zhang, & S. Vezzu

Abstract: At present, although the percentages of students with disabilities (SDs) and/or students who are English language learners (ELL) excluded from a NAEP administration are reported, no statistical adjustment is made for these excluded students in the calculation of NAEP results. However, the exclusion rates for both SD and ELL students vary substantially across jurisdictions at a given administration, and, in some cases, have changed substantially over time within a jurisdiction. Consequently, comparisons of performance based on reported NAEP scores may indeed be biased by differential exclusion and identification practices. Using only NAEP data, this report investigates plausible explanations for the observed heterogeneity among jurisdictions in exclusion rates. It also examines the operating characteristics of a particular class of methods that carry out statistical adjustments to NAEP's reported scores to address the possible bias due to differential exclusion rates. The final results of such adjustments are termed full-population estimates (FPEs). The conclusions are that there is both a strong likelihood of bias and that neither the current NAEP procedure nor the FPE methodologies constitutes an ideal solution. The

former because it assumes that all excluded students could not meaningfully participate in NAEP, and the latter because they implicitly assume that all students could obtain a proper NAEP score.

To order a copy of this report, write to R&DWeb@ets.org.

A Developmental Writing Scale

Report Number: RR-08-19

Author(s): Y. Attali & D. Powers

Abstract: This report describes the development of grade norms for timed-writing performance in two modes of writing: persuasive and descriptive. These norms are based on objective and automatically computed measures of writing quality in grammar, usage, mechanics, style, vocabulary, organization, and development. These measures are also used in the automated essay scoring system e-rater® V.2. Norms were developed through a large-scale data collection effort that involved a national sample of 170 schools, more than 500 classes from 4th, 6th, 8th, 10th, and 12th grades and more than 12,000 students. Personal and school background information was also collected. These students wrote (in 30-minute sessions) up to 4 essays (2 in each mode of writing) on topics selected from a pool of 20 topics. The data allowed us to explore a range of questions about the development and nature of writing proficiency. Specifically, this paper provides a description of the trajectory of development in writing performance from 4th grade to 12th grade. The validity of a single developmental writing scale is examined through a human scoring experiment and a longitudinal study. The validity of the single scale is further explored through a factor analysis (exploratory and confirmatory) of the internal structure of writing performance and changes in this structure from 4th grade to 12th grade. The paper also explores important factors affecting performance, including prompt difficulty, writing mode, and student background (gender, ethnicity, and English language background).

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-19.pdf

Automated Scoring of Short-Answer Open-Ended GRE® Subject Test Items

Report Number: RR-08-20, GREB-04-02

Author(s): Y. Attali, D. Powers, M. Freedman, M. Harrison, & S. Obetz

Abstract: This report describes the development, administration, and scoring of open-ended variants of GRE® Subject Test items in biology and psychology. These questions were administered in a Web-based experiment to registered examinees of the respective Subject Tests. The questions required a short answer of 1-3 sentences, and responses were automatically scored by natural language processing methods, using the c-rater™ scoring engine, immediately after participants submitted their responses. Participants received immediate feedback on the correctness of their answers, and an opportunity to revise their answers. Subsequent human scoring of the responses allowed an evaluation of the quality of automated scoring. This report focuses on the success of the automated scoring process. A separate report describes the feedback and revision results.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-20.pdf

Effect of Immediate Feedback and Revision on Psychometric Properties of Open-Ended GRE® Subject Test Items

Report Number: RR-08-21, GREB-04-05

Author(s): Y. Attali & D. Powers

Abstract: Registered examinees for the GRE® Subject Tests in Biology and Psychology participated in a Web-based experiment where they answered open-ended questions that required a short answer of 1-3 sentences. Responses were automatically scored by natural language processing methods (the c-rater™ scoring engine) immediately after participants submitted their responses. Based on natural language processing methods (the c-rater scoring engine), participants received immediate feedback on the correctness of their answers and an opportunity to revise their answers. A significant revision effect was found. Participants were able to correct many of their initial incorrect answers, resulting in higher revised scores. In addition, the reliability of revised scores was higher than initial scores, although the correlations of the initial and revised scores with the GRE Subject Test scores were similar.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-21.pdf

Robustness of a Value-Added Assessment of School Effectiveness

Report Number: RR-08-22

Author(s): H. Braun, Y. Qu, & C. Trapani

Abstract: This paper reports on a study conducted to investigate the consistency of the results between 2 approaches to estimating school effectiveness through value-added modeling. Estimates of school effects from the layered model employing item response theory (IRT) scaled data are compared to estimates derived from a discrete growth model based on the analysis of transitions along an ordinal developmental scale. The data were extracted from the longitudinal records maintained in the Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS-K) archive for students remaining in the same school from the beginning of kindergarten through the end of Grade 3. The results of different comparisons indicated that the estimates from the 2 approaches are moderately consistent.

To order a copy of this report, write to R&DWeb@ets.org.

Examining the Impact of Audio Presentation on Tests of Reading Comprehension

Report Number: RR-08-23

Author(s): C. Cahalan Laitusis, L. Cook, F. Cline, T. King, & J. Sabatini

Abstract: This study examined the impact of a read-aloud accommodation on standardized test scores of reading comprehension at Grades 4 and 8. Under a repeated measures design, students with and without reading-based learning disabilities took both a standard administration and a read-aloud administration of a reading comprehension test. Results show that the mean score on the audio version was higher than scores on the standard version for both groups of students at both grade levels. Students with reading-based learning disabilities at both levels benefited differentially more than students with no disability. This finding continues to hold after controlling for reading fluency and ceiling effects at both grades. The results also examined the relationship between test scores and teachers' ratings of reading comprehension to determine which measures are the best predictors of teachers' ratings of reading comprehension by grade and disability classification.

To order a copy of this report, write to R&DWeb@ets.org.

Theoretical and Empirical Standard Errors for Two Population Invariance Measures in the Linear Equating Case

Report Number: RR-08-24

Author(s): A. A von Davier, J. R. Manalo, & F. Rijmen

Abstract: The standard errors of the 2 most widely used population-invariance measures of equating functions, root mean square difference (RMSD) and root expected mean square difference (REMSD), are not derived for common equating methods such as linear equating. Consequently, it is unknown how much noise is contained in these estimates. This paper describes 2 methods for obtaining the standard errors for RMSD and REMSD. The delta method relies on an analytical approximation and provides asymptotic standard errors. The grouped jackknife method is a sampling-based method. Both methods were applied to a real data application. The results showed that there was very little difference between the standard errors found by the 2 methods.

To order a copy of this report, write to R&DWeb@ets.org.

The Influence of Strategies for Selecting Loglinear Smoothing Models on Equating Functions

Report Number: RR-08-25

Author(s): T. Moses & P. W. Holland

Abstract: This study addressed 2 issues of using loglinear models for smoothing univariate test score distributions and for enhancing the stability of equipercentile equating functions. One issue was a comparative assessment of several statistical strategies that have been proposed for selecting 1 from several competing model parameterizations. Another issue was an evaluation of the influence of the selection strategies on equating function accuracy. These issues were considered in a simulation study, where the accuracies of 17 selection strategies for loglinear models and their effects on equating function accuracies were assessed across a range of sample sizes, test score distributions, and population equating functions. The results differentiate the selection strategies in terms of their accuracies in selecting correct model parameterizations and define the situations where their use has the most important implications for equating function accuracy.

To order a copy of this report, write to R&DWeb@ets.org.

Equating of Mixed-Format Tests in Large Scale Assessments

Report Number: RR-08-26

Author(s): S. Kim, M. E. Walker, & F. McHale

Abstract: This study examined variations of the nonequivalent-groups equating design for mixed-format tests—tests containing both multiple-choice (MC) and constructed-response (CR) items—to determine which design was most effective in producing equivalent scores across the two tests to be equated. Four linking designs were examined: (a) an anchor with only MC items; (b) a mixed-format anchor containing both MC and CR items; (c) a mixed-format anchor incorporating CR item rescoring; and (d) a hybrid combining single-group and equivalent-groups designs, thereby avoiding the need for an anchor test. Designs using MC items alone or those using a mixed anchor without CR item rescoring resulted in much larger bias than the other two design approaches. The hybrid design yielded the smallest root mean squared error value.

To order a copy of this report, write to R&DWeb@ets.org.

Fitting the Structured General Diagnostic Model to NAEP Data

Report Number: RR-08-27

Author(s): X. Xu & M. von Davier

Abstract: Xu and von Davier (2006) demonstrated the feasibility of using the general diagnostic model (GDM) to analyze National Assessment of Educational Progress (NAEP) proficiency data. Their work showed that the GDM analysis not only led to conclusions for gender and race groups similar to those published in the NAEP Report Card, but also allowed flexibility in estimating multidimensional skills simultaneously. However, Xu and von Davier noticed that estimating the latent skill distributions will be much more challenging with this model when there is a large number of subgroups to estimate. To make the GDM more applicable to NAEP data analysis, which requires a fairly large subgroups analysis, this study developed a log-linear model to reduce the number of parameters in the latent skill distribution without sacrificing the accuracy of inferences. This paper describes such a model and applies the model in the analysis of NAEP reading assessments for 2003 and 2005. The comparisons between using this model and the unstructured model were made through the use of various results, such as the differences between item parameter estimates and the differences between estimated latent class distributions. The results in general show that using the log-linear model is efficient.

To order a copy of this report, write to R&DWeb@ets.org.

Tight But Loose: Scaling Up Teacher Professional Development in Diverse Contexts

Report Number: RR-08-29

Editor: E. C. Wylie

Abstract: This series of papers was originally presented as a symposium at the annual meetings of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) held between April 9, 2007, and April 13, 2007, in Chicago, IL. The authors represent school districts and departments of education across the United States, as well as researchers at Cleveland State University, Educational Testing Service (ETS), the Institute for Education in London, and the University of Wyoming at Laramie. All of the current ETS staff, along with Dylan Wiliam and Marnie Thompson, worked at ETS for several years on an iterative research and development program, out of which grew the Keeping Learning on Track® (KLT) program. These papers represents the thinking about the theory behind the KLT program, describes the range of contexts used to implement the program, and illustrates the inherent tensions between the desire to maintain fidelity to a theory of action and the need to demonstrate flexibility in order to accommodate local situations. Papers 2 through 6 present descriptions of five implementations in chronological order.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-29.pdf

Response to Assessment Feedback: The Effects of Grades, Praise, and Source of Information

Report Number: RR-08-30

Author(s): A. A. Lipnevich & J. K. Smith

Abstract: This experiment involved college students (N = 464) working on an authentic learning task (writing an essay) under 3 conditions: no feedback, detailed feedback (perceived by participants to be provided by the course instructor), and detailed feedback (perceived by participants to be computer generated). Additionally, conditions were crossed with 2 factors of grade (receiving grade or not) and praise (receiving praise or not). Detailed feedback specific to individual work was found to be strongly related to student improvement in essay scores, with the influence of grades and praise more complex. Overall, detailed, descriptive feedback was found to be most effective when given alone, unaccompanied by grades or praise. The results have implications for theory and practice of assessment.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-30.pdf

Effects of Calculator Availability on GRE® Quantitative Questions

Report Number: RR-08-31, GREB-03-09

Author(s): B. Bridgeman, F. Cline, & J. Levin

Abstract: In order to estimate the likely effects on item difficulty when a calculator becomes available on the quantitative section of the Graduate Record Examinations® (GRE®-Q), 168 items (in six 28-item forms) were administered either with or without access to an on-screen four-function calculator. The forms were administered as a special research section at the end of operational tests, with student volunteers randomly assigned to the calculator or no-calculator groups. Usable data were obtained from 13,159 participants. Test development specialists were asked to rate which items they thought would become easier with a calculator. In general, the specialists were successful in identifying the items with relatively large calculator effects, though even these effects were quite small. An increase of only about four points in the percent correct should suffice for the items identified as likely to show calculator effects with no adjustment needed for the majority of the items. Introduction of a calculator should have little or no effect on gender and ethnic differences.

Full report available from
http://www.ets.org/Media/Research/pdf/RR-08-31.pdf

Sample-Size Requirements for Automated Essay Scoring

Report Number: RR-08-32

Author(s): S. J. Haberman & S. Sinharay

Abstract: Sample-size requirements were considered for automated essay scoring in cases in which the automated essay score estimates the score provided by a human rater. Analysis considered both cases in which an essay prompt is examined in isolation and those in which a family of essay prompts is studied. In typical cases in which content analysis is not employed and in which the only object is to score individual essays to provide feedback to the examinee, it appears that several hundred essays are sufficient. For application of one model to a family of essays, fewer than 100 essays per prompt may often be adequate. The cumulative logit model was explored as a possible replacement of the linear regression model usually employed in automated essay scoring; the cumulative logit model performed somewhat better than did the linear regression model.

To order a copy of this report, write to R&DWeb@ets.org.

The Assessment of Information Literacy: A Case Study

Report Number: RR-08-33

Author(s): I. R. Katz, N. Elliot, Y. Attali, D. Scharf, D. E. Powers, H. Huey, K. Joshi, & V. Briller

Abstract: This study presents an investigation of information literacy as defined by the ETS iSkills™ assessment and by the New Jersey Institute of Technology (NJIT) Information Literacy Scale (ILS). As two related but distinct measures, both iSkills and the ILS were used with undergraduate students at NJIT during the spring 2006 semester. Undergraduate students (n = 331), first through senior years, took the iSkills and submitted portfolios to be judged by the ILS. First-year students took the Core iSkills assessment, which was designed to provide administrators and faculty with an understanding of the information and communication technology (ICT) literacy of a student doing entry-level coursework (n = 155). Upper classmen took the more difficult Advanced iSkills assessment, appropriate for rising juniors (n = 176). Across all class levels, iSkills scores varied as expected. First-year basic skills writing students performed at lower levels than first-year students enrolled in traditional composition and cultural history courses; seniors performed at higher levels than sophomores and juniors. Because the NJIT ILS scores were designed to be curriculum sensitive, portfolio scores did not similarly follow grade levels. Analyses revealed weak correlations between portfolio and Core iSkills scores and moderate correlations between portfolio and Advanced iSkills scores. As two associated yet distinct systems of inquiry designed to explore undergraduate student performance, the ETS iSkills assessment and the NJIT ILS—taken both individually and together—yield important information regarding student performance.

To order a copy of this report, write to R&DWeb@ets.org.

Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology

Report Number: RR-08-34, TOEFLiBT-05

Author(s): R. J. Tannenbaum & E. C. Wylie

Abstract: The Common European Framework of Reference (CEFR) describes language proficiency in reading, writing, speaking, and listening on a 6-level scale. In this study, English-language experts from across Europe linked CEFR levels to scores on three tests: the TOEFL® iBT test, the TOEIC® assessment, and the TOEIC Bridge™ test. Standard-setting methodology (a modified Angoff approach and a modified examinee paper selection approach) was used to construct the linkages. Linkages were established for TOEFL iBT at levels B1, B2, and C1. Linkages were established for TOEIC at levels A1 through C1, with the exception of Reading at the C1 level. The TOEIC Bridge test was linked to its three targeted levels of the CEFR. The report details the methods, procedures, and results of the study.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-34.pdf

Comparing Multiple-Group Multinomial Loglinear Models for Multidimensional Skill Distributions in the General Diagnostic Model

Report Number: RR-08-35

Author(s): X. Xu & M. von Davier

Abstract: The general diagnostic model (GDM) utilizes located latent classes for modeling a multidimensional proficiency variable. In this paper, the GDM is extended by employing a log-linear model for multiple populations that assumes constraints on parameters across multiple groups. This constrained model is compared to log-linear models that assume separate sets of parameters to fit the distribution of latent variables in each group of a multiple-group model. Estimation of these constrained log-linear models using iterative weighted least squares (IWLS) methods is outlined and an application to NAEP data exemplifies the differences between constrained and unconstrained models in the presence of larger numbers of group-specific proficiency distributions. The use of log-linear models for the latent skill space distributions using constraints across populations allows for efficient computations in models that include many proficiency distributions.

To order a copy of this report, write to R&DWeb@ets.org.

Development of Approximations to Population Invariance Indices

Report Number: RR-08-36

Author(s): J. Liu & X. Zhu

Abstract: The purpose of this paper is to explore methods to approximate population invariance without conducting multiple linkings for subpopulations. Under the single group or equivalent groups design, no linking needs to be performed for the parallel-linear system linking functions. The unequated raw score information can be used as an approximation. For other linking functions that are nonparallel-linear, linking only needs to be conducted for the total population. The difference of the standardized mean differences between each subpopulation and the total population across the old form and the new form can be used as an approximation of population invariance. Under the nonequivalent groups with anchor test design, conducting separate subpopulation linking and comparing them to the total population linking may still be the best way to estimate population invariance.

To order a copy of this report, write to R&DWeb@ets.org. Specify the title and report number in your request.

The Impact of Changes in the TOEFL® Examination on Teaching and Learning in Central and Eastern Europe: Phase 2, Coping With Change

Report Number: RR-08-37, TOEFLiBT-05

Author(s): D. Wall & T. Horák

Abstract: The aim of this report is to present the findings of the second phase in a longitudinal study of the impact of changes in the TOEFL® test on teaching and learning in test preparation classrooms. The focus of this phase was to monitor six teachers from five countries in Central and Eastern Europe as they received news about changes in the TOEFL and began thinking about how these might affect their teaching in the future. Data were gathered during the period of January to May 2005. The teachers responded to monthly tracking questions and tasks that explored their awareness of the old and new TOEFL tests, the features of their test preparation classes, their reactions to the most innovative parts of the new test, and their thoughts about the type of content and activities they would offer once the new TOEFL was operational in their countries. The report includes an analysis of the teachers' awareness, attitudes, and plans, and a discussion of the types of factors that could affect the shape and intensity of TOEFL washback in years to come.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-37.pdf

Evidence-Centered Assessment Design for Reasoning About Accommodations for Individuals With Disabilities in NAEP Reading and Math

Report Number: RR-08-38

Author(s): E. G. Hansen, R. J. Mislevy, & L. S. Steinberg

Abstract: Accommodations play a key role in enabling individuals with disabilities to participate in the National Assessment of Educational Progress (NAEP) and other large-scale assessments. However, it can be difficult to know how accommodations affect the validity of results, thus making it difficult to determine which accommodations should be allowed. This study describes recent extension of evidence-centered assessment design (ECD) for reasoning about the impact of accommodations and other accessibility features (e.g., universal design features) on the validity of assessment results, using examples from NAEP reading and mathematics. The study found that the ECD-based techniques were useful in analyzing the effects of accommodations and other accessibility features on validity. Such design capabilities may increase assessment designers' capacity to employ accessibility features without undermining validity.

To order a copy of this report, write to R&DWeb@ets.org.

Small-Sample Equating by the Circle-Arc Method

Report Number: RR-08-39

Author(s): S. A. Livingston & S. Kim

Abstract: This paper suggests two new, related methods for estimating a test-score equating relationship from small samples of test takers. These methods do not require the estimated equating transformation to be linear. Instead, they constrain the estimated equating curve to pass through 2 prespecified endpoints and a middle point determined from the data. Some preliminary results indicate that these methods outperform mean equating and other methods used for equating in small samples.

To order a copy of this report, write to R&DWeb@ets.org.

L-Bivariate and *L*-Multivariate Association Coefficients

Report Number: RR-08-40

Author(s): N. Kong & C. Lewis

Abstract: Given a system of multiple random variables, a new measure called the *L*-multivariate association coefficient is defined using (conditional) entropy. Unlike traditional correlation measures, the *L*-multivariate association coefficient measures the multiassociations or multirelations among the multiple variables in the given system; that is, the *L*-multivariate association coefficient measures the degree of the association for the given system. The *L*-multivariate association coefficient for the system of two random variables is also called the *L*-bivariate association coefficient. The association measured by the *L*-multivariate association coefficient is a general type of association, not any specific type of a linear or nonlinear association. Unlike the *K*-dependence coefficient, which is an asymmetrical measure, the *L*-multivariate association coefficient is a symmetrical measure. A direct application of the *L*-multivariate association coefficient is in variables selection or variables reduction. This paper also explores the relationship between the *L*-multivariate association coefficient and the *K*-dependence coefficient.

To order a copy of this report, write to R&DWeb@ets.org.

Outliers in Assessment

Report Number: RR-08-41

Author(s): S. J. Haberman

Abstract: Outliers in assessments are often treated as a nuisance for data analysis; however, they can also assist in quality assurance. Their frequency can suggest problems with form codes, scanning accuracy, ability of examinees to enter responses as they intend, or exposure of items.

To order a copy of this report, write to R&DWeb@ets.org.

A Study of Confidence and Accuracy Using the Rasch Modeling Procedures

Report Number: RR-08-42

Author(s): I. Paek, J. Lee, L. Stankov, & M. Wilson

Abstract: This study investigated the relationship between students' actual performance (accuracy) and their subjective judgments of accuracy (confidence) on selected English language proficiency tests. The unidimensional and multidimensional IRT Rasch approaches were used to model the discrepancy between confidence and accuracy at the item and test level and to assess disattenuated strength of association between accuracy and confidence. The analysis results indicate a pattern of overconfidence bias (i.e., overestimation of success rate), which was related to item difficulty. In addition, the strength of association between accuracy and confidence dimension was relatively high: The confidence dimension explained 45% and 52% of the variability in the accuracy dimension for the two tests employed in this study.

To order a copy of this report, write to R&DWeb@ets.org.

A Review of Recent Developments in Differential Item Functioning

Report Number: RR-08-43

Author(s): R. Mapuranga, N. J. Dorans, & K. Middleton

Abstract: In many practical settings, essentially the same differential item functioning (DIF) procedures have been in use since the late 1980s. Since then, examinee populations have become more heterogeneous, and tests have included more polytomously scored items. This paper summarizes and classifies new DIF methods and procedures that have appeared since the early 1990s and assesses their appropriateness for practical use. Widely used DIF methods are evaluated alongside these new methods for completeness, clarity, and comparability.

To order a copy of this report, write to R&DWeb@ets.org.

DIF Detection With Small Samples: Applying Smoothing Techniques to Frequency Distributions in the Mantel-Haenszel Procedure

Report Number: RR-08-44

Author(s): L. Yu, T. Moses, G. Puhan, & N. J. Dorans, Neil

Abstract: All differential item functioning (DIF) methods require at least a moderate sample size for effective DIF detection. Samples that are less than 200 pose a challenge for DIF analysis. Smoothing can improve upon the estimation of the population distribution by preserving major features of an observed frequency distribution while eliminating the noise brought about by irregular data points. This study applied smoothing techniques to frequency distributions and investigated the impact of smoothed data on the Mantel-Haenszel (MH) DIF detection in small samples. Eight sample-size combinations were randomly drawn from a real data set to make the study realistic and were replicated 80 times to produce stable results. The population DIF results were used as the criteria to evaluate sample estimates using root-mean square difference (RMSD), bias analysis, and Type II error rate. Loglinear smoothing was found to provide slight to moderate improvements in MH DIF estimation with small samples

To order a copy of this report, write to R&DWeb@ets.org.

Comparison of Multidimensional Item Response Models: Multivariate Normal Ability Distributions Versus Multivariate Polytomous Distributions

Report Number: RR-08-45

Author(s): S. J. Haberman, M. von Davier, & Y.-H. Lee

Abstract: Multidimensional item response models can be based on multivariate normal ability distributions or on multivariate polytomous ability distributions. For the case of simple structure in which each item corresponds to a unique dimension of the ability vector, some applications of the two-parameter logistic model to empirical data are employed to illustrate how, at least for the example under study, comparable results can be achieved with either approach.

Comparability involves quality of model fit as well as similarity in terms of parameter estimates and computational time required. In both cases, numerical work can be performed quite efficiently. In the case of the multivariate normal ability distribution, multivariate adaptive Gauss-Hermite quadrature can be employed to greatly reduce computational labor. In the case of a polytomous ability distribution, use of log-linear models permits efficient computations.

To order a copy of this report, write to R&DWeb@ets.org.

Understanding What the Numbers Mean: A Straightforward Approach to GRE® Predictive Validity

Report Number: RR-08-46, GREB-04-03

Author(s): B. Bridgeman, N. Burton, & F. Cline

Abstract: Descriptions of validity results for the GRE® General Test based solely on correlation coefficients or percentage of the variance accounted for are not merely difficult to interpret, they are likely to be misinterpreted. Predictors that apparently account for a small percentage of the variance may actually be highly important from a practical perspective. This study used 2 existing data sets to demonstrate alternative methods of showing the value of the GRE as an indicator of 1st-year graduate grades. The combined data sets contained 4,451 students in 6 graduate fields: biology, chemistry, education, English, experimental psychology, and clinical psychology. In one set of analyses, students within a department were divided into quartiles based on GRE scores and the percentage of students in the top and bottom quartiles earning a 4.0 average was noted. Students in the top quartile were 3 to 5 times as likely to earn 4.0 averages compared to students in the bottom quartile. Even after controlling for undergraduate grade point average quartiles, substantial differences related to GRE quartile remained.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-46.pdf

Measuring Learning Outcomes in Higher Education Using the Measure of Academic Proficiency and Progress (MAPP)

Report Number: RR-08-47

Author(s): O. L. Liu

Abstract: The Secretary of Education's Commission on the Future of Higher Education emphasizes accountability in higher education as one of the key areas of interest. The Voluntary System of Accountability (VSA) was developed to evaluate the effectiveness of general public college education. This study examines how student progress in college, indicated by the performance difference between freshmen and seniors after controlling for admission scores, can be measured using the Measure of Academic Proficiency and Progress (MAPP) test. A total of 6,196 students from 23 institutions were included in this study. Results indicated that MAPP was able to differentiate the performance between freshmen and seniors after controlling for SAT®/ACT scores. The institutions were classified into 10 groups on the basis of the difference in the actual vs. expected MAPP performance. This study provides an example of how MAPP can be used to evaluate value-added performance in college education. Issues such as student sampling and test-taking motivation are discussed.

To order a copy of this report, write to R&DWeb@ets.org.

Testing Accommodations for English Language Learners: A Review of State and District Policies

Report Number: RR-08-48, CBR-2008-06

Author(s): J. W. Young & T. King

Abstract: This report is a review and summary of current information regarding testing accommodations currently used in different states and districts for English language learners (ELLs). The federal No Child Left Behind (NCLB) Act of 2001 requires the inclusion of ELLs in assessments used by the states for accountability purposes. This represents a federal education requirement that did not exist prior to the enactment of NCLB. However, the policies for identification and reclassification of ELLs, appropriate testing accommodations, and testing requirements are state-level decisions. In order to validly and fairly assess the skills of ELL students, testing accommodations are made available where necessary by the states. However, there is no common set of standards across the states as to what are appropriate accommodations permitted for ELLs. Similarities and differences among states regarding ELL testing accommodations are documented in this review. Special attention is given to the ELL accommodation policies for states with high school exit examinations because these are the high-stakes exams, which have the clearest relevance in designing accommodation policies for ELLs in taking the SAT®.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-48.pdf

Design Patterns for Improving Accessibility for Test Takers With Disabilities

Report Number: RR-08-49

Author(s): E. G. Hansen & R. J. Mislevy

Abstract: There is a great need to help test designers determine how to make tests that are accessible to individuals with disabilities. This report takes design patterns, which were developed at SRI for assessment design, and uses them to clarify issues related to accessibility features for individuals with disabilities—such as low-vision and blindness—taking a test of reading. Design patterns appear useful in clarifying how variable features of a test design need to be matched to disability-related characteristics of test takers in order to ensure accessibility. Giving consideration to accessibility issues during the development and use of design patterns may help improve the validity and fairness of tests, as well as their accessibility for individuals with disabilities.

To order a copy of this report, write to R&DWeb@ets.org.

Development and Validity Evidence Supporting a Teamwork and Collaboration Assessment for High School Students

Report Number: RR-08-50

Author(s): X. Zhuang, C. MacCann, L. Wang, O. L. Liu, & R. D. Roberts

Abstract: Various policy papers and research studies assert that teamwork is one of the most important skills for students to learn if they are to become meaningful contributors to the 21st century workforce. However, outside of organizational psychology and adult populations, few reliable assessments of this construct exist, with suitable validity evidence scant or nonexistent. To redress this imbalance, teamwork assessments for high school students were developed using multiple methods: self-report ratings, situational judgment testing, and teacher reports. Exploratory factor, confirmatory factor, and latent class analyses were used to determine the structure of the scales. Measures showed reasonable reliability and promising validity evidence, relating to each other and to academic achievement, while remaining relatively independent from personality. The advantages and disadvantages of each methodology and the potential applications for identification and intervention, selection, and evaluation of training programs are discussed. This report also serves as an archival document for the teamwork and collaboration assessments that have been developed at ETS for high school students.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-50.pdf

Applying Content Similarity Metrics to Corpus Data: Differences Between Native and Non-Native Speaker Responses to a TOEFL® Integrated Writing Prompt

Report Number: RR-08-51

Author(s): P. Deane & O. Gurevich

Abstract: For many purposes, it is useful to collect a corpus of texts all produced to the same stimulus, whether to measure performance (as on a test) or to test hypotheses about population differences. This paper examines several methods for measuring similarities in phrasing and content and demonstrates that these methods can be used to identify population differences between native and non-native speakers of English in a writing task.

To order a copy of this report, write to R&DWeb@ets.org.

Investigating the Effectiveness of Collateral Information on Small-Sample Equating

Report Number: RR-08-52

Author(s): S. Kim, S. A. Livingston, & C. Lewis

Abstract: This paper describes an empirical evaluation of a Bayesian procedure for equating scores on test forms taken by small numbers of examinees, using collateral information from the equating of other test forms. In this procedure, a separate Bayesian estimate is derived for the equated score at each raw-score level, making it unnecessary to specify a parametric model for the equating function. Collateral information can come either from other forms of the same test or, possibly, from other tests having a similar structure. Our evaluation consisted of two resampling studies. Each study applied the Bayesian procedure to small samples drawn from large-sample data collected for an anchor equating. The large-sample equating function served as the criterion. The results of the two studies were somewhat inconsistent, leading to different conclusions regarding the use of the empirical Bayesian procedure with small samples.

To order a copy of this report, write to R&DWeb@ets.org.

Comparisons Among Designs for Equating Constructed-Response Tests

Report Number: RR-08-53

Author(s): S. Kim, M. E. Walker, & F. McHale

Abstract: This study examined variations of a nonequivalent groups equating design used with constructed-response (CR) tests to determine which design was most effective in producing equivalent scores across the two tests to be equated. Using data from a large-scale exam, the study investigated the use of anchor CR item rescoring in the context of classical equating methods. Four linking designs were examined: (a) an anchor set containing common CR items, (b) an anchor set incorporating common CR items rescored, (c) an external multiple-choice (MC) anchor test, and (d) an equivalent groups design incorporating CR items rescored (no anchor test). The use of CR items without rescoring or the use of an external MC anchor resulted in much larger bias than the other two designs. The use of a rescored CR anchor and the equivalent groups design led to similar levels of equating error.

To order a copy of this report, write to R&DWeb@ets.org.

Comparison of Subscores Based on Classical Test Theory Methods

Report Number: RR-08-54

Author(s): G. Puhan, S. Sinharay, S. J. Haberman, & K. Larkin

Abstract: Will reporting subscores provide any additional information than the total score? Is there a method that can be used to provide more trustworthy subscores than observed subscores? These 2 questions are addressed in this study. To answer the 2nd question, 2 subscore estimation methods (i.e., subscore estimated from the observed total score or subscore estimated using both the observed subscore and observed total score) are compared. Analyses conducted on 8 certification tests indicated that reporting subscores at the examinee level may not be necessary as they do not provide much additional information than the total score. However, at the institutional level (for institution size greater than 30), reporting subscores may not be harmful, although it may be redundant. Finally, results indicated that subscores estimated using both the observed subscore and observed total score were the most trustworthy and may be used if subscores were to be reported.

To order a copy of this report, write to R&DWeb@ets.org.

Cognitive Models of Writing: Writing Proficiency as a Complex Integrated Skill

Report Number: RR-08-55

Author(s): P. Deane, N. Odendahl, T. Quinlan, M. Fowles, C. Welsh, & J. Bivens-Tatum

Abstract: This paper undertakes a review of the literature on writing cognition, writing instruction, and writing assessment with the goal of developing a framework and competency model for a new approach to writing assessment. The model developed is part of the Cognitively Based Assessments of, for, and as Learning (CBAL) initiative, an ongoing research project at ETS intended to develop a new form of kindergarten through Grade 12 (K–12) assessment that is based on modern cognitive understandings; built around integrated, foundational, constructed-response tasks that are equally useful for assessment and for instruction; and structured to allow multiple measurements over the course of the school year. The model that emerges from a review of the literature on writing places a strong emphasis on writing as an integrated, socially situated skill that cannot be assessed properly without taking into account the fact that most writing tasks involve management of a complex array of skills over the course of a writing project, including language and literacy skills, document-creation and document-management skills, and critical-thinking skills. As such, the model makes strong connections with emerging conceptions of reading and literacy, suggesting an assessment approach in which writing is viewed as calling upon a broader construct than is usually tested in assessments that focus on relatively simple, on-demand writing tasks.

To order a copy of this report, write to R&DWeb@ets.org.

The Redesigned TOEIC® (Listening and Reading) Test: Relations to Test Taker Perceptions of Proficiency in English

Report Number: RR-08-56

Author(s): D. E. Powers, H.-J. Kim, & V. Z. Weng

Abstract: To facilitate the interpretation of test scores from the redesigned TOEIC® (listening and reading) test as a measure of English language proficiency, we administered a self-assessment inventory to TOEIC examinees in Japan and Korea that gathered perceptions of their ability to perform a variety of everyday English language tasks. TOEIC scores related relatively strongly to test-taker self-reports for both reading and listening tasks. The results were, with few exceptions, extraordinarily consistent, with examinees at each higher TOEIC score level being more likely to report that they could successfully accomplish each of the everyday language tasks in English. The pattern of correlations also showed modest discriminant validity of the listening and reading components of the redesigned TOEIC, suggesting that both sections contribute to the measurement of English language skills.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-56.pdf

Mapping State Standards to the NAEP Scale

Report Number: RR-08-57

Author(s): H. Braun & J. Qian

Abstract: This report describes the derivation and evaluation of a method for comparing the performance standards for public school students set by different states. It is based on an approach proposed by McLaughlin and associates, which constituted an innovative attempt to resolve the confusion and concern that occurs when very different proportions of students in various states are declared to have met a standard with the same label. Our method, like McLaughlin's, employs equipercentile methods to map state standards on to a common scale, that associated with the National Assessment of Educational Progress (NAEP). We have also derived error estimates that take into account both NAEP's complex sampling design and measurement errors. The method was applied to two data sets, and the results were qualitatively similar to those obtained by McLaughlin's method. The paper notes the superior statistical properties of the proposed method and presents evidence that supports the viability and general utility of this approach.

To order a copy of this report, write to R&DWeb@ets.org.

Establishing the Validity of TOEIC Bridge™ Test Scores for Students in Colombia, Chile, and Ecuador

Report Number: RR-08-58

Author(s): S. Sinharay, Y. Feng, L. Saldivia, D. E. Powers, A. Ginuta, A. Simpson, & V. Z. Weng

Abstract: The validity of TOEIC Bridge™ scores as a measure of English language skill was examined from the standpoint of a unified concept of test validity. In this study, more than 6,000 test takers in 3 Latin American countries (Chile, Colombia, and Ecuador) took 1 form of the TOEIC Bridge test, and their scores were compared to additional information about the students (teacher judgments, self-assessments, and performance on other academic achievement measures). The evidence collected was generally quite consistent with the interpretation of TOEIC Bridge scores as indicators of the English language competencies for the students examined.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-58.pdf

Notes on a General Framework for Observed Score Equating

Report Number: RR-08-59

Author(s): T. Moses & P. W. Holland

Abstract: The purpose of this paper is to extend von Davier, Holland, and Thayer's (2004b) framework of kernel equating so that it can incorporate raw data and traditional equipercentile equating methods. One result of this more general framework is that previous equating methodology research can be viewed more comprehensively. Another result is that the standard error of equated score difference (SEED) has a wider application than originally proposed. The methods described in this paper are empirically evaluated in an accompanying simulation study (Moses & Holland, 2007).

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-59.pdf

An Evaluation of Statistical Strategies for Making Equating Function Selections

Report Number: RR-08-60

Author(s): T. Moses

Abstract: Nine statistical strategies for selecting equating functions in an equivalent groups design were evaluated. The strategies of interest were likelihood ratio chi-square tests, regression tests, Kolmogorov-Smirnov tests, and significance tests for equated score differences. The most accurate strategies in the study were the likelihood ratio tests and the significance tests for equated score differences.

Full report available from
http://www.ets.org/Media/Research/pdf/RR-08-60.pdf

Linking With Continuous Exponential Families: Single-Group Designs

Report Number: RR-08-61

Author(s): S. J. Haberman

Abstract: Continuous exponential families are applied to linking forms via a single-group design. In this application, a distribution from the continuous bivariate exponential family is used that has selected moments that match those of the bivariate distribution of scores on the forms to be linked. The selected continuous bivariate distribution then yields continuous univariate marginal distributions for the two forms. These marginal distributions then provide distribution functions and quantile functions that may be employed in equating. Normal approximations are obtained for the sample distributions of the conversion functions.

To order a copy of this report, write to R&DWeb@ets.org.

Automated Scoring of Spontaneous Speech Using SpeechRater v1.0

Report Number: RR-08-62

Author(s): X. Xi, D. Higgins, K. Zechner, & D. M. Williamson

Abstract: This report presents the results of a research and development effort for SpeechRaterSM Version 1.0 (v1.0), an automated scoring system for the spontaneous speech of English language learners used operationally in the Test of English as a Foreign Language™ (TOEFL®) Practice Online assessment (TPO). The report includes a summary of the validity considerations and analyses that drive both the development and the evaluation of the quality of automated scoring. These considerations include perspectives on the construct of interest, the context of use, and the empirical performance of the SpeechRater in relation to both the human scores and the intended use of the scores. The outcomes of this work have implications for short- and long-term goals for iterative improvements to SpeechRater scoring.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-62.pdf

Studies of a Latent-Class Signal-Detection Model for Constructed-Response Scoring

Report Number: RR-08-63

Author(s): L. T. DeCarlo

Abstract: Rater behavior in essay grading can be viewed as a signal-detection task, in that raters attempt to discriminate between latent classes of essays, with the latent classes being defined by a scoring rubric. The present report examines basic aspects of an approach to constructed-response (CR) scoring via a latent-class signal-detection model. The model provides a psychological framework for CR scoring and includes rater parameters with a clear cognitive basis. Simulations are used to examine how well rater parameters and latent-class sizes are recovered as well as the accuracy of classification. The relation of rater parameters to agreement statistics and classification accuracy is examined. The effects of using a balanced, incomplete block design are compared to those for a fully crossed design. The model is applied to several ETS datasets.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-63.pdf

Subscores and Validity

Report Number: RR-08-64

Author(s): S. J. Haberman

Abstract: In educational testing, subscores may be provided based on a portion of the items from a larger test. One consideration in evaluation of such subscores is their ability to predict a criterion score. Two limitations on prediction exist. The first, which is well known, is that the coefficient of determination for linear prediction of the criterion score by the subscore cannot exceed the reliability coefficient of the subscore. The second limitation is on incremental validity. The coefficient of determination for linear prediction of the criterion score by both the total score and the subscore is at least as great as the coefficient of determination for linear prediction of the criterion score by only the total score. Incremental validity may be measured by the difference between these two coefficients of determination. This difference is no greater than the reliability of the residual from linear prediction of the subscore by the total score.

To order a copy of this report, write to R&DWeb@ets.org.

Effective and Scalable Teacher Professional Development: A Report of the Formative Research and Development

Report Number: RR-08-65

Author(s): E. C. Wylie, C. J. Lyon, & E. Mavronikolas

Abstract: This study is a qualitative analysis of data collected during a yearlong series of teacher learning community meetings and classroom observations. The participants are middle and high school mathematics teachers from 2 school districts. Teachers were introduced to the research behind formative assessment

and how to apply that research to their teaching during a 3-day summer workshop. Teachers then met monthly in small school-based groups to deepen their understanding of formative assessment and to talk about their own classroom experiences. The analyses focused on how teachers' understanding changed over time, how the teacher learning communities supported the teachers, how learning translated into classroom practice, and the factors that supported or hindered the development of teachers' understanding and practice. Lessons learned during the study itself and from subsequent analyses of the data had a significant impact on the development of ETS's Keeping Learning on Track® program.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-65.pdf

Factor Structure of the TOEFL® Internet-Based Test Across Subgroups

Report Number: RR-08-66, TOEFLiBT-07

Author(s): L. J. Stricker & D. A. Rock

Abstract: This study assessed the invariance in the factor structure of the Test of English as a Foreign Language™ Internet-based test (TOEFL® iBT) across subgroups of test takers who differed in native language and exposure to the English language. The subgroups were defined by (a) Indo-European and Non-Indo-European language family, (b) Kachru's classification of outer and expanding circles of countries (based on prevalence of English use in educational and business contexts), and (c) years of classroom instruction in the English language. The same factor structure (four first-order factors corresponding to the test sections and a single higher-order factor encompassing these factors) was identified in each subgroup. The results support the present scoring scheme for the TOEFL iBT assessment and suggest that the test functions the same way for diverse subgroups of test takers.

Full report available from:
http://www.ets.org/Media/Research/pdf/RR-08-66.pdf

Consistency of SAT® I: Reasoning Test Score Conversions

Report Number: RR-08-67

Author(s): S. J. Haberman, H. Guo, J. Liu, & N. J. Dorans

Abstract: This study uses historical data to explore the consistency of SAT® I: Reasoning Test score conversions and to examine trends in scaled score means. During the period from April 1995 to December 2003, both Verbal (V) and Math (M) means display substantial seasonality, and a slight increasing trend for both is observed. SAT Math means increase more than SAT Verbal means. Several statistical indices indicate that, during the period under study, raw-to-scale conversions are very stable, although conversions for extreme raw score points are less stable than are other conversions.

To order a copy of this report, write to R&DWeb@ets.org.

Educational Assessment Using Intelligent Systems

Report Number: RR-08-68

Author(s): V. J. Shute & D. Zapata-Rivera

Abstract: Recent advances in educational assessment, cognitive science, and artificial intelligence have made it possible to integrate valid assessment and instruction in the form of modern computer-based intelligent systems. These intelligent systems leverage assessment information that is gathered from various sources (e.g., summative and formative). This paper analyzes the role of educational assessment in intelligent systems, summarizes the characteristics of successfully deployed intelligent systems, and describes an evidence-based approach to incorporating valid and reliable assessments into enhanced intelligent systems.

To order a copy of this report, write to R&DWeb@ets.org.

Monitoring and Fostering Learning Through Games and Embedded Assessments

Report Number: RR-08-69

Author(s): V. J. Shute, M. Ventura, M. Bauer, & D. Zapata-Rivera

Abstract: To reveal what is being learned during the gaming experience, this report proposes an approach for embedding assessments in immersive games, drawing on recent advances in assessment design. Key to this approach are formative assessment to guide instructional experiences and evidence-centered design to systematically analyze the assessment argument (including the claims about the learner and the evidence that supports or fails to support those claims). Elements of this approach that have been applied in a nongame setting are shown and ideas are discussed for applying the approach to an existing immersive game setting. Finally, the report offers suggestions for extending and applying this approach for existing games and the design of new ones.

To order a copy of this report, write to R&DWeb@ets.org.

Reliability of Scaled Scores

Report Number: RR-08-70

Author(s): S. J. Haberman

Abstract: The reliability of a scaled score can be computed by use of item response theory. Estimated reliability can be obtained even if the item response model selected is not valid.

To order a copy of this report, write to R&DWeb@ets.org.

The Fusion Model for Skills Diagnosis: Blending Theory With Practicality

Report Number: RR-08-71

Author(s): S. Hartz & L. Roussos

Abstract: The current paper presents the development of the Fusion Model Skills Diagnosis System, which can help integrate standardized testing into the learning process with both skills-level examinee parameters for modeling examinee skill mastery and skills-level item parameters, giving information about the diagnostic power of the test. The development of the Fusion Model System involves advancements in modeling, parameter estimation, model fitting methods, and model fit evaluation procedures, which are described in detail in the paper. To document the accuracy of the estimation procedure and the effectiveness of the model fitting and model fit evaluation procedures, the current paper also presents a series of simulation studies. Special attention is given to evaluating the robustness of the Fusion Model System to violations of various modeling assumptions. The results demonstrate that the Fusion Model System is a promising tool for skills diagnosis that merits further research and development.

To order a copy of this report, write to R&DWeb@ets.org.

Listening. Learning. Leading.®

www.ets.org