Listening. Learning. Leading.®

ETS Research Spotlight

ISSUE 3, MARCH 2010

Acknowledgments

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

Reviews and Editing: Dan Eignor, Jinghua Liu, Hyeon-Joo Oh, Michael Zieky, Jeff Johnson, Eileen Kerrigan

Cover Design: Marita Gray

Layout: Jeff Johnson

Production Coordination: Bill Petzinger

ETS Research Spotlight
Issue 3

The views expressed in this report are those of the authors and do not necessarily reflect the views of the officers and trustees of Educational Testing Service.

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

March 2010
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: RDWeb@ets.org

Table of Contents

Foreword

ETS's mission statement asserts that "our products and services measure knowledge and skills, promote learning and educational performance, and support education and professional development for all people worldwide." We could not make this claim without the capabilities of our Research & Development (R&D) division. Issue 3 of ETS Research Spotlight features three concrete examples of how our scientists and psychometricians strive not only to ensure quality in assessments that we produce, but also to advance the field of educational measurement.

The first article, *Relationship Between Scores on the TOEIC® Speaking and Writing Tests and Test Takers' Perceptions of Their English Proficiency*, describes efforts to link test scores to so-called can-do statements. The goal is to help the TOEIC test's score users to make useful and appropriate interpretations about score meaning.

In *First Language of Examinees and Empirical Assessment of Fairness*, we report on statistical methods of maintaining standards for fairness, validity, and reliability even as test-taking populations change.

Finally, *Highlights from the Cognitively Based Assessment* of, for, *and* as *Learning (CBAL) Project in Mathematics* illustrates an example of assessment innovation: The CBAL project aims to introduce a new model for K – 12 assessment, unifying accountability assessment, formative assessment, and teacher professional development.

These reports represent just a sample of the breadth of ETS's assessment-related research. More reports are available in ReSEARCHER, ETS's searchable database, at http://search.ets.org/custres. If you have questions about any of the work in our research portfolio, please contact us at RDWeb@ets.org.

Ida Lawrence
Senior Vice President
ETS Research & Development

Relationship Between Scores on the TOEIC® Speaking and Writing Tests and Test Takers' Perceptions of Their English Proficiency[1]

Donald E. Powers, Hae-Jin Kim, Feng Yu, Vincent Z. Weng, and Waverely VanWinkle

To facilitate the interpretation of test scores from the TOEIC® Speaking and Writing tests as measures of English-language proficiency, researchers administered a self-assessment inventory to examinees in Japan and Korea, to gather their perceptions of their ability to perform everyday English-language tasks. TOEIC scores related relatively strongly to these test-taker self-reports. At each higher score level, examinees were more likely to report that they could successfully accomplish each of the everyday language tasks in English. The pattern of correlations also revealed modest discriminant validity of the speaking and writing measures, suggesting that both measures contribute uniquely to the assessment of English-language proficiency.

As the professional standards for educational and psychological testing make clear, the most fundamental concern in test evaluation is validity. One important source of validity evidence is the degree to which test scores relate to important variables that are external to the test. For a variety of reasons, however, variables (validation criteria) that are both defensible and compelling are often difficult to come by. For example, the various criteria that have been used to validate academic admissions tests (e.g., course grades, faculty ratings, qualifying examinations, degree attainment) all suffer from certain limitations. Their meaning may be unclear, they may not be comparable for all test takers, they may have undesirable psychometric properties (e.g., restricted range, skewed distributions, or low reliability), and they may be difficult to obtain (Hartnett & Willingham, 1980).

Self-assessments of various sorts — self-reports, checklists, self-testing, mutual peer assessment, diaries, log books, behaviorally anchored questionnaires, global proficiency scales, and "can-do" statements (Oscarson, 1989, 1997) — have proven to be useful validation criteria in a variety of low-stakes contexts, especially in the assessment of language skills. In this regard, Upshur (1975) noted that language learners often have access to the full spectrum of their successes and failures, whereas other assessments (or assessors) may have a much narrower view. Similarly, Shrauger and Osberg (1981) noted that, being active observers of their own behavior, people often have extensive data — often more than is available to external evaluators — on which to base their judgments. Shrauger and Osberg (1981) concluded that there is substantial support, both empirical and conceptual, for the notion that self-assessors frequently have both the information and the motivation to make effective judgments about themselves.

Like other criteria, however, self-assessments also have certain limitations. The main limitations arise because people may lack objectivity when reporting about themselves, or they may be unaware of some of their most important characteristics. In addition, they may intentionally exaggerate their skills and abilities, or they may attempt to present themselves in socially desirable ways. It is probably unwise, therefore, to trust the results of self-assessments completely when the stakes are high.

The TOEIC® test was developed to measure the ability to listen and read in English, using a variety of contexts from real-world settings. ETS added TOEIC Speaking and Writing tests to the TOEIC product line in order to directly assess the ability to speak and write in English in a workplace setting. This addition was in response to multinational corporations' need for employees with high-level speaking and writing skills.

The research described here was designed to provide evidence of the validity of the TOEIC Speaking and Writing tests as measures of English-language proficiency. We planned to establish this by investigating the relationship between scores on the measures and TOEIC test takers' reports of their ability to perform selected everyday English speaking and writing tasks in the workplace.

Method

In fall 2008, after assembling a self-report can-do inventory of speaking and writing tasks, we administered the inventory to individuals who took the TOEIC Speaking and Writing tests in Japan and Korea. We followed several steps in the development of this inventory. First, we assembled a preliminary list

[1] This article is based on ETS Research Report No. RR-09-18, which appears in the reference list as Powers, Kim, Yu, Weng, & VanWinkle (2009). The full report is available from the ETS ReSEARCHER database at http://search.ets.org/custres/.

of tasks for review by major clients in Japan and Korea. This list drew heavily from a list developed by Ito, Kawaguchi, and Ohta (2005) as well as from previous research (e.g., Duke, Kao, & Vale, 2004; Tannenbaum, Rosenfeld, Breyer, & Wilson, 2007). From these sources, we selected can-do task statements and translated them from English into Japanese and Korean. In language assessment, can-do statements imply a claim of what test takers should be able to do at a given level of proficiency. Native speakers of Japanese or Korean checked the translations.

Next, we invited TOEIC clients in Japan and Korea to review the preliminary list. These clients were relatively large companies that have significant language-training programs, and are therefore well versed in communication problems encountered in the workplace. For each task listed in the inventory, clients rated the importance of being able to perform the task with regard to the kind of job (or family of jobs) for which they were reporting. The specific question was, "How important is it that a worker be able to perform this task competently in order to perform his/her job satisfactorily?" Responses were on a 6-point scale (0 = Does not have to perform this task as part of the job, 1 = Slightly important, 2 = Somewhat important, 3 = Important, 4 = Very important, 5 = Extremely important).

After they indicated their ratings, respondents were asked to think about the job or family of jobs for which they were reporting and to list any important job tasks that were not included on the preliminary list. In addition, they were encouraged to indicate changes or alternative wording for any of the tasks that seemed unclear. In total, 23 company representatives from Korea and 24 from Japan returned responses. Between the two countries, the agreement on task importance was reasonably good, with average ratings of tasks correlating .67 for speaking and .70 for writing.

Respondents suggested a number of additional tasks, several of which we added to the inventory. However, some suggested tasks were unique to particular industries or jobs. Because these tasks had limited applicability to the market in general, we did not add them to the inventory. Also, we deleted the listed tasks that respondents had rated lowest in importance. The final version of the inventory comprised 40 common language tasks (can-do statements) for speaking and 29 for writing. In the fall of 2008, this final inventory was administered in Japan and Korea to test takers who were taking the TOEIC Speaking and Writing tests.

In completing the inventory, test takers used a 5-point scale to rate how easily they could perform each task: 1 = not at all, 2 = with great difficulty, 3 = with some difficulty, 4 = with little difficulty, and 5 = easily. Respondents were encouraged to respond to each statement, but they were allowed to omit a task statement if they thought it did not apply to them or they were unable to make a judgment.

Results

We obtained data from 2,947 test takers in Korea and 867 in Japan. TOEIC Speaking scores were available for 3,518 participants; TOEIC Writing scores were available for 1,472 participants. Approximately 46% of the participants were female. More than three-fourths (78%) of the participants had either completed or were currently pursuing a bachelor's degree, another 14% had completed or were pursuing a graduate degree, and about 5% had completed or were pursuing an associate's degree at a 2-year college. The study sample was nearly equally divided between full-time students (43%) and full-time employees (42%). About 10% of all respondents reported being unemployed; 5% of respondents reported that they either worked or studied part-time. Employed participants reported holding a wide variety of jobs.

The correlation between TOEIC Speaking and TOEIC Writing scores was high (.71), as was the correlation between the speaking and writing can-do reports (.87). More importantly, speaking can-do reports (total over all tasks) correlated relatively strongly (.54) with TOEIC Speaking scores, and writing can-do reports correlated equally strongly with TOEIC Writing scores (.52). TOEIC Speaking scores correlated slightly less with writing can-do reports (.49) than with speaking can-do reports, and TOEIC Writing scores correlated slightly less with speaking can-do reports (.51) than with writing can-do reports. This pattern suggests very modest discriminant validity of the two TOEIC scores.

To indicate how test performance relates to each can-do task, we have presented (in Table 1 [p. 6] for speaking and Table 2 [p. 7] for writing) task-by-task results for a small, representative sample of tasks (mean response on the 5-point scale). The numbers shown in the tables are the proportions of test takers at each of several score intervals who said that they could perform the task either easily or with little difficulty. Table 1 shows the number of participants at each TOEIC Speaking score level who said they were able to perform various speaking tasks easily in English. Table 2 displays similar data by TOEIC Writing score level. As Tables 1 and 2 show, for each task, higher test performance is associated with a greater likelihood of reporting

Table 1. **Percentages of TOEIC Test Takers, by Speaking-Score Level, Who Indicated They Could Perform Various English-Speaking Tasks Easily or With Little Difficulty**

	Score level						
I can:	0-50	60-70	80-100	110-120	130-150	160-180	190-200
make/change/cancel an appointment to see a person	19	32	43	65	78	91	100
telephone a company to place (or follow-up) an order for an item	16	22	34	56	67	83	96
have "small talk" with a guest about topics of general interest (e.g., the weather) before discussing business	10	24	35	57	69	83	94
explain (to a co-worker or colleague) how to operate a machine or device (e.g., photocopier, PC, audio player) that I am familiar with	11	25	29	43	51	68	86
state and emphasize my opinion during a discussion or meeting	2	6	13	26	36	51	73
serve as an interpreter for top management on various occasions such as business negotiations and courtesy calls	2	3	6	11	18	28	47
Sample size for score interval	65	176	658	819	1,333	417	50

successful task performance. This was true for each of the many other tasks that are not shown. (Results for all tasks are available in Powers, Kim, Yu, Weng, & VanWinkle, 2009.)

Because some test-score users preferred a more narrative presentation of results, we also prepared an alternative presentation, categorizing the tasks that test takers at various test-score levels (a) are likely to be able to perform with little difficulty, (b) are likely to be able to perform with difficulty, and (c) are unlikely to be able to perform at all. We used the following convention to classify tasks into these three levels. Test takers at a given score level were considered likely to be able to perform a particular task (probably can do) if at least 50% of them reported that they could perform the task either easily or with little difficulty. If at least 50% of test takers at a score level said they could not perform a task at all, or could perform it only with great difficulty, then they were considered as being unlikely to be able to perform the task (probably cannot do). If a task could not be classified as either probably can do or probably cannot do by these criteria, it was classified as probably can do with difficulty if at least 50% of test takers said they could perform the task with little difficulty, some difficulty, or great difficulty. Using these criteria, all speaking and all writing tasks could be placed into one (and only one) of the three categories. (See Table 3 [p. 8] for a sample of tasks at one TOEIC Speaking score level and Table 4 [p. 8] for a sample of tasks at one TOEIC Writing score level.)

Discussion/Implications

One kind of evidence that has proven useful in elucidating the meaning, or validity, of language-test scores has come from examinees themselves, in the form of self-assessments of their own language skills. Although self-assessments may sometimes be susceptible to distortion (either unintentional or deliberate), results from them have been shown to be valid in a variety of contexts (see, e.g., Falchikov & Boud, 1989; Harris & Schaubroeck, 1988; Mabe & West, 1982), especially in the assessment of language skills (LeBlanc & Painchaud, 1985; Upshur, 1975; Shrauger & Osberg, 1981). It has even been asserted (e.g., Upshur, 1975; Shrauger & Osberg, 1981) that, in some respects, language learners often have more complete knowledge of their linguistic successes and failures than do third-party assessors.

For this study, a large-scale data collection effort was undertaken to establish links between (a) test-takers' performance on the TOEIC Speaking and Writing tests and (b) self-assessments

Table 2. **Percentages of TOEIC Test Takers, by Writing-Score Level, Who Indicated They Could Perform Various English-Writing Tasks Easily or With Little Difficulty**

	Score level					
I can:	0-80	90-100	110-130	140-160	170-190	200
write an email requesting information about hotel accommodations	12	37	48	70	81	91
write clear directions on how to get to my office	14	23	37	58	69	86
translate documents (e.g., business letters, manuals) into English	11	18	24	39	54	81
write discussion notes during a meeting or class and summarize them	11	17	19	37	53	79
write a formal letter of thanks to a client	11	20	24	38	50	71
prepare text and slides (in English) for a presentation at a professional conference	7	11	17	28	44	69
Sample size for score interval	44	85	313	590	363	77

of their ability to perform a variety of common, everyday language tasks in English. Results revealed that, for both speaking and writing, TOEIC scores were relatively strongly related to test takers' self-assessments, both overall and for each individual task. For instance, conventional standards consider the magnitude of the correlations observed in the study reported here to fall into the large range (.50 and above) with respect to effect size (Cohen, 1988). Moreover, the correlations that were observed here compare very favorably with those typically observed in validity studies that use other kinds of validation criteria, such as course grades, faculty ratings, and degree completion. For example, in a recent very large-scale meta-analysis of graduate-level academic admissions tests, Kuncel and Hezlett (2007) reported that, over all the different tests that they considered, first-year grade average — the most predictable of several criteria available — correlated, on average, about .45 with test scores. The correlations observed here also compared favorably with those (in the .30s and .40s) found between overall student self-assessments and performance on the TOEFL iBT™ exam (Powers, Roever, Huff, & Trapani, 2003).

In addition, the pattern of correlations among the measures also indicated modest discriminant validity of the TOEIC Speaking and Writing measures, suggesting that each contributes uniquely to the measurement of English-language skills. This result is consistent with a recent factor-analytic study of a similar test (the TOEFL iBT) by Sawaki, Stricker, and Oranje (2008), in which the correlation (r =.71) suggested relatively highly related, but distinct, speaking and writing factors.

In the present study, we were not able to evaluate the soundness of test-taker self-reports as a validity criterion. However, in comparable studies that we have conducted recently in similar contexts, can-do self-reports have exhibited several characteristics that suggest that they are reasonably trustworthy validity criteria, especially for low-stakes research, in which examinees have no incentive to intentionally distort their reports. For example, we have found that examinees rank-order the difficulty of tasks in accordance with our expectations (Powers, Bravo, & Locke, 2007; Powers et al., 2008) and that they exhibit reasonably stable agreement about task difficulty when self-reports are collected again on later occasions (Powers et al., 2008). In addition, the current study's results are consistent with previous meta-analytic summaries (e.g., Ross, 1998) that have documented substantial correlations between various criterion measures and the self-ratings of learners of English as a second language.

In conclusion, the current study provides evidence of the validity of TOEIC Speaking and Writing scores by linking them to test takers' assessments of their ability to perform a variety of everyday (often job-related) English-language activities. The practical implication of these linkages lies in their ability to facilitate the interpretation and use of TOEIC scores. The results strongly suggest that TOEIC Speaking and Writing scores can distinguish between test takers who are

Table 3. **Sample Can-Do Table for TOEIC Speaking — Classification of Test-Taker Perceptions at Scaled Score Level 110–120**

Probably can do

- Make/change/cancel an appointment to see a person
- Have “small talk” with a guest about topics of general interest (e.g., the weather) before discussing business
- Telephone a company to place (or follow-up) an order for an item

Probably can do with difficulty

- Explain (to a co-worker or colleague) how to operate a machine or device (e.g., photocopier, PC, audio player) that I am familiar with
- State and emphasize my opinion during a discussion or meeting

Probably cannot do

- Serve as an interpreter for top management on various occasions such as business negotiations and courtesy calls

Table 4. **Sample Can-Do Table for TOEIC Writing — Classification of Test-Taker Perceptions at Scaled Score Level 140–160**

Probably can do

- Write an e-mail requesting information about hotel accommodations
- Write clear directions on how to get to my office

Probably can do with difficulty

- Translate documents (e.g., business letters, manuals) into English
- Write discussion notes during a meeting or class and summarize them
- Write a formal letter of thanks to a client
- Prepare text and slides (in English) for a presentation at a professional conference

Probably cannot do

- None

likely to be able to perform these tasks and those who are not. According to most conventional standards, the relationships that we detected are practically meaningful. To the degree that the language tasks studied here are important for success in a global business environment, using the TOEIC test to recruit, hire, or train prospective employees should be a beneficial business strategy.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Duke, T., Kao, C., & Vale, D. C. (2004, April). *Linking self-assessed English skills with the Test of English for International Communication (TOEIC)*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*, 395-430.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*, 43-62.

Hartnett, R. T., & Willingham, W. W. (1980). The criterion problem: What measure of success in graduate education? *Applied Psychological Measurement, 4*, 281-291.

Ito, T., Kawaguchi, K., & Ohta, R. (2005). *A study of the relationship between TOEIC scores and functional job performance: Self-assessment of foreign language proficiency* (TOEIC Research Rep. No. 1). Tokyo: Institute for International Business Communication.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students’ success. *Science, 315*, 1080.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly, 19*, 673-687.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*, 280-296.

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing,* 6, 1-13.

Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *The encyclopedia of language and education, Vol. 7, Language testing and assessment* (pp. 175-187). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Powers, D. E., Bravo, G., & Locke, M. (2007). *Relating scores on the Test de français international™ (TFI™) to language proficiency in French* (ETS Research Memorandum No. RM-07-04). Princeton, NJ: Educational Testing Service.

Powers, D. E., Bravo, G. M., Sinharay, S., Saldivia, L. E., Simpson, A. G., & Weng, V. Z. (2008). *Relating scores on the TOEIC Bridge™ to student perceptions of proficiency in English* (ETS Research Memorandum No. RM-08-02). Princeton, NJ: Educational Testing Service.

Powers, D. E., Kim, H.-J., Yu, F., Weng, V. Z., & VanWinkle, W. (2009). *The TOEIC speaking and writing tests: Relations to test-taker perceptions of proficiency in English* (ETS Research Rep. No. RR-09-18). Princeton, NJ: ETS.

Powers, D. E., Roever, C., Huff, K. L., & Trapani, C. S. (2003). *Validating LanguEdge Courseware scores against faculty ratings and student self-assessments* (ETS Research Rep. No. RR-03-11). Princeton, NJ: Educational Testing Service.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*, 1-20.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-based test (iBT): Exploration in a field trial sample* (ETS Research Rep. No. RR-08-09). Princeton, NJ: Educational Testing Service.

Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others of psychological assessment. *Psychological Bulletin, 90*, 322-351.

Tannenbaum, R. J., Rosenfeld, M., Breyer, J., & Wilson, K. M. (2007). *Linking TOEIC scores to self-assessments of English-language abilities: A study of score interpretation.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Upshur, J. (1975). Objective evaluation of oral proficiency in the ESOL classroom. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing 1967-1974* (pp. 53-65). Washington, DC: TESOL.

First Language of Examinees and Empirical Assessment of Fairness[1]

Sandip Sinharay, Longjuan Liang, and Neil J. Dorans

It is important to examine the effect of language proficiency on standard psychometric procedures used to ensure fairness. This paper describes an approach to examining how the results of two of these standard fairness procedures, differential item functioning (DIF) and score equating, are affected by inclusion or exclusion in the analysis sample of those who report that English is not their first language. Furthermore, the authors explore the sensitivity of the results of the two procedures to shifts in population composition. Data from a large-volume testing program are employed for illustrative purposes. The equating results were not affected by inclusion or exclusion of such examinees in the analysis sample, or by shifts in population composition. The effect on DIF results, however, varied across focal groups.

Insufficient language proficiency might interfere with test performance. If an examinee does not possess the degree of language proficiency needed to understand what a question is asking on a test that is not intended to measure language proficiency (such as a mathematics test), this may adversely affect fair and valid measurement of the construct of interest for that examinee. Hence, it is important to understand better the relationship of language proficiency to the two basic empirical procedures used to ensure test fairness: differential item functioning (DIF) and equating.

DIF procedures have been in place since the 1980s. Equating procedures have been in place for an even longer period. When equating procedures were initially implemented, the testing samples in the United States were mostly homogeneous with respect to their native language, English. As a consequence, equatings were not likely to be affected by the inclusion of non-native speakers in the testing sample. Hence, equatings were, and still are, for the most part performed using the full examinee sample. DIF results, in contrast, appear to have been sensitive to language proficiency. See Dorans and Kulick (1986), who described a study that used the self-reported *English First Language* (EL1)[2] and *Not English First Language* (NEL1) categorizations to explain what appeared to be many DIF items on the SAT Mathematics section for Asian Americans. This finding suggested that cleaner DIF results could be achieved if EL1 was used in addition to total score to provide a better matching variable in DIF analyses.

Research Question

The composition of the U.S. population has changed since the above-mentioned practices were adopted and is likely to keep changing. The report, *America's Perfect Storm*, by Kirsch, Braun, Yamamoto, and Sum (2007), noted, among other facts about a changing U.S. population, that immigration has accounted for an increasingly large fraction of U.S. population growth over the past few decades, and that the Hispanic share of the U.S. population is expected to grow from 14 percent in 2005 to slightly more than 20 percent by 2030. Hakuta and Beatty (2000) noted that between 1990 and 1997, the number of U.S. residents not born in the United States increased by 6 million, or 30%. In addition, to provide fair assessment and uphold standards on instruction for every child in the United States, both federal (e.g., No Child Left Behind Act of 2001) and state legislation now require the inclusion of all students, including English-language learners, in large-scale assessments (e.g., Abedi, 2002). Because of these changes, there has been an increase in the proportion of NEL1 examinees taking these large-scale assessments. Hence, there is a need to revisit the issue of choice of the examinee sample to be used in DIF and equating.

Ideally, for this purpose, we would like to identify the population of test takers who possess at least the level of proficiency

[1] This article is based on ETS Research Report Nos. RR-09-05 and RR-09-11. The original reports appear in the reference list of this article as Liang, Dorans, & Sinharay (2009) and Sinharay, Dorans, & Liang (2009), respectively. Both reports are available from the ETS ReSEARCHER database at http://search.ets.org/custres.

[2] Some of the works cited in this article, including the original ETS Research Reports upon which this article is based, abbreviate the variables English First Language and Not English First Language as EFL and NEFL, respectively. To avoid confusion with the term English as a Foreign Language, which language researchers also abbreviate as EFL, this article uses the term EL1 to stand for English First Language and NEL1 to stand for Not English First Language.

in English presumed to be necessary to provide a fair assessment of the construct of interest. Let us denote this population as *Sufficiently Proficient in English* (SPE). However, very few large-scale testing programs collect information on the examinees' English proficiency. Instead, most testing programs ask examinees whether English is one of their first languages or English is one of their best languages; that information is inadequate to classify an examinee as SPE. In this paper, we are limited to studying EL1 and NEL1 populations, instead of the SPE and *Not Sufficiently Proficient in English* (NSPE) populations that are of real interest.

The goal of this paper is to describe a set of methods that will provide information to a testing program about whether or not inclusion of NEL1 examinees in the analysis (a) affects the results of DIF analyses and equating given current proportions of NEL1 examinees, and (b) will affect the results of DIF analyses and equating if the proportion of NEL1 examinees in the test-taking sample increases in the future. This set of methods, which includes DIF analysis and score equity assessment (SEA) (Dorans, 2004), is applied to several data sets from large-scale assessments.

Our research aims to make a unique contribution regarding test fairness for English-language learners by focusing on two of the three facets of fairness discussed in Dorans (2008): DIF and SEA. DIF analysis examines fairness at the individual item level, while SEA examines fairness at the test score level. The third facet, differential prediction, examines whether test scores, in conjunction with other information such as high school grades, predict performance on an external criterion, such as grade in a college course, in the same way across different groups.

Description of the Data

We analyzed data from the PSAT/NMSQT® for illustrative purposes. The test has three sections: Critical Reading, Mathematics, and Writing. We focus only on the Mathematics section in this paper. The main reason for studying PSAT/NMSQT data is the presence of a significant number of NEL1 examinees among the hundreds of thousands of examinees taking the test. We report results from one recent administration of the PSAT/NMSQT: a Saturday administration. Those interested in more results should see Liang, Dorans, and Sinharay (2009) for equating results, or Sinharay, Dorans, and Liang (2009) for DIF results. Research has shown that examinee performance on mathematics items written in English can be affected by examinees' difficulty in reading the items (Abedi & Lord, 2001; Martiniello, 2008).

***Table 1*. Sample Demographics: Percentage of NEL1 Examinees in Each Group Studied for DIF Analysis**

Group	Percentage NEL1 Examinees
Combined[a]	7.8
Males	7.7
Females	7.9
Whites	1.6
Blacks	2.8
Asians	30.0
Hispanics	27.7

[a] The Combined group consists of all EL1 and NEL1 examinees, as well as test takers who did not answer the language background question.

The PSAT/NMSQT Mathematics section has 28 multiple-choice items and 10 student-produced response items (College Board, n.d.). A form of the test is equated to the mathematics section of an SAT® parent form using 22 common items. This equating is performed on the total sample (that is, sophomores and juniors) who took the PSAT/NMSQT form and the juniors and seniors who took the parent SAT form.

With the PSAT/NMSQT, the examinees are asked, during the examination, about the language they first learned to speak. They can answer (a) English, (b) English and another, or (c) Another. However, they need not answer the question. The PSAT/NMSQT operational DIF analysis is performed on sophomores and juniors who choose either the first option or second option to the question; in this paper, this subsample is used as the EL1 sample. The sophomores and juniors who choose the third option to the question are defined as the NEL1 sample. Both NEL1 and EL1 sophomores and juniors comprise the new form PSAT/NMSQT equating samples.

Table 1 shows the percentages of NEL1 examinees in each of the groups studied using DIF analysis (Combined, Males, Females, Whites, Blacks, Asians, and Hispanics). The Hispanic and Asian groups contain the largest proportion of NEL1 examinees. About 1 in 3 members of this group is a NEL1 examinee. These NEL1 examinees were excluded from the operational DIF analysis. For descriptive statistics about these data, see Sinharay et al. (2009).

Methods

DIF Analysis on the Observed Sample

We performed a DIF analysis on the EL1 subsample with either White examinees as a reference group or Male examinees as a reference group for any focal group with sufficient sample size. We then ran the same DIF analyses (Female/Male, Black/White, Asian/White, and Hispanic/White) on the Combined sample (the combination of EL1 and NEL1) as well. We then compared the two sets of DIF statistics (from Combined and from EL1-only) for Female/Male, Black/White, Asian/White, and Hispanic/White. We reported the results for the Mantel-Haenszel (MH D-DIF) statistic (e.g., Dorans & Holland, 1993).

To compare the two sets of DIF statistics, we used graphical plots and simple correlations. We also used the mean difference (MD), which is the difference between the arithmetic mean of the two sets of statistics and root mean squared difference (RMSD). Suppose the first set of DIF statistics are $X_i, i=1,2,...I$, and the second set of DIF statistics are $Y_i, i=1,2,...I$. The mean difference is defined as

$$MD = \frac{1}{I}\sum_i X_i - \frac{1}{I}\sum_i Y_i$$

and the RMSD is defined as

$$RMSD = \sqrt{\frac{1}{I}\sum_i (X_i - Y_i)^2}\ .$$

If there is no difference between the DIF statistics from the EL1 and Combined samples, this would suggest that the testing program can use the Combined sample to perform all of its DIF analyses instead of using the EL1 examinees only; that will increase the statistical power of the DIF analyses. A difference would indicate that the testing program should continue the practice of performing DIF analyses on the EL1 examinees only.

DIF Analyses on the Synthetic Subsamples

The above analyses provided DIF results for the currently observed percentage of NEL1 examinees in the test-taking sample, but did not reveal how DIF results would change if the percentage of NEL1 examinees in the test-taking sample changed to a higher value — say, 20%. It is important to know how DIF results would be affected if the proportion of NEL1 examinees increased in the examinee sample, especially in the light of the above-mentioned findings of Hakuta and Beatty (2000) and Kirsch et al. (2007). Hence, to study the DIF results for percentages of NEL1 examinees higher than that currently observed, we create synthetic subsamples from the observed data by combining the NEL1 group with a simple random sample from the EL1 group. This allowed us to study the effects of specific percentages of NEL1 examinees on DIF results. For example, consider the Saturday form of the PSAT/NMSQT that is taken by 40,160 NEL1 examinees and 474,735 EL1 examinees. The percentage of NEL1 examinees among all examinees was 7.8%[3]. However, if we were to draw a random sample of 40,160 examinees from the 474,735 EL1 examinees, and then combine them with the 40,160 NEL1 examinees, it would yield a synthetic subsample with 50%[4] NEL1 examinees. We created synthetic subsamples with proportions of NEL1 examinees in the subsamples ranging from .1 to .9 in increments of .1, and performed DIF analyses on these synthetic subsamples. Such analyses allow us to study the sensitivity of the DIF results to the percentage of NEL1 examinees in the samples.

Equating on the Observed Sample

For each set of test data, we performed equating on three examinee samples: EL1, NEL1, and total (EL1 + NEL1 + Those who did not provide an answer to the language question = Combined group + Those who did not answer the language question). Note that operationally, most testing programs perform equating on the total sample. We then compared the three sets of results by comparing the EL1 equating sample to the total equating sample, and the NEL1 equating sample to the total equating sample. This is similar to an SEA, which examines subpopulation invariance of equating functions to assess fairness at the test score level (Dorans, 2004). As noted by Dorans, DIF examines whether items function in the same way across different subpopulations by checking that their relationship to total test performance is the same. In contrast, SEA examines whether scores on tests built to the same set of specifications can be related to each other in the same way across different subpopulations. If these scores are not related in the same way across different subpopulations, then different versions of the test cannot be viewed as being interchangeable.

The tests were equated using the non-equivalent groups with anchor test (NEAT) design. We employed the chained equipercentile method to equate the new form to the old form for each test, once with the EL1 subsample (that is, including in the equating sample only the EL1s for both the current and old form) and once with the NEL1 subsample. We also performed the equating using the current and old form total groups. We

[3] 100x40,160/(40,160+474,735) = 7.8%

[4] 100x40,160/(40,160+40,160) = 50%

then computed the differences of the equating functions based on the subsamples and on the total sample.

To compare any two conversions, the Root Expected Square Difference (RESD) index (Dorans & Liu, 2009),

$$RESD_g = \sqrt{\sum_m f_{gm}\left[s_g(m) - s_P(m)\right]^2}$$

was used, where *g* represents the subsample, which is EL1 or NEL1 in this study, *P* represents the total sample, and *m* represents the score level. The quantities *sg(m)* and *sP(m)* represent the scale score of subsample *g* and the scale score of the total sample at score level *m*. The weight *fgm* is the relative frequency of subsample *g* at score level *m* and is used so that scale scores with higher frequency receive larger weights.

Equating on the Synthetic Subsamples

As with our DIF analyses, synthetic subsamples were created to study the equating results for certain specific percentages of NEL1 examinees. Simple random samples of several sizes were drawn from the EL1 examinees and combined with all of the NEL1 examinees to create synthetic subsamples with varying proportions of NEL1 examinees. The equating functions for the two subgroups and the total group were then compared for each of the synthetic subsamples.

Table 2 shows the proportion of NEL1 students in each synthetic subsample. In each of Subsamples 1 to 8, we attempted to simulate a gradual increase in the size of the NEL1 subpopulation from the administration of the old form to the new form: The NEL1 proportion for the new form is chosen to be 5% higher than that of the old form in each pair of synthetic subsamples to simulate a consistent but gradual growth in the NEL1 proportions in the total group in the time gap between the administration of the old and new forms. For comparison, Subsample 9 represents a much more dramatic difference — a new form sample of 50% EL1 and 50% NEL1 and an old form sample of 90% EL1. The nine synthetic subsamples allowed us to study the sensitivity of the equating results to the percentage of NEL1 examinees.

Results

DIF Results for the Observed Sample

Table 3 shows the values of the correlation, RMSD, and MD for the MH D-DIF statistic for the four focal/reference group combinations for the Saturday form.

Table 2. **Proportions of NEL1 Test Takers in Synthetic Subsamples**

Synthetic Subsample No.	Proportion for Old Form	Proportion for New Form
1	.10	.15
2	.15	.20
3	.20	.25
4	.25	.30
5	.30	.35
6	.35	.40
7	.40	.45
8	.45	.50
9	.10	.50

Table 3. **Correlation, RMSD, and MD for the MH D-DIF Statistic for EL1 and Combined Samples**[a]

DIF Analysis	Correlation	RMSD	MD
Female/Male	1.000	0.014	0.002
Black/White	1.000	0.011	0.001
Hispanic/White	.990	0.060	-0.004
Asian/White	.995	0.102	0.004

[a] Data are from the Saturday form of the PSAT/NMSQT Mathematics Test.

The table shows that the association between the DIF statistics obtained in the EL1 group and those obtained in the Combined group is quite strong, which means that the ordering of items with respect to DIF is the same regardless of whether the DIF analysis is performed on the EL1 group or on the Combined group.

DIF Results for Synthetic Subsamples

Figure 1 (p. 14) shows how a change in the proportion of NEL1 examinees affects the MH D-DIF statistic for items on the Saturday form. The figure has two panels — the top panel shows the results for correlation and the bottom panel shows the results for RMSD. Both panels show, for proportions of NEL1 examinees ranging from .1 to .9, the values of the correlation or RMSD showing association between the DIF statistics for the Combined sample versus those for the EL1 subsample for Female/Male DIF, Black/White DIF, Asian/White DIF, and Hispanic/White DIF.

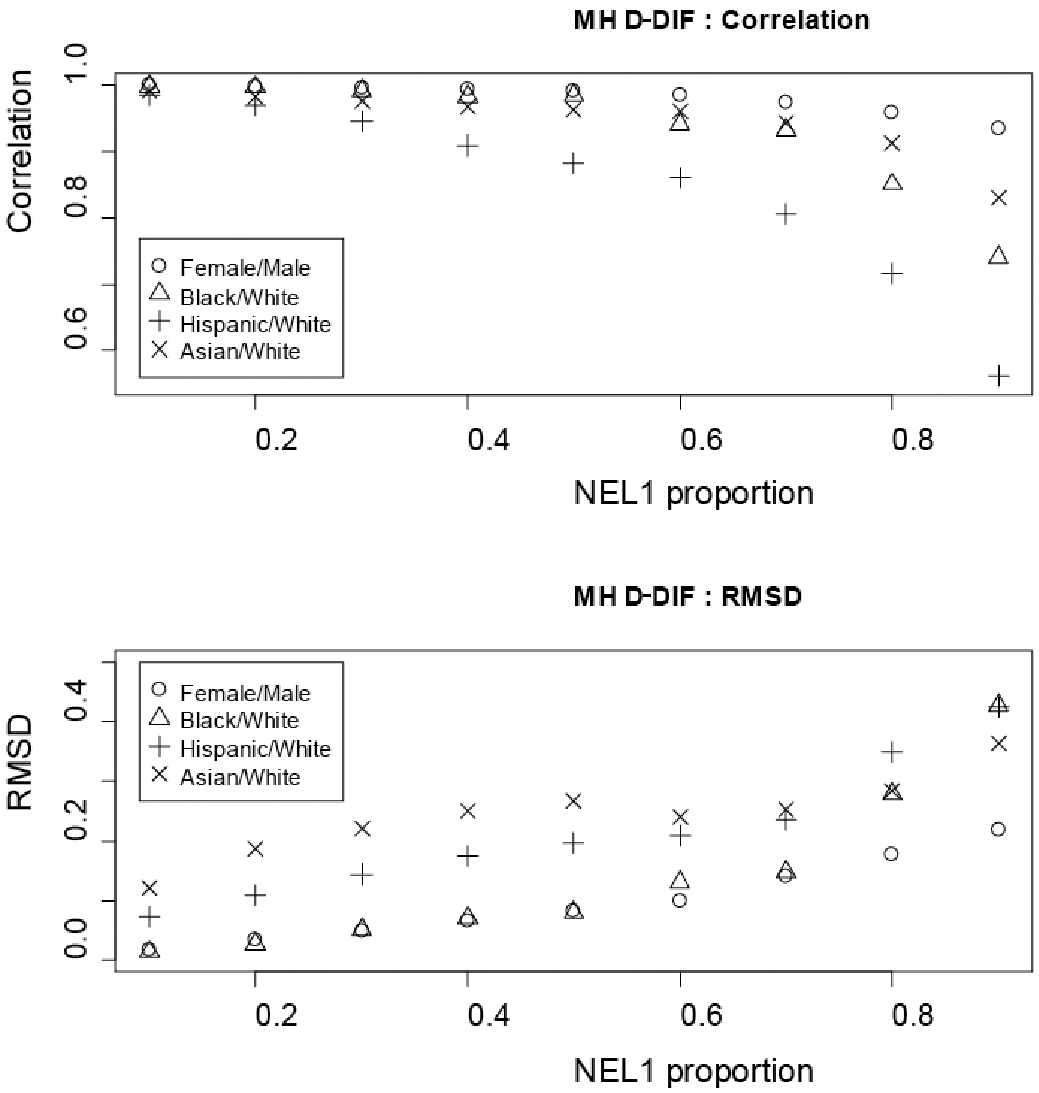


Figure 1*. Effect of Change in NEL1 Proportion on DIF Results for the Synthetic Subsamples for the Saturday form: Female/Male, Black/White, Asian/White, and Hispanic/White*.

Examination of Figure 1 leads to the following conclusions:

- The association between the MH D-DIF statistics from the Combined and EL1 groups becomes weaker (correlation decreases and RMSD increases) as the NEL1 proportion increases, which is expected because an increase in the NEL1 proportion causes the Combined sample to be increasingly different from the EL1 subsample.
- Among the four focal/reference groups, the relationship between the DIF statistics from the Combined and EL1 pair is the strongest for Female/Male DIF statistics, and the weakest for Hispanic/White DIF statistics. For example, the correlations are all close to 1 for Female/Male DIF statistics, while the correlation may be as low as .6 for Hispanic/White DIF statistics. This finding indicates that the Female/Male DIF statistics are basically the same in the EL1 and Combined groups, but that the same statistics for the other groups are not.
- Regarding the practical question of whether the DIF analysis should be performed on the Combined group or on the EL1 group, Figure 1 shows that the DIF results for the Combined and EL1 groups are essentially the same across subgroups for the proportion of NEL1 currently seen in actual PSAT/NMSQT data (about 7.8%). However, as the

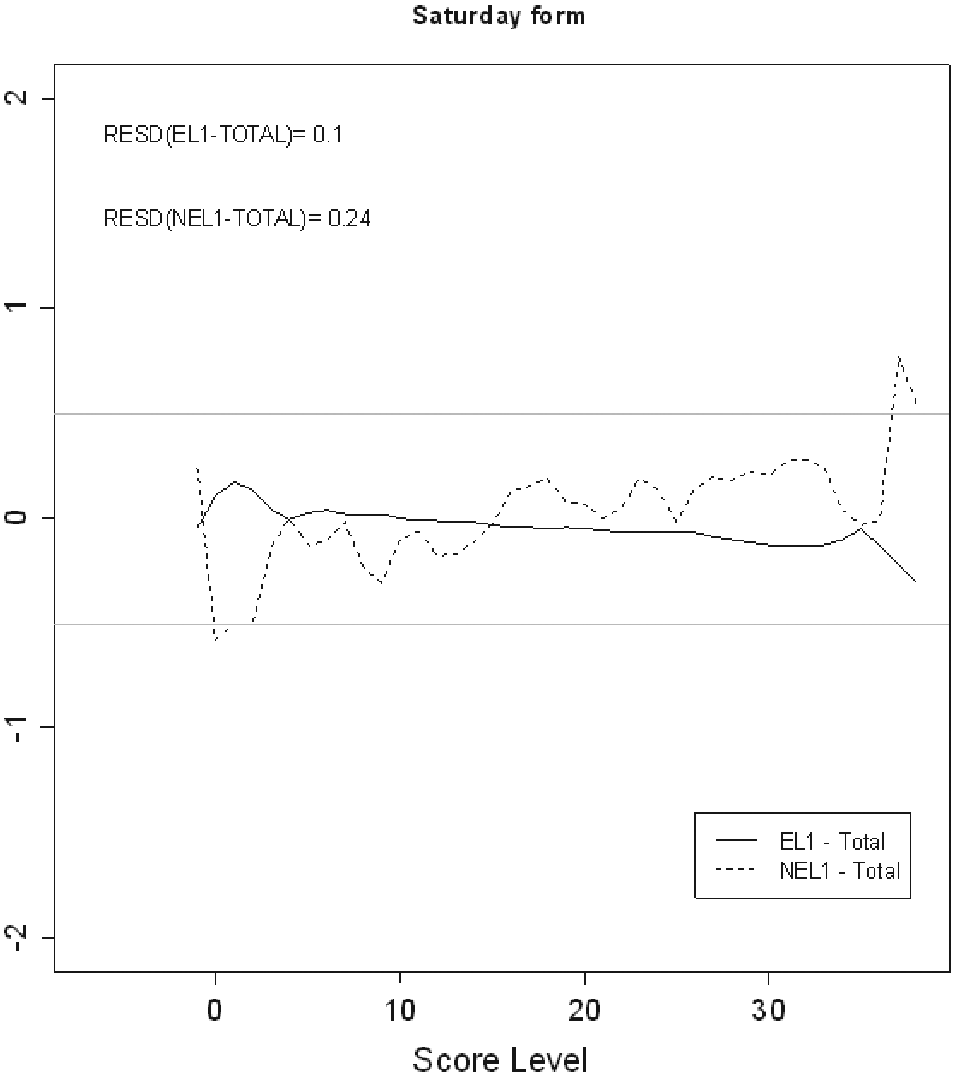


***Figure 2*. Differences in the equated scaled scores for the EL1 and NEL1 subsamples, and the total sample for the Saturday form.**

proportion of NEL1 examinees increases, for example, beyond 0.4, the DIF statistics for the Combined and EL1 groups differ somewhat for ethnic/racial groups, but do not differ for Female/Male groups.

Equating[5] Results for the Observed Sample

Operationally, unsmoothed chained equipercentile method (with an internal anchor test) is used to link the PSAT/NMSQT to the SAT using the total sample, regardless of the examinees' response to the first language question. We employed the same equating design and method to link the PSAT/NMSQT scores to the SAT scores, but, in addition to performing the linking on the total sample, we also performed the linking on the EL1 subsample (where we use only the EL1 examinees for the PSAT/NMSQT form and the SAT form to which the PSAT/NMSQT is linked) and on the NEL1 subsample (where we use only the NEL1 examinees for the PSAT/NMSQT form and the SAT form to which the PSAT/NMSQT is linked).

Figure 2 shows the differences in the raw-to-scale conversions for the total sample and the EL1 and NEL1 subsamples for the Saturday form. The two horizontal lines shown in each panel of Figure 2 correspond to the Minimal Difference That Might Matter criterion (Dorans & Liu, 2009), or DTM for short. The DTM is considered to be half of an unrounded report score

[5] Strictly speaking, new PSAT/NMSQT forms are not equated to each other. Instead they are linked via a design in which each PSAT/NMSQT form has been linked via what Holland and Dorans (2006) call calibration to one of two longer, more reliable SAT forms that are directly equated to each other. We use the term equating loosely here.

Table 4. **Effect of the Choice of the Equating Sample on the Summary Statistics of the Equated Scores for the Saturday Form**

Group	N	Linking	Mean	SD	Mean Diff	RESD	% ESS |DIFF| ≥ 0.5[b]	% Examinees |DIFF| ≥ 0.5
Total	154,371	TGL	51.49	10.16				
EL1	144,570	TGL	51.40	9.99	-0.05	0.10	0%	0%
		SGL	51.35	9.94				
NEL1	7,963	TGL	53.94	12.41	0.08	0.24	10.0%	7.6%

[b] % ESS |DIFF| ≥ 0.5 represents the percentage of absolute differences in equated scaled scores that are greater than or equal to the DTM.

unit. Since the reporting scale of the PSAT/NMSQT is 20 – 80 with a 1-point increment, the DTM is 0.5 in this study. Any difference with absolute value less than the DTM is negligible for most practical purposes because it is smaller than the rounding effect observed for reported scores.

Figure 2 shows that the conversion for the EL1 subsample is not much different from that for the total sample. The absolute differences of the two conversions are always less than the DTM. From an operational point of view, use of the total sample or the EL1 subsample should yield the same results, except for rounding. This small difference is partly due to a large proportion of the EL1 examinees in the total sample.

The differences in the conversions based on the NEL1 subsample and on the total sample are also plotted for informational purposes. Only a few of these differences, all for extremely high or low score-points, are larger than the DTM.

Table 4 shows, for the Saturday form, the extent to which the summary statistics of the equated scores of the EL1 and NEL1 subsamples are affected by whether the total group conversion or subgroup-specific conversion is used to produce scale scores. In the table, TGL represents results based on the total group conversion and SGL represents results based on the subgroup conversion. For example, the second and third rows of numbers of the table indicate that for the Saturday form, if the equating of PSAT/NMSQT to SAT is performed on the total group, the mean and SD of the equated scores of the EL1 examinees are 51.40 and 9.99, respectively, but if the equating of PSAT/NMSQT to SAT is performed on the EL1 group only, the mean and SD of the equated scores of the EL1 examinees are 51.35 and 9.94, respectively. Mean Diff is the difference in means for a subgroup when the SGL or TGL is applied. The table also shows the percentage of equated scale score differences that are greater than or equal to the DTM, % ESS |DIFF| ≥ 0.5, and the percentage of examinees with equated scale score differences that are greater than or equal to the DTM, % Examinees |DIFF| ≥ 0.5.

The table demonstrates that there is hardly any difference between the results from the total group conversion and the subgroup conversion in the EL1 subsample. The mean difference for the EL1 subsample is -0.05. The RESD for the EL1 subsample is 0.10, which is much smaller than the DTM. However, the mean will go up slightly if, instead of total group conversion, subgroup conversion is used for the NEL1 subsample. The mean difference (SGL-TGL) is 0.08. The RESD of 0.24 is still smaller than .5 score units. The percentage of examinees with equated scale score differences that are greater than or equal to .5 score units is only 7.6% for NEL1 when subgroup conversion is based on the NEL1 examinees, demonstrating that only a small proportion of the NEL1 examinees will be affected if subgroup conversion is used instead of total group conversion.

Equating Results for Synthetic Subsamples

Figure 3 plots the RESD of the conversions based on the EL1/NEL1 subsample and the total sample against the NEL1 proportion for the two forms.

The figure shows that in general, as the proportion of NEL1 examinees increases, the differences in the conversions based on the EL1 subsample and on the total sample become more and more noticeable and the RESDs increase. In contrast, the differences in the conversions based on the NEL1 subsample and on the total sample become smaller as the NEL1 proportion increases. The RESD values, however, do not exceed the DTM of 0.5, even for the ninth synthetic subsample that represents an extreme difference between the NEL1 proportions in the old and new forms.

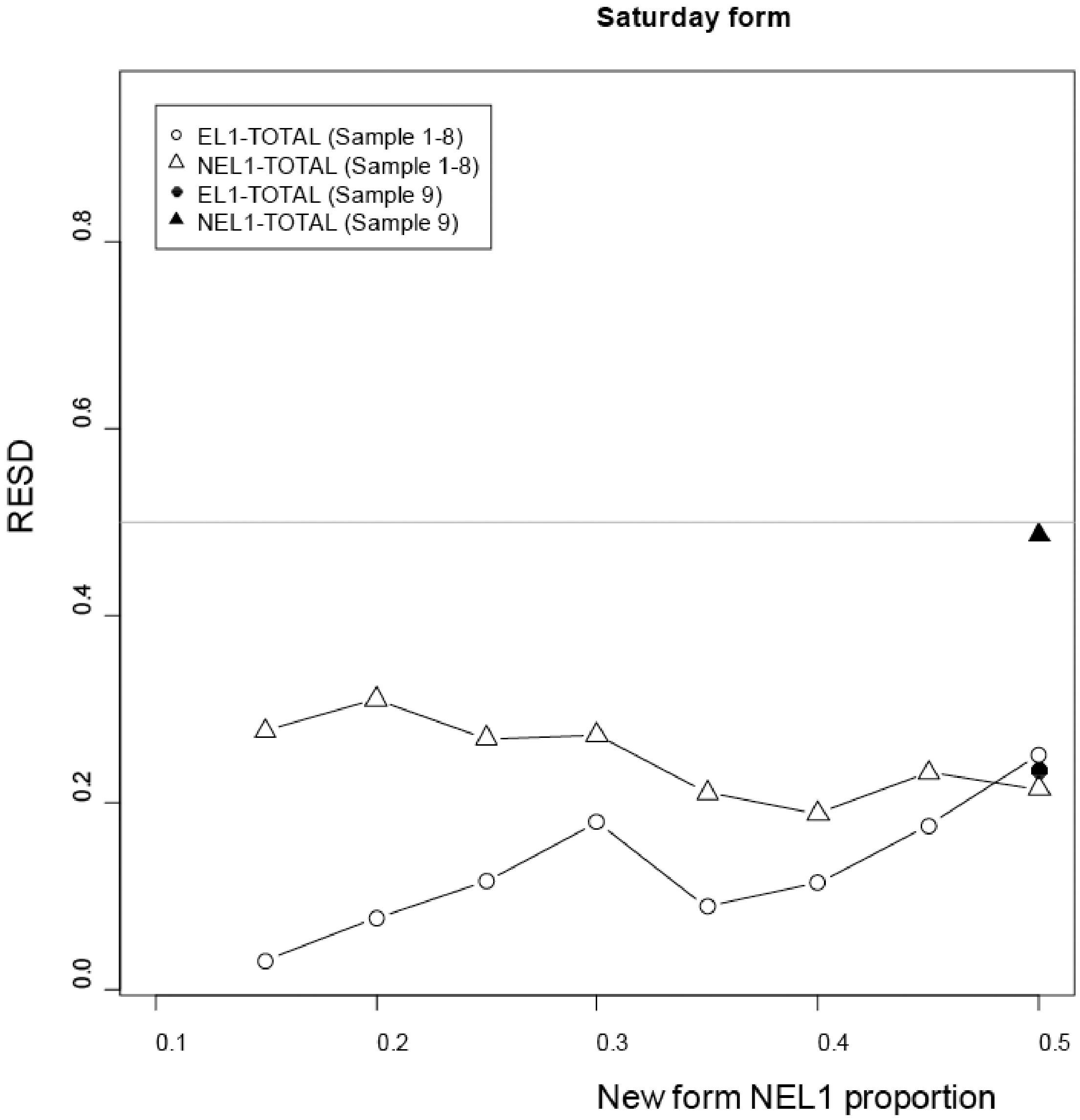


Figure 3. **RESD for synthetic subsamples with varying NEL1 proportions for the Saturday form***.*

Discussion and Conclusions

We used data from one form of the PSAT/NMSQT Mathematics examination to illustrate our suggested method of examining the sensitivity of DIF and equating procedures to the proportion of NEL1 examinees in the analysis sample.

For these data, DIF results across subgroups do not seem to be affected much by the examinee's first language status when the proportion of NEL1 examinees is around the proportion currently observed in actual data (about 7.8% for the form of PSAT/NMSQT used here); that is, the DIF results are virtually the same irrespective of whether it is performed on the EL1 sample or the Combined sample. However, as the proportion of NEL1 examinees increased, for example, beyond 0.4, the DIF statistics for ethnic/racial DIF exhibited some sensitivity to this increase. These results suggest that for PSAT/NMSQT, it does not matter now whether the DIF analysis is performed on the EL1 group or on the Combined group, but it may matter in the future if the proportion of NEL1 examinees taking the PSAT/NMSQT increases. Hence we recommend monitoring the proportion of NEL1 examinees taking the PSAT/NMSQT and periodically performing the analyses suggested here.

Equating results from this study show that, from an operational point of view, when averaged across all score levels, it does not matter whether the equating is based on the EL1 subsample or on the total sample. Furthermore, the equating results are hardly affected by an increase in the NEL1 proportion. This should be good news to the testing program.

The data used for illustrative purposes here have limitations. First, the criterion used in this study to categorize examinees as EL1 or NEL1 was the examinees' self-report on a question asking if English was their first language. This question is printed on the test form and is ready-to-use information. However, it is not an accurate measure of examinees' true English proficiency, which is the construct that is likely to affect DIF and equating analyses. Future investigations of the sensitivity of DIF procedures to language proficiency should include a more direct measure of English proficiency.

Second, the magnitude of DIF in these data is very small. While this demonstrates the high quality of the PSAT/NMSQT items (they are in a sense double pretested, once prior to their use on the SAT and once as SAT items prior to their use with PSAT/NMSQT), it is difficult to study effects on DIF.

Third, the PSAT/NMSQT is one of the few tests that have a large number of NEL1 examinees, but it is only one test that was studied.

Finally, the forecasts we made about increases in the size of language proficiency were simplistic and not meant to predict what might actually happen.

Having stated the limitations of these illustrative data, it is nonetheless important to examine the effects of language proficiency on the outcome of fairness procedures. The process we employed merits application wherever feasible, as do other processes that investigate the sensitivity of methods to shifts in the composition of the population.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8*, 231-257.

Abedi, J., & Lord, C. (2001). The language factors in mathematics tests. *Applied Measurement in Education, 14*, 219-234.

College Board (n.d.). *PSAT/NMSQT®: What's on the test?* Retrieved March 1, 2010, from http://www.collegeboard.com/student/testing/psat/about/ontest.html

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43-48.

Dorans, N. J. (2008, March). *Three facets of fairness.* Paper presented at the annual Meeting of the National Council on Measurement in Education, New York.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Dorans, N. J., & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT Mathematics test data across several administrations* (ETS Research Report No. RR-09-08). Princeton, NJ: Educational Testing Service.

Hakuta, K., & Beatty, A. (Eds.). (2000). *Testing English-language learners in U.S. schools.* Washington, DC: National Academy Press.

Holland, P. W., & Dorans, N. J. (2009). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: American Council on Education/ Praeger

Kirsch, I., Braun, H., Yamamoto, K., & Sum, A. (2007). *America's perfect storm: Three forces changing our nation's future*. Princeton, NJ: Educational Testing Service.

Liang, L., Dorans, N. J., & Sinharay, S. (2009). *First language of examinees and its relationship to equating* (ETS Research Report No. RR-09-05). Princeton, NJ: Educational Testing Service.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*, 333-368.

Sinharay, S., Dorans, N. J., & Liang, L. (2009). *First language of examinees and its relationship to differential item functioning* (ETS Research Report No. RR-09-11). Princeton, NJ: Educational Testing Service.

Highlights from the Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL) Project in Mathematics[1]

Edith Aurora Graf, Karen Harris, Elizabeth Marquez, James H. Fife, and Margaret Redman

This article describes the early design and development stages of the mathematics strand of the Cognitively Based Assessment of, for, *and* as *Learning (CBAL) project. The project aims to create a model that synergistically unifies three components: accountability assessment, formative assessment, and professional support. Bennett and Gitomer (2009) describe the rationale for such a model. Prototype CBAL mathematics assessments were developed according to an evidence-centered design approach (e.g., Mislevy, Steinberg, & Almond, 2003). The article describes a model for mathematics competency in Grades 6–8 and includes an excerpt from the evidence model and sample tasks. The competency model's design is based on research in cognitive psychology and mathematics education and on national and state standards. A review of the underlying research is available from Graf (2009).*

This article provides an overview of the early stages of design and development for the mathematics strand of the Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL) project. The CBAL project is intended to provide an innovative model of how states might measure student achievement in ways that both satisfy the needs of policy makers and support classroom instruction. The project is in progress, and the assessment framework, task specifications, and tasks continue to evolve as the work matures. Many of the materials presented in Graf, Harris, Marquez, Fife, and Redman (2009) have changed since that report's release, so we provide updated versions of these materials here.

Multiple prototype CBAL assessments for reading, writing, and mathematics have been created for use at the middle-school level following the evidence-centered design approach developed by Mislevy and colleagues (e.g., see Mislevy, Steinberg, & Almond, 2003). Within each content area, a *competency model* specifies the relationships among the target concepts and skills. The competency models were informed by research in cognitive psychology and education, as well as by state and national standards. The CBAL system includes three components: accountability assessment, formative assessment, and professional support. In order to facilitate conceptual coherence among components within a domain, all three are based on the same competency model. As discussed in Bennett and Gitomer (2009), CBAL is intended to extend traditional assessment systems by measuring aspects of student competency important for educational policy making (assessment *of* learning), providing information that may be used to adapt instruction (assessment *for* learning), and offering a worthwhile educational experience in and of itself (assessment *as* learning).

There are similarities in design among the three content areas of CBAL, but there are also some important domain-specific differences. The first step in the CBAL project was to review research for the purpose of characterizing competency in each of the three domains. For reviews of the supporting research in each of the three domains, the reader is referred to O'Reilly and Sheehan (2009) for reading, to Deane et al. (2008) for writing, and to Graf (2009) for mathematics.

A Model of Mathematics Competency

The *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 2000) addresses both *content* and *process* strands of mathematics in prekindergarten through Grade 12. Consistent with this framework, CBAL Mathematics is intended to assess both content and process aspects of mathematical competency. The initial focus of CBAL Mathematics is at the middle school level (Grades 6 through 8). It is frequently pointed out that most mathematics curricula address far too many topics in too little depth. In an attempt to respond to this problem, NCTM released the *Curriculum Focal Points for Prekindergarten*

[1] This article is based on ETS Research Memorandum No. RM-09-07. The original report appears in the reference list of this article as Graf, Harris, Marquez, Fife, & Redman (2009). To request a copy of the full report, send an e-mail to RDweb@ets.org. The article is also related to ETS Research Report No. RR-09-42 (Graf, 2009), which is available from the ETS ReSEARCHER database at http://search.ets.org/custres.

Through Grade 8 Mathematics (NCTM, 2006), a streamlined specification of mathematics content that is essential to cover in prekindergarten through Grade 8.

Figure 1, on p. 21, shows a recent version of the competency model for CBAL Mathematics. The lower half of the model shows content competencies (represented by constituents of the high-level competency Understand and Use Content-Specific Procedures and Language), while the upper half of the model shows process competencies (represented by constituents of the high-level competency Use Cross-Cutting Mathematical Processes). The main content areas include numbers and operations, algebra, geometry, measurement, and data analysis and probability. These content areas correspond to the content strands used in *Principles and Standards* (NCTM, 2000) as well as those used in most state standards. All of these areas are essential to any middle school mathematics curriculum, though different areas may receive more or less focus at different levels.

Consistent with recent recommendations, an emphasis on numbers and operations, algebra, and the connections between them is supported at this level. As discussed in *Adding it Up: Helping Children Learn Mathematics* (Kilpatrick, Swafford, & Findell, 2001), the study of numbers and operations is foundational to other topics in mathematics. If students have difficulty with numbers and operations, they are also likely to have difficulty with algebra, measurement, geometry, and data analysis and probability. Reports by both the RAND mathematics study panel (RAND Mathematics Study Panel & Ball, 2003) and the National Mathematics Advisory Panel (U.S. Department of Education, 2008) made recommendations for how to support mathematics instruction and focused on algebra in particular. Research also suggests that students tend to regard numbers and operations and algebra as disconnected branches of mathematics (e.g., see Borchert, 2003, and Lee & Wheeler, 1989). Since this perception is not easily altered, long-term instruction that encourages students to develop the connections between them is advisable.

Because mathematics is an interconnected discipline, it is possible to focus on foundational skills without excluding any important content. For example, although the National Mathematics Advisory Panel focused on algebra, the report included two critical foundation benchmarks for whole number fluency, six benchmarks for fraction fluency, and three benchmarks for geometry and measurement. While these are not algebraic topics, they are essential to the development of algebraic understanding. Vennebush, Marquez, and Larsen (2005) described examples of tasks associated with other content areas that lend themselves to algebraic approaches. They also illustrated how tasks in other content areas can be revised to encourage algebraic reasoning to a greater extent.

The lower half of the model was also heavily influenced by *Curriculum Focal Points for Prekindergarten Through Grade 8 Mathematics* (NCTM, 2006); several of the nodes in this portion of the model are named for focal point titles. For example, "Developing an understanding of and using formulas to determine surface areas and volumes of three-dimensional shapes" is the title of a focal point for Grade 7 (NCTM, 2006, p. 19) and "Analyzing two- and three-dimensional space and figures by using distance and angle" is a focal point for Grade 8 (NCTM, 2006, p. 20).

As Figure 1 shows, modeling, representation, and argument are all considered important process competencies in mathematics assessment. Modeling involves using mathematics to solve real-world problems, and as such is a central competency assessed by many CBAL Mathematics tasks. Lesh and Lamon (1992, Figure 2, p. 22) emphasized two components of mathematical modeling in particular: interpretation and prediction. Interpretation involves characterizing a situation in mathematical terms, as a model. The model is usually a simplification of the real-life situation it represents, and its development necessarily involves some assumptions about which aspects of the situation are important to capture. Decisions regarding the assumptions behind a model are an extremely important part of the interpretation process. Prediction involves applying a mathematical model back to the real-world situation. In the competency model, the term Apply Models is used instead of Make Predictions, since prediction suggests that use of a model is limited to predicting a future event, while the intention here is to define a competency that is somewhat more general.

Modeling and representation are highly related skills; in an earlier version of the competency model, they were combined in a single competency. Even at that stage, however, it was recognized that modeling and representation are not identical: A model is a mathematical characterization, whereas a representation is a medium that can be used to convey such a characterization. As an example, a situation may be modeled as a linear function, which can be represented in a table, in a graph, or as an equation. The main distinction between the modeling and representation competencies is that the modeling competency is focused on the process of developing, applying, and revising a *mathematical characterization* of events, while the representation competency is focused on depictions of mathematical relationships, which may or may not be models of real events. Representation is one of the process strands in the *Principles and Standards* (NCTM, 2000), where it is suggested that using alternate representations supports learning because different representations provide different windows for understanding a particular concept. It is also suggested that different representa-

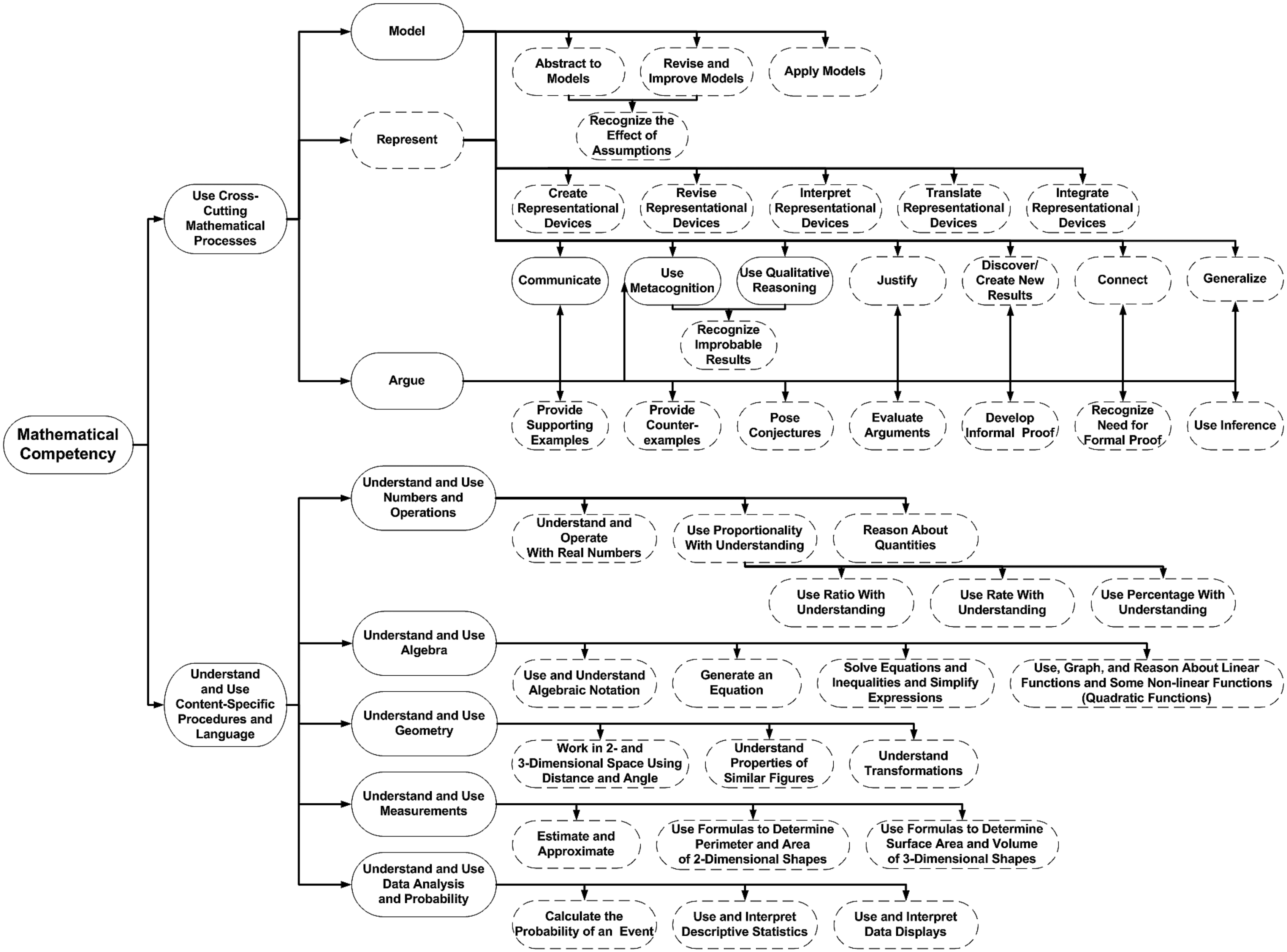

Mathematical Competency
Use Cross-Cutting Mathematical Processes
Understand and Use Content-Specific Procedures and Language
Model
Represent
Argue
Abstract to Models
Revise and Improve Models
Apply Models
Recognize the Effect of Assumptions
Create Representational Devices
Revise Representational Devices
Interpret Representational Devices
Translate Representational Devices
Integrate Representational Devices
Communicate
Use Metacognition
Use Qualitative Reasoning
Recognize Improbable Results
Justify
Discover/ Create New Results
Connect
Generalize
Provide Supporting Examples
Provide Counter-examples
Pose Conjectures
Evaluate Arguments
Develop Informal Proof
Recognize Need for Formal Proof
Use Inference
Understand and Use Numbers and Operations
Understand and Operate With Real Numbers
Use Proportionality With Understanding
Reason About Quantities
Use Ratio With Understanding
Use Rate With Understanding
Use Percentage With Understanding
Understand and Use Algebra
Use and Understand Algebraic Notation
Generate an Equation
Solve Equations and Inequalities and Simplify Expressions
Use, Graph, and Reason About Linear Functions and Some Non-linear Functions (Quadratic Functions)
Understand and Use Geometry
Work in 2- and 3-Dimensional Space Using Distance and Angle
Understand Properties of Similar Figures
Understand Transformations
Understand and Use Measurements
Estimate and Approximate
Use Formulas to Determine Perimeter and Area of 2-Dimensional Shapes
Use Formulas to Determine Surface Area and Volume of 3-Dimensional Shapes
Understand and Use Data Analysis and Probability
Calculate the Probability of an Event
Use and Interpret Descriptive Statistics
Use and Interpret Data Displays

Figure 1. A recent version of the competency model used in the mathematics strand of the Cognitively Based Assessment *of, for,* and *as* Learning (CBAL) project.

Competency	Abstract to Models (e.g., see Lesh & Lamon, 1992)
Definition	Develop a mathematical model by: (1) identifying features and relationships of a real-world situation that are relevant to a particular goal, and (2) characterizing those features and relationships in mathematical terms. During stage (1), the student disentangles aspects of the situation that are important to capture from those that can be simplified or ignored. Because any mathematical model is necessarily a simplification, the abstraction process should include an account of the model's assumptions and limitations.
Response Type	Mathematical statement
Scorable Features of Work Product	Correctness; whether or not the statement is in the correct form; whether or not the statement is in the requested form; whether or not the statement is close to correct (e.g., sign error or other very common and recognizable error); whether or not the statement captures features and relationships consistent with the proposed model.
Level 1	When the student produces a mathematical statement that represents a situation, the statement is either incorrect or clearly of the wrong type (e.g., the student provides a numeric response when asked for an expression or equation).
Level 2	When representing a situation as a mathematical statement, the student usually produces a statement of the correct type, and it usually includes the correct variables, though it may not be correct (e.g., he or she may commit the variable reversal error, as described in Clement, Lochhead, & Monk, 1981, and Clement, Lochhead, & Soloway, 1979). More generally, the statement may not capture the proposed model.
Level 3	When representing a situation as a mathematical statement, the student almost never generates a statement in incorrect form and usually generates a simple statement (involving no more than one variable and two operations) correctly. The statement is consistent with the proposed model.
Level 4	When representing a situation as a mathematical statement, the student will consistently generate a correct mathematical statement in the correct form. The statement is consistent with the proposed model. In generating the statement, the student attends to subtleties, such as the domain to which the statement applies, and notes exceptions.

***Figure 2*. Example of an evidence model for the Abstract to Models competency.**

tions are better suited to different purposes, and in addition to developing facility with translating among alternate representations, students should learn to recognize which representations are best applied to a particular purpose.

Iterative revision is important both to modeling and to representation. Typically, mathematical models are revised as certain assumptions are found to be unrealistic or as the model is found to be in error. Similarly, one may experiment with several different representations before selecting one that is best for solving a problem or emphasizing a particular pattern. The process of iterative revision should be encouraged from an early stage; when guided by a teacher, even young children can improve data models and representations through classroom discussion (Lehrer & Schauble, 2000).

The term argument is used broadly to refer to informal arguments and descriptions of solutions as well as more formal presentations of reasoning and proof. Ideally, the development of a mathematical argument is an exploratory enterprise (e.g., see Schoenfeld, 1994). Students should try out examples, pose conjectures, and be willing to pursue several approaches before finding the path to solution. Generalizing and formalizing an argument usually comes at a later stage in the argument's development. In Grades 6 through 8, students should use a variety of examples to formulate conjectures. Michener (1978) distinguished among several types of examples and their role in instruction. In addition to discussing the importance of different types of supporting examples, she also discussed the importance of counterexamples to teaching and learning. When students are formulating conjectures, supporting examples can suggest a pattern. A single counterexample, however, refutes the conjecture, leading students to reformulate it or perhaps to abandon it entirely. Understanding the roles of both types of examples is important in the development of mathematical arguments. At this stage, students may also be able to construct informal proofs. For example, Bastable and Schifter (as cited in Kaput, 1999) found that third-graders were able to show that whole numbers are commutative under multiplication by rotating an m by n array 90 degrees to produce an n by m array.

An Evidence Model for CBAL Mathematics

The CBAL Mathematics evidence model specifies how student responses are used to interpret performance with respect to each target competency. CBAL Mathematics tasks use a variety of response types, including those in which the student is expected to provide numbers (either a single number or a set of numbers), mathematical statements (either a single statement or a set of statements, as in a derivation), text, graphs, or tables, or to make selections. Even for a particular competency, the nature of the provided evidence is expected to vary widely across response types. For this reason, the model characterizes evidence for each competency by response type combination. Figure 2, on p. 22, shows an example of such a characterization.

Although the design of the CBAL Mathematics evidence model is informed by research, the specification and ordering of the levels are still under revision. Future versions of the evidence model will draw from research on *developmental models* (Harris, Bauer, & Redman, 2008), sometimes referred to as *learning progressions*, in a variety of mathematical content areas. Empirical validation of the evidence model should examine several questions. One question concerns whether the level descriptors are sufficient to characterize the breadth of observed student performance. It is possible that student responses will suggest the need for (a) additional levels and/or (b) additions to the current level descriptors. Another question concerns the extent to which the level descriptors are necessary; it is possible that student responses will suggest that (a) some levels should be eliminated and/or (b) some aspects of the current descriptors are never or rarely in evidence. Also, it will need to be evaluated whether the level descriptors are correctly ordered and whether they can be reliably assigned based on students' responses.

Excerpts from Sample Tasks Developed for CBAL Mathematics

In this section, excerpts from CBAL Mathematics tasks are briefly discussed to provide a sense of the nature of the tasks. For thorough descriptions, see the full report on which this article is based (Graf, Harris, Marquez, Fife, & Redman, 2009). CBAL Mathematics tasks are intended to provide both rich information and learning opportunities. As such, they are computer-based, and many include interactive elements or simulations. Most of the questions within tasks are constructed-response types and require the application of multiple concepts and skills in a real-world context.

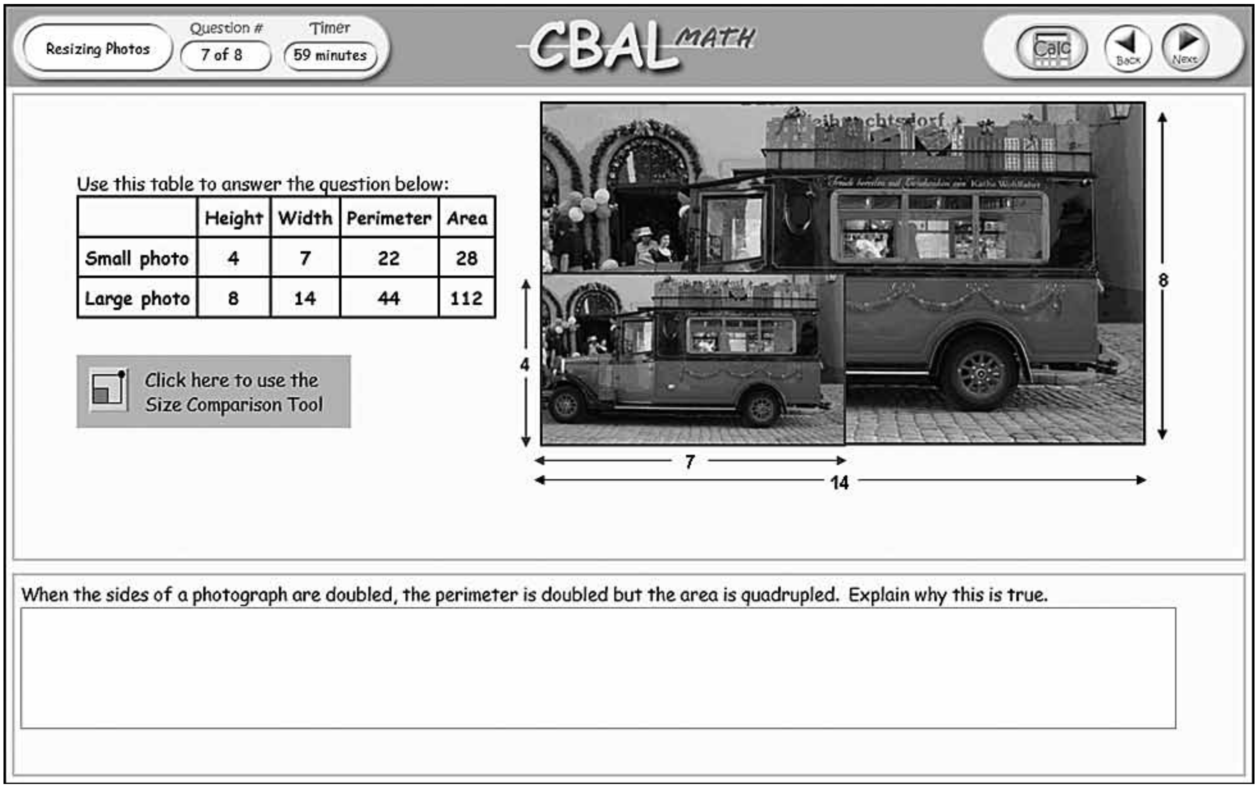
Resizing Photos

Resizing Photos is an extended task from a prototype accountability test developed for Grade 8. The task was designed to assess a large number of content and process competencies, but the focus is on using proportionality with understanding and mathematical argument. In particular, the task addresses the concept of *scale factor* in two dimensions. The scale factor refers to the ratio between corresponding measures of two similar shapes. For example, if one rectangle has dimensions $l \times w$ and a similar rectangle has dimensions $kl \times kw$, then the perimeter for the similar rectangle is $k(2l + 2w)$, and its area is k^2lw. Thus, the scale factor for the sides and the perimeter is k, and the scale factor for the area is k^2.

Early questions in *Resizing Photos* ask the student to consider whether rectangles are still in proportion after being resized in a variety of ways. Later questions in the task address the concept of scale factor for the sides, perimeter, and area. For example, one early question in the task shows the student an 8 × 10 in. photo[2] that has been proportionally enlarged to fit 16 × 20 in. photo paper. The student is asked to explain why the 16 × 20 in. paper will keep the 8 × 10 in. photo in proportion. In a subsequent question, the student is asked to explain why a 5 × 7 in. photo *cannot* be proportionally enlarged to exactly fit 15 × 20 in. paper. Thus, in these two questions, the student is asked to argue why something *is true* in one case and why something *cannot be true* in another. Early questions in the task require the student to calculate perimeters and areas for rectangles, in preparation for the culminating questions at the end of the task, which ask the student to explain the relationship between the scale factor for the sides and the scale factors for perimeter and area.

One of the later questions in *Resizing Photos* is shown in the top panel of Figure 3 (p. 24). To assist in answering this question,

[2] in. = inches; 1 inch is approximately 2.54 cm.

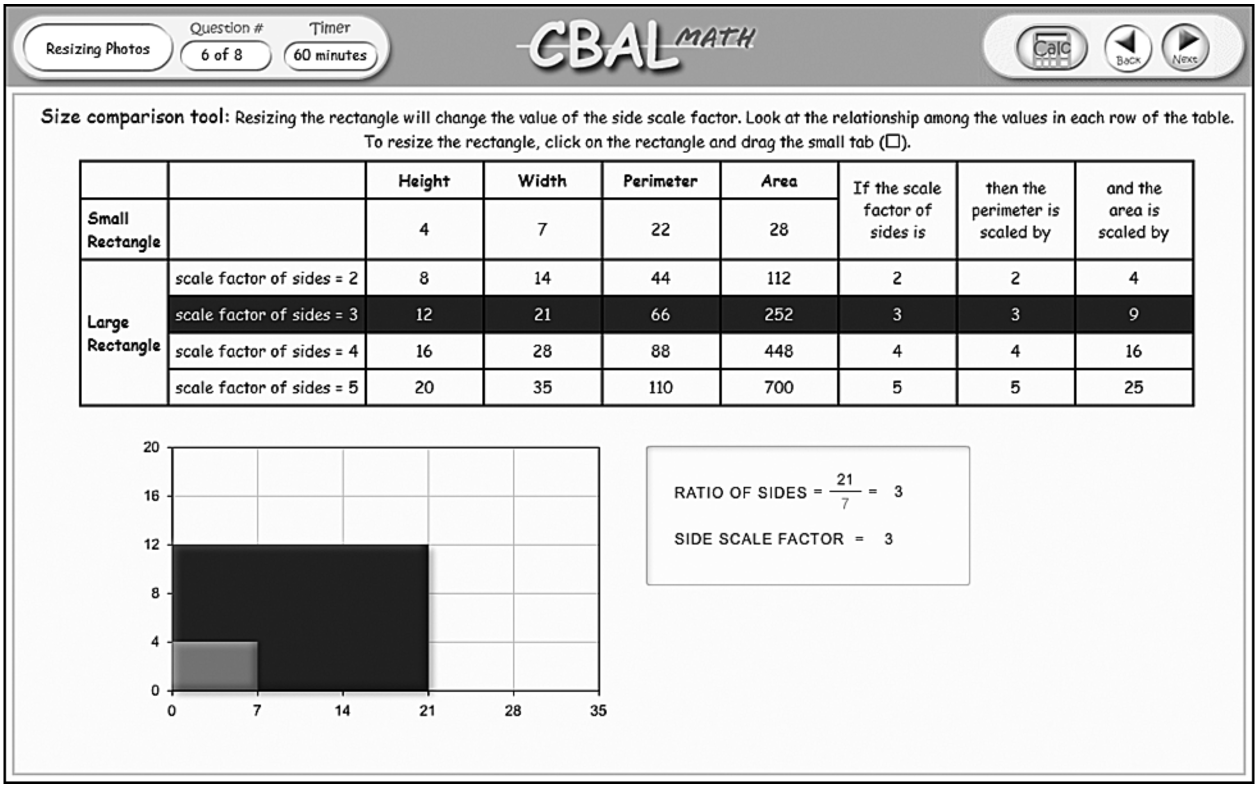


Use this table to answer the question below:

	Height	Width	Perimeter	Area
Small photo	4	7	22	28
Large photo	8	14	44	112

Click here to use the Size Comparison Tool

When the sides of a photograph are doubled, the perimeter is doubled but the area is quadrupled. Explain why this is true.



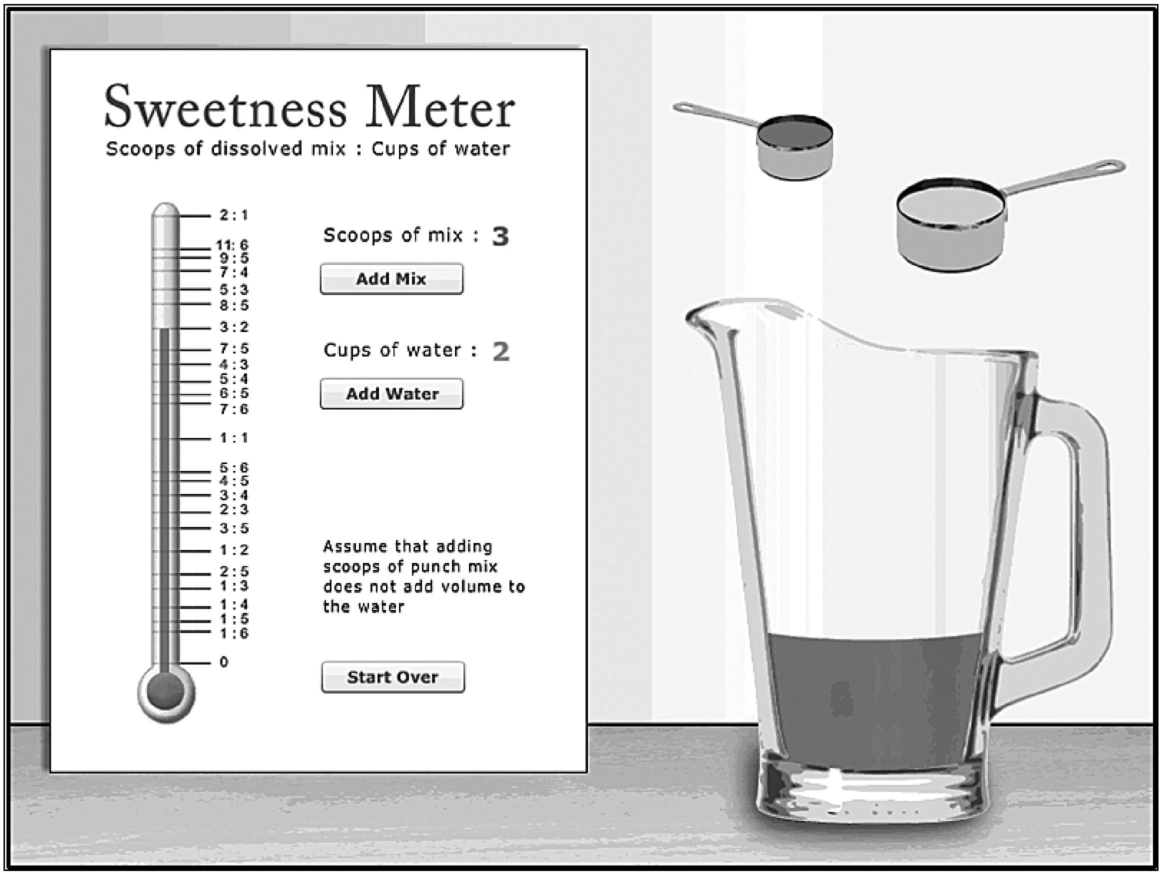
Size comparison tool: Resizing the rectangle will change the value of the side scale factor. Look at the relationship among the values in each row of the table. To resize the rectangle, click on the rectangle and drag the small tab (□).

		Height	Width	Perimeter	Area	If the scale factor of sides is	then the perimeter is scaled by	and the area is scaled by
Small Rectangle		4	7	22	28			
Large Rectangle	scale factor of sides = 2	8	14	44	112	2	2	4
	scale factor of sides = 3	12	21	66	252	3	3	9
	scale factor of sides = 4	16	28	88	448	4	4	16
	scale factor of sides = 5	20	35	110	700	5	5	25

RATIO OF SIDES = $\frac{21}{7}$ = 3

SIDE SCALE FACTOR = 3

Figure 3. **Screen captures from the CBAL Mathematics *Resizing Photos* task.**



Sweetness Meter

Scoops of dissolved mix : Cups of water

Scoops of mix : 3

Cups of water : 2

Assume that adding scoops of punch mix does not add volume to the water

***Figure 4.* A screen capture of the Sweetness Meter from the CBAL Mathematics *Proportional Punch* task.**

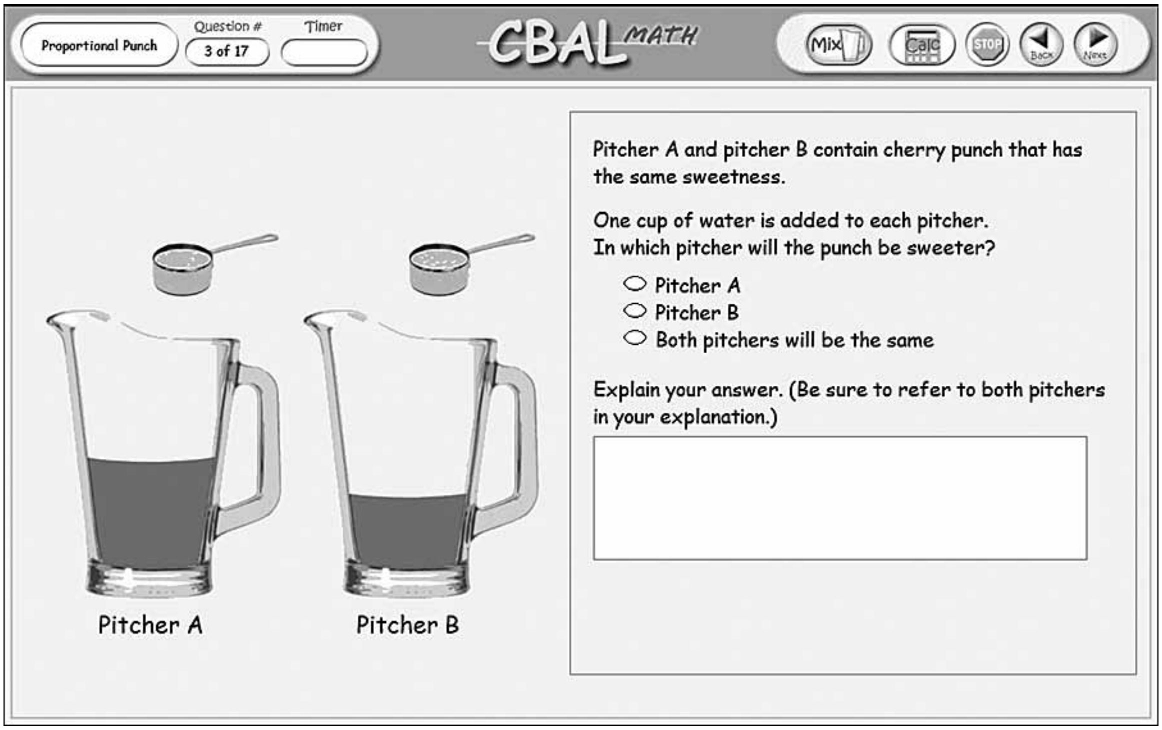
the student can access the *Size Comparison Tool*. Clicking a button brings up the tool (see the bottom panel of Figure 3). The small rectangle in the foreground serves as a reference. As the student drags from the upper right corner of the small rectangle, an enlarged rectangle appears behind it, and the corresponding row in the table is highlighted. The bottom panel of Figure 3 shows the state of the tool when the sides of the small rectangle have been enlarged by a scale factor of three. Understanding the relationship between the scale factor for the sides and the scale factor for the area is very difficult for students; the *Size Comparison Tool* makes it visually apparent why scaling the sides by a factor of k should scale the area by a factor of k^2.

Proportional Punch

Proportional Punch is a formative assessment task developed for Grade 7. Like *Resizing Photos,* the task is primarily concerned with whether students can use proportionality with understanding. It explores various aspects of proportional reasoning by considering different ratios of punch mix to water. The task progresses from qualitative questions about the relative “sweetness” (or strength) of different mixtures to simple quantitative questions involving specific ratios of scoops of mix to cups of water to more difficult quantitative questions involving more complicated ratios. The task includes a “sweetness meter” simulation tool that students use to discover equivalent ratios (see Figure 4).

After introductory questions to familiarize the student with the simulation tool, the task itself begins with questions that ask the student to reason qualitatively about the effect of adding more water or more punch mix to different volumes of punch with the same sweetness, i.e., the same ratio of scoops of mix to cups of water. Figure 5 (p. 26) shows two questions from the current version of *Proportional Punch*. The question in the top panel is an example of a qualitative question, while the question in the bottom panel is one of a series of quantitative questions in which the student is given different ratios of scoops of mix to cups of water and is asked to identify which mixture is sweeter.

The task then moves to a series of questions that elicit student understanding of proportionality at higher developmental lev-

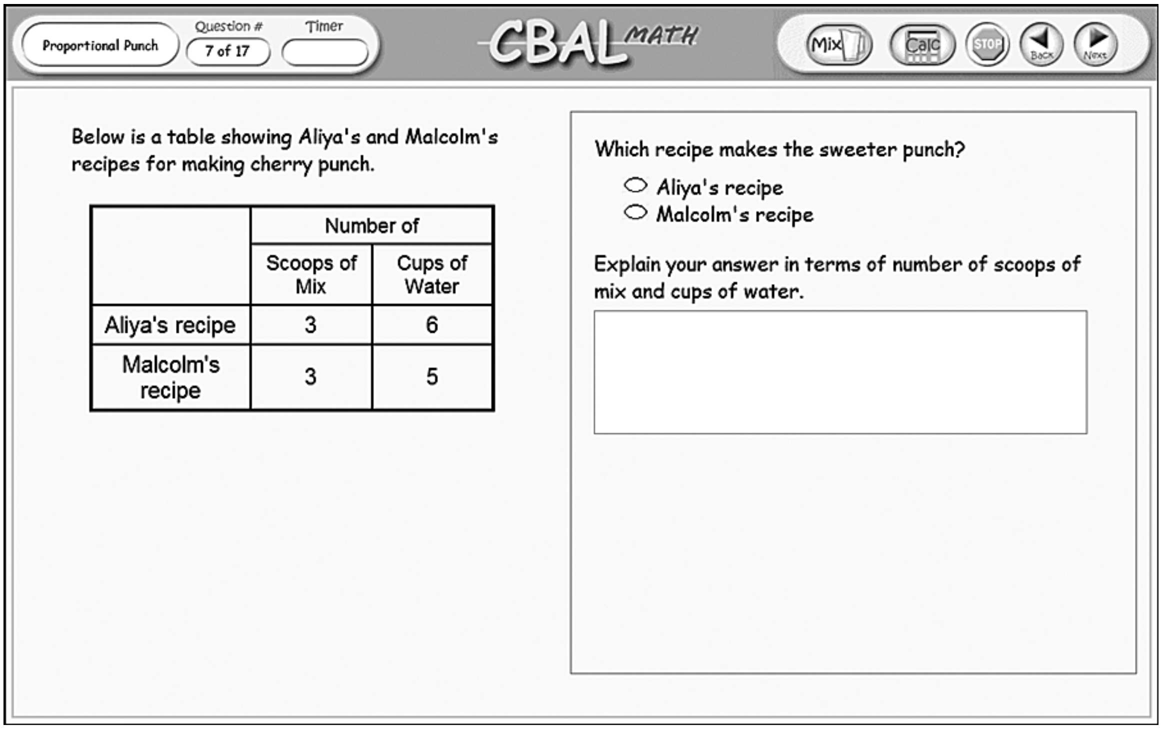


Pitcher A and pitcher B contain cherry punch that has the same sweetness.

One cup of water is added to each pitcher.
In which pitcher will the punch be sweeter?

○ Pitcher A
○ Pitcher B
○ Both pitchers will be the same

Explain your answer. (Be sure to refer to both pitchers in your explanation.)



Below is a table showing Aliya's and Malcolm's recipes for making cherry punch.

	Number of	
	Scoops of Mix	Cups of Water
Aliya's recipe	3	6
Malcolm's recipe	3	5

Which recipe makes the sweeter punch?

○ Aliya's recipe
○ Malcolm's recipe

Explain your answer in terms of number of scoops of mix and cups of water.

Figure 5. **Screen captures from the *Proportional Punch* task.**

els. For example, in several questions, the student is asked to determine the ratios of mix to water that have the same proportion as 1:3 when there are 3, 6, 9, and 12 cups of water. These questions assess the student's understanding of the multiplicative nature of proportionality in a context in which counting strategies can be used to identify the pattern. The student is then asked to extend these results to determine the number of scoops of mix needed to produce the 1:3 ratio when mixed with 36 cups of water. This question encourages the student to find a more efficient procedure than counting.

Later questions eliminate the scaffolding provided in earlier questions. For example, the student is asked to find how many cups of water should be combined with 28 scoops of punch mix when following a recipe that uses a ratio of 2 scoops of mix to 3 cups of water. One of the later questions provides an opportunity to reflect on what has been learned: "How would you compare two recipes to determine which one made the sweeter punch without actually making any punch?"

This question requires the student to generalize concepts addressed earlier in the task.

Moving Sidewalks

Moving Sidewalks is a formative assessment task designed to measure how well students use and understand algebra, and how well they translate among equivalent representations. The task is set in an airport: Rider A has just come in on a flight and wants to leave the gate on a moving sidewalk; Rider B has an outbound flight and wants to approach the gate on the sidewalk moving in the opposite direction. The sidewalks move at the same constant speed and the riders step onto them at the same time. The opening frame from the current version of the *Moving Sidewalks* simulation is shown in the top panel of Figure 6 (p. 28). To operate the simulation, the student can use the start and stop buttons or the sliders below. The lower panel of Figure 6 shows the last frame when the slider was set to stop the sidewalks at 6.5 seconds.

Moving Sidewalks was designed to address two grade levels. According to the *Learning Results: Parameters for Essential Instruction* (Maine Department of Education, 2007) for the state of Maine, in Grade 7 it is expected that "Students *understand* and use directly proportional relationships, $y = kx$," and in Grade 8 it is expected that "Students *understand* and use the basic properties of linear relationships, $y = kx + b$" (p. 29). Many of the questions involve modeling a rider's distance from the gate. With respect to time, Rider A's distance from the gate can be modeled as a directly proportional relationship, and Rider B's distance from the gate can be modeled as a linear function with a positive intercept and negative slope. Thus it is appropriate to ask Grade 7 students questions about Rider A and to ask Grade 8 students questions about both riders.

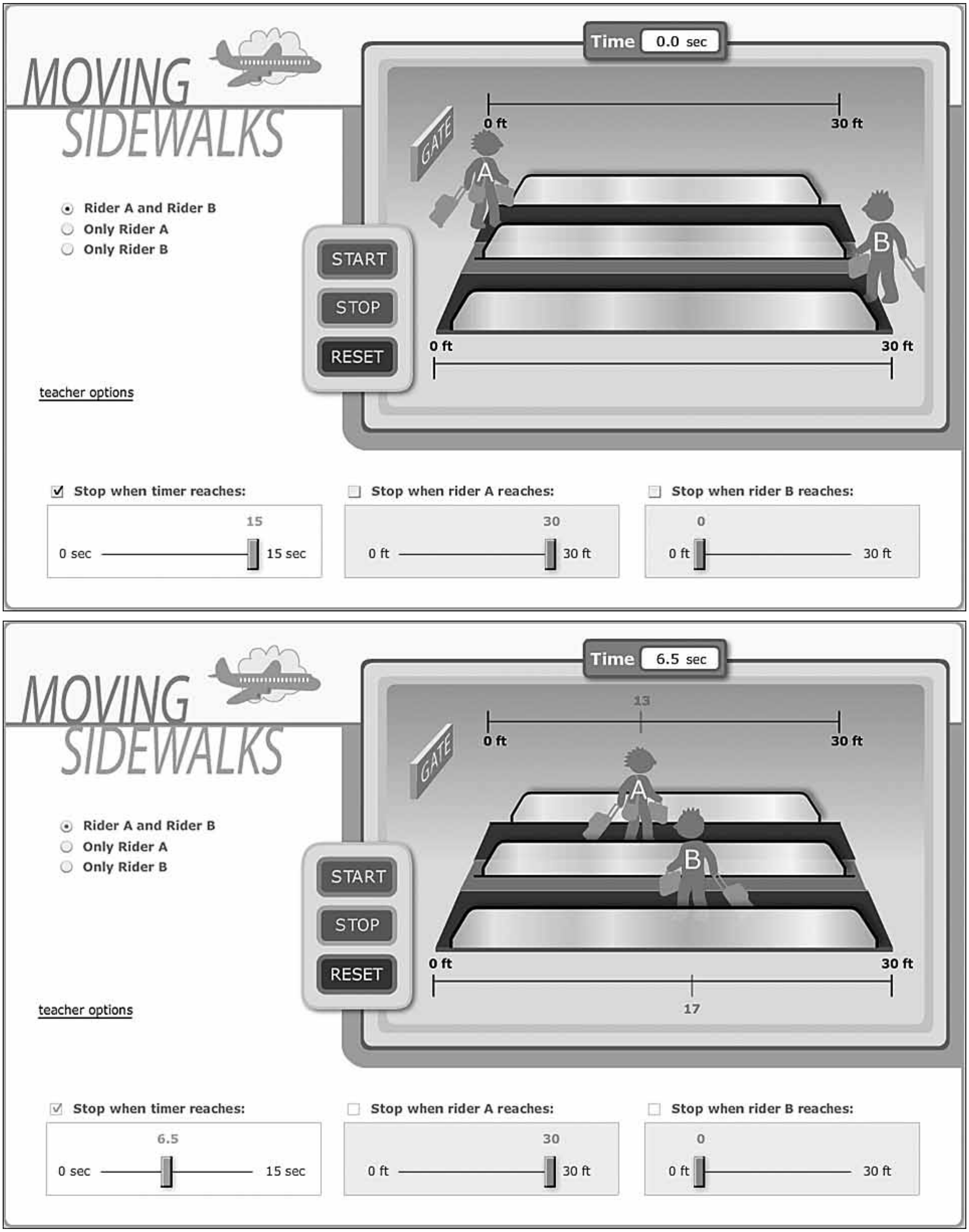
The task progresses from some basic familiarization questions to tabular representations, followed by graphical representations and finally algebraic equations. The student uses equivalent representations to express the same model ($y = kx$ or $y = kx + b$). More challenging questions can also be included: for example, the student can be asked to consider where the two riders meet. By default, the simulation is configured so that the sidewalks are 30 feet in length[3] and they each move at 2 feet per second. In the current version, however, the length of the sidewalks, the unit of length, and the rate at which each sidewalk is moving are all parameters. This allows for greater flexibility in the use of the simulation; for example, a student can compare scenarios in which the sidewalks move at the same rate to scenarios in which they move at different rates.

Earlier versions of the tasks described in this section as well as other CBAL Mathematics tasks have been piloted; an analysis of some of the data appears in Graf, Harris, Marquez, Fife, & Redman (2009).

Conclusions and Next Steps

The CBAL assessment system consists of three main components: an accountability component, a formative component, and a teacher professional development component. Within each domain (of reading, writing, and mathematics), the three components are based on a common competency model. In mathematics, the competency model is based on research in cognitive psychology and mathematics education. The CBAL Mathematics assessment design components (the competency model, the evidence model, task models and tasks) continue to evolve. The tasks described in this article are updated from earlier versions; many new tasks have since been created. Work on the project thus far has suggested areas where additional research and development is needed. For example, the evidence model should be further supported by research on how mathematical competencies develop over time. For each of a number of mathematical competencies, developmental models (Harris, Bauer, & Redman, 2008), sometimes referred to as learning progressions, are being constructed. The intent is that the developmental models will support the design of tasks as well as the interpretation of student responses. For example, in the design of a formative task, skills that develop early and come more easily to students can be leveraged to assist the transition

[3] 30 feet is approximately 9.14 meters — a very short sidewalk, but this parameter can be changed.



MOVING SIDEWALKS

Time 0.0 sec

● Rider A and Rider B
○ Only Rider A
○ Only Rider B

☑ Stop when timer reaches: 15 (0 sec – 15 sec)
□ Stop when rider A reaches: 30 (0 ft – 30 ft)
□ Stop when rider B reaches: 0 (0 ft – 30 ft)

MOVING SIDEWALKS

Time 6.5 sec

● Rider A and Rider B
○ Only Rider A
○ Only Rider B

☑ Stop when timer reaches: 6.5 (0 sec – 15 sec)
□ Stop when rider A reaches: 30 (0 ft – 30 ft)
□ Stop when rider B reaches: 0 (0 ft – 30 ft)

Figure 6. **Screen captures from the *Moving Sidewalks* task.**

to more advanced skills. If it is possible to identify levels of understanding or common types of errors that are particular to a stage of development, this should assist in the interpretation of student responses.

Another area of research is focused at the task level: CBAL mathematics tasks involve real-world contexts and many of the questions within tasks are constructed-response types. Although such tasks have the potential to provide very rich information about the nature of student understanding, they are also more difficult to develop and score than traditional tasks because their levels of complexity may introduce sources of construct-irrelevant difficulty. Research is underway to examine alternate approaches to revising questions with the goal of reducing construct-irrelevant difficulty.

Acknowledgements

We thank Malcolm Bauer, Jeff Haberstroh, and Sarah Ohls of the CBAL mathematics team for contributing to the updates presented in this article as well as for their ongoing efforts. We thank Marlene Supernavage for her work on task development, and Debbie Pisacreta, Janet Stumper, and Tom Hutchinson for work on art and design. For their continued work on the technology, we thank Jason Bonthron, Mike Ecker, and Mike Wagner. We thank the many colleagues who provided reviews of the original report, including Isaac Bejar, Randy Bennett, Gloria Dion, Irv Katz, Don Powers, and Cindy Tocci. Finally, we thank Kim Fryer, Darlene Rosero, and Ayleen Stelhorn for their editorial work on the original report and this article. Any errors are sole responsibility of the authors.

References

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43-61). New York: Springer.

Borchert, K. (2003). *Dissociation between arithmetic and algebraic knowledge in mathematical modeling*. Unpublished doctoral dissertation, University of Washington, Seattle, WA.

Clement, J., Lochhead, J., & Monk, G. S. (1981). Translation difficulties in learning mathematics. *American Mathematical Monthly*, *88*(4), 286-290.

Clement, J., Lochhead, J., & Soloway, E. (1979). *Translating between symbol systems: Isolating a common difficulty in solving algebra word problems* (TR 79-19). Amherst, MA: University of Massachusetts, Department of Physics and Astronomy, Cognitive Development Project.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill* (ETS Research Report No. RR-08-55). Princeton, NJ: Educational Testing Service.

Graf, E. A. (2009). *Defining mathematics competency in the service of cognitively based assessment for grades 6 through 8* (ETS Research Report. No. RR-09-42). Princeton, NJ: Educational Testing Service.

Graf, E. A., Harris, K., Marquez, E., Fife, J., & Redman, M. (2009). *Cognitively based assessment of, for, and as Learning (CBAL) in mathematics: A design and first steps toward implementation* (ETS Research Memorandum No. RM-09-07). Princeton, NJ: Educational Testing Service.

Harris, K., Bauer, M. I., & Redman, M. (2008). *Cognitive based developmental models used as a link between formative and summative assessment*. Paper presented at the annual meeting of the International Association for Educational Assessment (IAEA), Cambridge, UK.

Kaput, J. J. (1999). Teaching and learning a new algebra with understanding. In E. Fennema & T. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 133-155). Mahwah, NJ: Lawrence Erlbaum Associates.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Lee, L., & Wheeler, D. (1989). The arithmetic connection. *Educational Studies in Mathematics, 20*, 41-54.

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (Vol. 5, pp. 101-159). Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R., & Lamon, S. J. (1992). Assessing authentic mathematical performance. In *Assessment of authentic performance in school mathematics* (pp. 17-57). Washington, DC: American Association for the Advancement of Science.

Maine Department of Education. (2007). *Learning results: Parameters for essential instruction, mathematics section* [Regulation]. Available from the Maine Department of Education Learning Results Web site: http://www.maine.gov/education/lres/

Michener, E. R. (1978). Understanding understanding mathematics. *Cognitive Science, 2*, 361-383.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics*. Retrieved February 23, 2009, from http://www.nctm.org/standards/focalpoints.aspx?id=298

O'Reilly, T., & Sheehan, K. (2009). *Cognitively based assessment of, for, and as learning: A framework for assessing reading competency* (ETS Research Report No. RR-09-26). Princeton, NJ: Educational Testing Service.

RAND Mathematics Study Panel, & Ball, D. L. (2003). *Mathematical proficiency for all students: Toward a strategic research and development program in mathematics education* (No. MR-1643.0-OERI). Santa Monica, CA: The Office of Educational Research and Improvement (OERI).

Schoenfeld, A. H. (1994). Reflections on doing and teaching mathematics. In A. H. Schoenfeld (Ed.), *Mathematical thinking and problem solving* (pp. 53-70). Hillsdale, NJ: Lawrence Erlbaum Associates.

U. S. Department of Education. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Retrieved June 15, 2009, from http://www.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

Vennebush, P., Marquez, E., & Larsen, J. (2005). Embedding algebraic thinking throughout the mathematics curriculum. *Mathematics Teaching in the Middle School, 11*(2), 86-93.

Listening. Learning. Leading.®

www.ets.org

756860