Applied Psychological Measurement

http://apm.sagepub.com

Equating and Linking of Performance Assessments

Eiji Muraki, Catherine M. Hombo and Yong-Won Lee

Applied Psychological Measurement 2000; 24; 325
DOI: 10.1177/01466210022031787

Equating and Linking of Performance Assessments

Eiji Muraki, Catherine M. Hombo, and Yong-Won Lee
Educational Testing Service

Performance assessments (PA) are used in various contexts of large-scale educational assessment. It is often desirable to compare examinee performance on different forms of an assessment or on the same forms administered at different times. An overview of linking methods applied to PA is presented: major issues and recent developments in linking PAs are discussed, three common linking designs (single group, randomly equivalent groups, and nonequivalent groups with anchor items) are compared, and two major linking methodologies [classical and item response theory (IRT)] are evaluated from the PA perspective. Also described are two classical equating methods (linear and equipercentile) and several IRT equating methods (item response function, vertical, common population, and multiple-group). Areas for future research are identified. *Index terms: item response theory, large-scale assessment, National Assessment of Educational Progress, performance assessment, test equating, test linking.*

Performance assessment (PA) can be broadly defined as a procedure in which examinees are required to complete tasks or processes that demonstrate their ability to apply knowledge and skills, or to put knowledge and understanding into action in simulated or real-life situations (Messick, 1996; Nitko, 1996; Payne, 1997; Wiggins, 1993). PAs encompass a wide variety of test formats, such as constructed-response, essay, demonstration, oral presentation, role-playing, exhibits, portfolios, and direct observation. Examinees are asked to produce, create, or perform something during a period of time, and then either the process, the product, or both are evaluated against performance standards (Messick, 1996; Oosterhof, 1994).

PA is believed to be suitable for the assessment of higher-order thinking or problem-solving skills. In PA, the structure of responses is defined by the examinees, and such constructed responses can be scored for multiple levels of quality, rather than only as correct or incorrect. As a result, examinees can demonstrate skills that are not easily assessed with multiple-choice (MC) items (Messick, 1996). PA has also been perceived by teachers, educators, and students as having greater relevance to real-life situations—that is, PA is viewed as more authentic than traditional MC items (Haertel & Linn, 1996; Sax, 1997). For example, a doctor-patient simulation for assessing the clinical and interaction skills of medical students bears greater resemblance to a real situation they might face as practitioners; important skills are assessed that factual responses to written assessment items fail to tap. PA focuses on problem solving, reasoning, and the ability to integrate knowledge and information, rather than only on providing isolated bits of knowledge and information.

Despite the strengths of PA, however, this new test format poses a number of serious challenges for test equating and comparability of tests. First, the measurement constructs in performance tests are very likely to be multidimensional and unstable across contexts (Bond, Moss & Carr, 1996; Haertel & Linn, 1996; Mislevy, 1992b). The complexity of the tasks themselves could be the prime

Applied Psychological Measurement, Vol. 24 No. 4, December 2000, 325–337

reason, but context effects, practice effects, and other construct-irrelevant variance also contribute to the instability of test dimensionality across contexts. Thus, dimensionality requirements for equating might be violated in PA. Second, there are quite often no useful common anchor items (if there are any at all) between test forms. Also, the number of items used in PA is typically much smaller than is used in MC tests, which can result in an inadequate sampling of the construct domain (Dunbar, Koretz, & Hoover, 1991) and very unstable scores.

In many cases, PA items are not reusable, due to the "easy-to-memorize" nature of tasks and test security considerations (Haertel & Linn, 1996; Loyd, Engelhard, & Crocker, 1996). Although MC items in the same test can be used as an external anchor, they might not be an effective anchor (DeMauro, 1992), because PA and MC items are frequently designed to measure somewhat different constructs. MC and PA items are often not on the same scale, and their reliability often differs substantially (Miller & Legg, 1993).

Third, PA items are usually polytomously scored, based on the subjective ratings of judges, which adds to the complexity of equating procedures. For example, choices must be made about the appropriate relative weight to give the items. Intrajudge rating inconsistency and interjudge rater severity differences can serve as additional sources of error (Kolen & Brennan, 1995).

Recent developments in linking and equating methodology provide new approaches for use with PAs. A new conceptual framework has been developed for linking scores from two tests when the dimensionality assumptions for equating are violated as well as strictly met. Linn (1993) and Mislevy (1992a) proposed five different levels of score linking: equating, calibration, projection, statistical moderation, and social moderation. Equating is assumed to be possible only when the two tests are from the same set of test specifications, of equal lengths, and measure the same construct. However, test scores from two tests can still be linked in this framework when (1) the two tests measure the same constructs but are neither of equal length nor of the same accuracy (calibration), and (2) the two tests measure different constructs (projection and moderation).

Equating is defined here as the strictest form of establishing a translation between scores on two or more assessments, assuming that the tests are developed from the same test specifications. Linking is defined more broadly as a scaling method used to achieve comparability of scores from two different tests. Because the statistical requirements of equating are rarely met, most techniques described are referred to as linking methods (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999).

From a technical point of view, psychometric modeling [especially item response theory (IRT)] has advanced to a point where new models can handle polytomous items, multifaceted analysis, and even multidimensional tests (Ackerman, 1996; Hambleton, 1989; Linacre, 1988; Muraki, 1993; van der Linden & Hambleton, 1997). These newly developed models are being actively explored and applied to linking performance tests. For example, Baker (1992, 1993) developed the equating procedures for polytomously scored items based on Bock's (1972) nominal response and Samejima's (1969) graded response IRT models. Similarly, Muraki & Chang (1994) and Muraki & Hombo (1999) attempted to implement Muraki's (1992) generalized partial-credit models for test equating. Huynh & Ferrara (1994) compared partial-credit IRT equating and traditional equipercentile equating for constructed-response items. Multifaceted IRT models have also been explored for equating performance tests (Myford, Marr, & Linacre, 1996; Stahl, Lunz, & Wright, 1991). Multidimensional IRT equating methods have also been explored for test equating (Li, 1997).

Equating and Linking Designs

In many situations, it is desirable to compare examinee performance on different forms of an assessment or on forms administered at different times. The assessments might not have the same overall difficulty or score distribution, bringing into question whether the scores have the same meaning. Equating scores is intended to make the scores have the same meaning, regardless of when the assessment was administered or which form of the assessment an examinee took (Green, 1995).

Equating is defined by Kolen & Brennan (1995, p. 2) as "a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably." Equated scores on alternate test forms can then to be compared, and differences in examinee scores after equating can be attributed to differences in ability instead of differences in difficulty between test forms or other, construct-irrelevant, sources of variance. Equating is not intended to adjust for differences in content between assessments, and should be applied only to tests that are designed to the same specifications. Equating procedures should also possess properties of symmetry, equity, and group invariance.

Linking tests has a different purpose than equating, although similar statistical procedures are used for both. Linking establishes comparability, to the extent possible, between tests of different frameworks and test specifications (Feuer et al., 1999).

Data Collection Designs

Various designs can be used to collect assessment data, and the design selected will affect the equating methodology selected. Three common data collection designs (DCDs)—single group, randomly equivalent groups, and nonequivalent groups with anchor items—are described below. The use of each design in the specific context of PA data varies according to requirements of the design.

Single group. In this design, a single group of examinees takes both (or all) forms of an assessment. The statistical relationship between scores on the forms provides evidence as to the comparability of the content and difficulty of tasks. However, because the tasks can take several hours or days to complete, concerns about fatigue are serious (Gordon, Engelhard, Gabrielson, & Bernknopf, 1996; Loyd et al., 1996). An extended rest period might be required between administration of the forms. Also, the tasks might be dependent on each other; the performance on one might influence performance on the other. This problem is especially likely if the tasks are intended to be relatively parallel and to assess the same construct.

Local independence between items is a strong assumption of many data analysis models, and violations of this assumption pose a threat to valid linkages. In many forms of equating, item parameters and item weights influence the accuracy of linking. In optimal IRT scoring, item weights are a function of item parameters. If there is only a small chance of an examinee guessing the response (as is true with PA items), the optimal weights are proportional to the IRT item discrimination parameter. These parameters tend to be inflated by the presence of local item dependence; this can have noticeable effects on the scoring weights of such items, making the weights less than optimal in reality (Yen, 1993).

The order of administration usually counterbalances forms, to control order and fatigue effects (Crocker & Algina, 1986, Kolen & Brennan, 1995). If all forms of an assessment are available at the same time, as well as a sufficient number of examinees, then the single-group design can be used in a single administration session. Counterbalancing administration order is desirable if order effects are anticipated. Concerns about fatigue due to session length can be alleviated if the same group of examinees is available for all forms at different times and locations. However, very long delays between form administration can lead to changes in examinee ability, and should be avoided.

Single-group designs are uncommon in practice, because the requirements are difficult to meet. Often, administering both forms of an assessment to a single group of examinees is not practical;

the form being equated might not be developed when the base form is administered. Most test forms show some order effects. The length of test forms and probable fatigue can render this design impractical.

Randomly equivalent groups. In this DCD, examinees are randomly assigned a form of the assessment to complete (Kolen & Brennan, 1995; Yen & Ferrara, 1997). Randomly equivalent groups should lead to results that are theoretically comparable. Differences in performance between the groups can be attributed to differences in the difficulty of the form taken. This DCD minimizes the time required for each examinee, which is a strong consideration when PA tasks are being administered.

The randomly equivalent groups design is desirable if the following conditions can be met: (1) a sufficient number of examinees can complete the assessment concurrently, (2) examinees can be randomly assigned to the forms, and (3) all forms to be equated are available simultaneously.

This DCD is infrequently achieved in practice—the level of control needed to randomly assign examinees is rare, particularly if the forms are given over an extended time period. To some extent, in large-scale testing programs the population is self-assigned to a particular test form (e.g., candidates can select spring or fall administration of the Scholastic Aptitute Test). This is less of a concern in programs that have required administration for all examinees (e.g., state high school exit exams). Also, the examinee sample might be selected on a purposeful, nonrandom basis where population representation is the overriding criterion, as is the case in National Assessment of Educational Progress (NAEP; Donoghue et al., 1996). In NAEP, it is determined whether the assessed groups are equivalent in ability—a necessary assumption for the randomly equivalent design.

Nonequivalent groups with anchor items. In this DCD, the assessment forms are created to have some subset of test items in common. The forms are administered to different groups that are selected so that an assumption of random equivalence cannot be supported. Items can contribute to the examinee total score (referred to as *internal* items), or they might not contribute to the score (referred to as *external* items). External anchor items are frequently administered as a separate, timed block of items; internal anchor items are often interspersed throughout a form (Kolen & Brennan, 1995).

Anchor items should represent, as far as possible, the content and properties of the assessment forms to be equated, and they must be identical on both (or all) forms. Also, the items should be located in approximately the same positions on the forms. This DCD requires more time than the randomly equivalent groups design, depending on the number of anchor items, but less time than the single-group design.

Problems occur when the context, position, order, and timing of anchor items are altered. The "NAEP Reading Anomaly," showing surprisingly large declines in reading ability for the population measured between assessment cycles in 1984 and 1986, was apparently based partly on inconsistency in anchor item presentation. The anchor items were embedded in test booklets that contained blocks of items from differing subject areas, the time allotted for completion of the anchor items was altered, and the anchor items were presented in a different order (Zwick, 1991). A substantial proportion of the apparent observed differences in reading ability were ultimately attributed to these sources.

The most common equating design used in large-scale assessment is some variant of the nonequivalent-groups-with-anchor-items design. Secure anchor items can be retained over several administration cycles in many testing programs. Item release requirements can complicate the design, but if this obstacle can be overcome, anchor designs seem to function well.

Classical and IRT Linking Methodologies

The DCD used for linking and equating is usually the nonequivalent-groups-with-anchor-items design. Linking and equating methodologies are based on classical or IRT methods by their underlying assumptions about data. IRT is now commonly used in large-scale assessment. However, IRT methodology requires acceptance of some stringent assumptions about the data structure, access to and expertise in the use of specialized analysis software, and very large sample sizes (Hambleton & Swaminathan, 1985; Muraki & Bock, 1998). These requirements exclude assessment programs of smaller scope, limited resources, and less-specialized staff, which still use classical methods.

Classical Equating Methods

Classical test theory provides useful, but somewhat limited, information about examinee performance. Examinees are presumed to have a "true" score on the measure. However, this score is affected by random components, called errors of measurement, which determine the reliability of the measure. In classical methodology, two scores are considered equivalent if they "measure the same trait with equal reliability and the percentile ranks corresponding to the scores are equal" (Crocker & Algina, 1986, p. 457).

Linear equating. Linear equating assumes that, apart from differences in means and standard deviations (SDs), score distributions on two forms of a test are the same. This allows difficulty differences to vary along the score scale of the two assessments. Given this assumption, scores on the two forms can be matched using their z scores. The linear conversion is defined in terms of the mean (μ) and SD (σ) of the two scores (X and Y):

$$\frac{x-\mu(X)}{\sigma(X)}=\frac{y-\mu(Y)}{\sigma(Y)} . \tag{1}$$

When the SDs are equal, linear equating becomes the same as mean equating described by Kolen & Brennan (1995). Linear equating is appropriate if the score distributions differ only in mean and SD. If there are differences beyond these first two moments, or if the shape of the distributions differs, then linear equating is inappropriate.

One important consideration in the design of PAs is the scoring scale(s) to be used. PAs are often rated both holistically and on specific facets of the performance. Often the individual scales are of such limited range that score distributions are unstable and unsuitable for distribution-based equating methods. If multiple scales are aggregated to form a composite or total score, the score range might be large enough to permit linear equating. However, caution is needed when assuming that score distributions will be reasonably stable over multiple assessment populations. Large score distribution differences can invalidate a linear equating because the equating transformation would differ for another dataset.

A nonequivalent groups design with anchor items is used in the Advanced Placement (AP) program (College Entrance Examination Board, 1988) to equate tests across assessment years. A form of linear equating is most commonly used, although equipercentile methods have been used as well. Most AP exams have PA (free-response) and MC components, but the equating is done solely on the MC section of the exam. All PA items are released each year, and therefore cannot be retained as anchor items.

Equating only on the MC items requires assuming that "[a]ny trends over time in the candidates' achievement of the knowledge and skills measured by the multiple-choice section will be matched by similar trends in their achievement of the knowledge and skills measured by the free-response section" (College Entrance Examination Board, 1988). Performance on the free-response section might be influenced by multidimensionality, and might not be well-represented by the conversion developed from the MC section. Free-response items might also have different reliabilities than MC items, because human judges rate the free responses, introducing another source of error variance not present in MC scoring.

Equipercentile equating. Equipercentile equating involves determining which scores in a distribution have the same percentile rank. Those scores are then declared equivalent. The percentile rank for each number-correct or scale score is determined on both test forms, and a percentile-rank score curve is created for each assessment. This graph is used to locate a corresponding base score for any equated score in the range. Because actual scores on assessments are discrete rather than continuous, a method is required to approximate a continuous distribution of scores (e.g., Holland & Thayer, 1989).

Equipercentile equating relies heavily on the presumption that scores will have sufficient variance to allow the formation of a stable statistical distribution. PAs could have so few items and/or such a limited score range that this might not be the case. When PA items are scored on multiple subscales, and a composite, overall score is created, there might be sufficient variation. However, many PA items are scored on very limited, holistic scales, and the entire assessment can be composed of only one or two scored responses. In this situation, equipercentile equating might be inappropriate.

IRT Linking Methods

As its name implies, IRT focuses on modeling examinee responses at the item, rather than the test, level (Kolen & Brennan, 1995). Variance in performance across the ability (θ) scale can be examined, supporting the intuitive sense that θ affects performance in ways that true score and classical models cannot (Crocker & Algina, 1986). However, IRT makes strong assumptions (e.g., unidimensionality and local independence) that might not be supported by PA data.

PAs are not expected to measure a single trait. The multidimensional nature of the elicited performance is a basis for claims of the greater "real-world" validity of PAs. Local independence means that examinee responses to items are statistically independent, once θ is controlled. Both local independence and unidimensionality are violated to some extent in every response dataset; as long as the violations are not extreme, the model appears to be robust. However, violations generated by PA might be extreme enough that IRT models cannot be applied to PA data.

True and observed score equating. For IRT "true-score" equating, equated scores are based on estimated θ, determined by translating the "true score" [estimated number correct (ENC) or estimated proportion correct (EPC)] into the θ metric, then mapping that θ metric into the ENC/EPC metric on the second test form. For a given ENC/EPC on the base form, an iterative process involving the IRT model can be used to determine the examinee θ corresponding to the ENC/EPC. This θ estimate is then used to locate the appropriate ENC/EPC on the equated form (Kolen & Brennan, 1995).

The ENC/EPC of an examinee is assumed to be known—this does not exist in reality. This procedure produces an estimated ENC/EPC relationship, which is then applied to the observed scores. This has been justified by demonstrating that the true- and observed-score conversions are similar (Lord & Wingersky, 1984). However, Kolen (1981) found that observed- and true-score conversions produced somewhat discrepant results. An IRT observed-score linking method uses the IRT model to specify an estimated observed-score distribution for each form of an assessment. Equipercentile methods then link the scores on the forms. For this observed-score procedure, the form of the θ distribution must be specified (Kolen & Brennan, 1995).

Item and test response function methods. When item and θ parameters are estimated, the item response function (IRF; also referred to as item characteristic curve) is invariant up to a linear transformation (Hambleton & Swaminathan, 1985). Either the θ or item parameter metric must be fixed. Once the basis of the equating has been established, a method of establishing the transformation must be selected. The transformation can be based on the IRFs (Haebara, 1980) or the test response function (TRF, or test characteristic curve; Stocking & Lord, 1983).

IRFs represent the probability of a correct response across the θ scale. Items on test forms can be matched and used as the basis of the linking transformation if they are believed to assess the same trait and are reasonably parallel. The linking transformation minimizes the differences between the values of the IRFs on different test forms, and thus the distance between the IRFs. In this approach, the influence on the transformation of items that have different item parameter estimates but similar IRFs is reduced, compared to other methods. However, IRF methods do not take into account estimation errors in the item parameters.

The TRF is defined as the sum, over items, of the probability of a correct response across the θ continuum. Stocking & Lord (1983) defined a method using the TRF as the basis of linking. In this method, the difference to be minimized is the squared difference between TRFs across the θ scale, cumulated over examinees.

Vertical equating. When achievement level batteries are administered to multiple grade/age groups of examinees, it is sometimes desirable to report scores on a common scale. The batteries typically are designed to have different content, selected to be appropriate for the examinees at that level (Kolen & Brennan, 1995). A scaling test with samples of content across all levels might be appropriate. Alternatively, items considered appropriate across pairs of levels can be administered. For example, in an assessment with students from 4th, 8th, and 12th grade, certain items would appear on both the 4th and 8th grade assessments. Other items would appear on both the 8th and 12th grade assessments. The results allow the three grade levels to be placed on a single, common score scale. This procedure was used prior to 1993 in NAEP.

Common-population linking. For convenience and clarity in the following explanation, it is assumed that the procedure is linking two assessments administered across some time period, with Assessment 2 administered second and then linked to Assessment 1. Item parameters are estimated for the items administered to examinees as part of both assessment cycles. The anchor-item linking procedure is used. Estimation can be done with data from both assessments, or with only Assessment 2 data. Item parameter estimation with the Assessment 1 data is done at the time of collection. If data from both populations are used, the common items can either be constrained to have identical parameter values, or the minimized difference between the two sets of common item parameters can be allowed to drift, which is defined as differential change in item parameter values over time (Goldstein, 1983). Then, scale linking is performed to linearly transform the new parameter estimates to the scale of Assessment 1. This is also used to place the Assessment 2 data on the same scale as Assessment 1.

Common-population linking is preferred over common-item linking because the score distributions are less sensitive to changes in individual items (Donoghue & Mazzeo, 1992). At the pre-equating stage, if some linking items clearly function differently across assessment years, these items can be treated as separate items for each year. The remaining linking items are constrained to have identical IRFs across assessment years during the joint calibration. If sampling weights are used, they must be restandardized to ensure that each year has a similar sum of weights, and so has an approximately equal influence in the joint calibration (Donoghue, Isham, & Worthington, 1996).

Multiple-group IRT model. In both common-item and common-population linking, it is possible that the score bias, introduced by the linking method, will compound across assessment years because of the pairwise linking process. If this occurs, judgments about the performance of different populations might become less valid as a trend line is continued.

Bock, Muraki, & Pfeiffenberger (1988) proposed a multiple-group dichotomous IRT model with a drift parameter (DRIFT) and estimated the model parameters, means, and SDs of the multiple θ distributions over years by the marginal maximum likelihood EM method. For example, the difficulty level of certain test items might increase as technological or cultural changes occur during the useful life of a scale. The item parameter drift model can also be viewed as another form of a differential item functioning (DIF) model, where a group variable is temporal, rather than demographic subgroup membership.

Bock et al. (1988) found numerous examples of DRIFT among items in a form of the College Board's AP Physics Examination, which was administered annually over a ten-year period. This particular method is currently implemented in BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), and is described by Bock & Zimowski (1997). Thissen, Steinberg, & Wainer (1993) used this multiple-group dichotomous IRT model as a DIF method. They noted that the method might be useful in routine test development.

Muraki (1999) applied the multiple-group generalized partial-credit model to the NAEP writing assessments. PC-PARSCALE 3.5 (Muraki & Bock, 1998) has the capability of analyzing parallel and nonparallel DIF and DRIFT of dichotomously and polytomously scored items.

With multiple-group IRT, it is possible to simultaneously scale the items in all populations for which data are available. The θ distribution is used from one assessment cycle to anchor those of the other populations and provide the scale of measurement. This procedure is not a pairwise linking method. Thus, more than two groups can be simultaneously linked. In this method, there is the potential to eliminate bias in scale linking, even under the most difficult conditions encountered operationally and in the extreme quantiles of the population. Hedges & Vevea (1997) found that multiple-group IRT methods have considerable merit for equating and scale linking. The problem of potentially compounding errors across assessment cycles is eliminated, because all groups are scaled simultaneously.

The multiple-group IRT model assumes that different groups have different distributions with means and SDs that are not necessarily normal. These empirical posterior distributions are estimated simultaneously with the estimation of the model's fixed parameters (including DRIFT). To obtain the parameters, the constraint is imposed for the multiple-group model, which implies that the overall difficulty levels of a test or a set of common items given to both the base group and a temporal group, respectively, are the same. Only the item-by-group interaction is considered DRIFT; average DRIFT for all items is considered a change in the population of potential examinees, and is absorbed in the estimate of the year-group mean (Bock et al., 1988). Thus, the item difficulty parameters for the other temporal groups are adjusted. Any overall difference in terms of test or subtest difficulty is assumed to be the difference in θ level for subgroups. The θ difference among groups can then be estimated by the posterior distributions.

Comparison of common-population and multiple-group linking procedures. The intercept and slope of the linear transformation function are used in the final rescaling step of common-population linking. The center of the distribution is shifted and the scale of measurement is altered, resulting in a rescaled score distribution. The critical step in common-population equating is the simultaneous calibration that produces item parameters. In this step, the linking items are constrained to be identical. This is a rather strong assumption that is not always supported by the results of the calibration. Some linking items are removed from their bank and are treated as separate between assessment waves, based on visual inspection. This involves professional judgment, but is still subjective in nature. Also, common-population linking is implemented only across two consecutive waves of data. Thus, linking errors can accumulate through later years if the method is used across several assessment cycles.

In multiple-group IRT linking, items common to assessment waves are not necessarily constrained to exhibit identical behavior. Decisions about significant changes in item parameters across time are made using a statistical, less-subjective criterion, the SDIF^2 values (Muraki, 1999; Muraki & Engelhard, 1989). Another advantage of this method is that it will function correctly with samples of unequal size or differing composition, whereas sampling weights are required for correct implementation of the common-population procedure. The development of these sampling weights involves considerable effort. The capability of multiple-group linking to encompass a series of assessment data waves is expected to eliminate the possible compounding of linking errors across multiple temporal links.

Conclusions

Several promising lines of research have begun on the effective linking of complex performance data. They fall into two general classes: the applicability of existing methodology to PA data and the development of new techniques specifically for PA data.

IRT models have been developed primarily for MC items. In addition to this traditional format, PAs require varied item formats. Several polytomous IRT models have been developed for constructed-response items, and their characteristics have been investigated (Hemker, Sijtsma, Molenaar, & Junker, 1997). However, the appropriateness of polytomous IRT models for certain response formats has not been investigated thoroughly. The degree of association between constructed responses and MC items needs to be explored from the psychological as well as psychometric point of view. Methods for obtaining optimal portions of various response formats given testing time constraints and content requirements need to be devised.

The unidimensionality assumption that is part of most dichotomous and polytomous IRT models is not strictly met in most PAs. Each performance item might purely measure a single cognitive trait, but dimensionality can vary across a set of items. Alternatively, each single item might measure several trait dimensions simultaneously, but the proportion of each trait dimension can vary across a set of PA items. Furthermore, depending on the interaction between the subgroup and test item characteristics, the dimensionality in assessments might change for subgroups or across administration times.

Several multidimensional extensions (Bock, Gibbons, & Muraki, 1988; Muraki & Carlson, 1995; Reckase, 1985) of dichotomous and polytomous IRT models have been proposed. The investigation of linking methods based on multidimensional IRT models has just begun (Li, 1997). However, as the applicability of multidimensional IRT equating methods is explored for PA, there is some basic research that needs to be done. Multidimensional IRT equating methods need to be studied in the context of measuring integrated skills or joint assessments of dimensionally distinct skills. For instance, tasks can be created so as to measure reading and writing skills simultaneously in the same task (Grabe, 1997). Both integrative skill tests (Mislevy, 1995) and multidimensional IRT equating procedures (Li, 1997) are still at an early stage of development. Their intersection seems to be a promising area of study for possible use with PA data.

Local item dependence among items is a more serious issue in PA than in MC tests and could have an impact on IRT equating methods (Carey, 1996). The construction of a testlet including a set of locally dependent items as a unit was proposed as a way to overcome this problem (Yen, 1993). In MC reading comprehension tests, for instance, a set of items in the passage-based testlet was dichotomously scored and aggregated at the testlet level to minimize the impact of local item dependence (Lee, 1998). In PA, however, each item is usually scored polytomously. Aggregating polytomously scored items in the testlet adds complexity to IRT analysis; the appropriate IRT models are not yet available for such situations. Local item dependence in PA is closely related to construct-irrelevant variance among performance tasks and multidimensionality of the test that cannot be attributable to the assessment construct. Therefore, the issue of local dependence should be examined further in the context of multidimensional modeling, as well as testlet models for polytomous items.

Raters' judgments for PA item scoring is a new facet that is not used for MC test items. When multiple raters score an examinee's constructed responses, a multifaceted model with a rater parameter is often used for scoring and scaling PAs. By applying Wright & Linacre's (1992) multifaceted IRT model, rater parameters can be estimated separately, eliminating the raters' rating effect from the estimation of other model parameters. Examinee scores based on the multifaceted model are, therefore, theoretically free from irrelevant rater parameter effects. In most of the multifaceted models, the rater effect is assumed to be a fixed linear component in the IRT logit. However, the present authors have observed that the raters' rating severity is influenced by characteristics of the individual examinee responses. It is too simplistic to treat the rater effect as a fixed and separable entity from the other model effects. More realistic parameterization of rater effects is necessary to maintain the stability among linked PAs, particularly when the common items involve the raters' judgments. Bock, Brennan, & Muraki (1998) proposed a correction to the IRT likelihood due to rater reliability and interdependencies among raters. Their method is the first serious attempt at synthesizing classical item statistics and IRT. Further research on this synthesis will help refine the linking methods used in PAs.

Common raters have been studied as a potential equating element between test forms through rater calibration, especially when there are no good anchor items, as in a single-item writing test. This possibility has been explored to some extent by Stahl et al. (1991) and Myford et al. (1996). Because of the complexity of human rating behavior and its error structure, this issue needs further attention. Burnstein et al. (1999) developed the Electronic Essay Rater (e-rater). The e-rater was designed to automatically analyze essay features based on writing characteristics, specified by a six-point rating scale. They found that exact or adjacent agreement between the e-rater and a human rater was quite high. Because the e-rater's performance was stable across administrations, using it might provide a new way to equate or link PAs.

More fundamental issues should also be considered. First, there is the question of whether constructed response items should be treated as substitutes for MC items or whether they provide unique or additional information about cognitive performance that MC items cannot measure. Second, many—if not most—PA items are scored on a holistic scale with a very limited range. Equating methodology that works well within these constraints is clearly needed.

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Measurement in Education, 20,* 311–329.

Baker, F. B. (1992). Equating under the graded response model. *Applied Psychological Measurement, 16,* 87–96.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17,* 239–251.

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

Bock, R. D., Brennan, R. L., & Muraki, E. (1998). *The information in multiple ratings.* Unpublished manuscript.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25,* 275–285.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response* *theory* (pp. 433–448). New York: Springer-Verlag.

Bond, L., Moss, P., & Carr, P. (1996). Fairness in large-scale performance assessment. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 117–140). Washington DC: National Center for Education Statistics.

Burnstein, J., Kukick, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1999). *Automated scoring using a hybrid feature identification technique.* Unpublished manuscript.

Carey, P. (1996). *A review of psychometric and consequential issues related to performance assessment* (TOEFL Research Monograph No. MS-3). Princeton NJ: Educational Testing Service.

College Entrance Examination Board (1988). *The College Board technical manual for the Advanced Placement program.* http://www.collegeboard.org/ap/techman/chap3/score.htm. Princeton NJ: Educational Testing Service.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth TX: Harcourt, Brace, & Jovanovich.

DeMauro, G. E. (1992). *Investigation of the appropriateness of the TOEFL test as a matching variable to equate TWE topics* (TOEFL Research Report No. 37). Princeton NJ: Educational Testing Service.

Donoghue, J. R., Isham, S. P., & Worthington L. H. (1996). Data analysis for the reading assessment. In N. L. Allen, D. L. Kline, & C. A. Zelenak (Eds.), *The NAEP 1994 technical report* (Report No. NCES 97-897, pp. 267–308). Washington DC: National Center for Education Statistics.

Donoghue, J. R., & Mazzeo, J. (1992, April). *Comparing IRT-based equating procedures for trend measurement in a complex test design.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289–303.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington DC: National Academy Press.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20,* 369–377.

Gordon, B., Engelhard, G., Gabrielson, S., & Bernknopf, S. (1996). Conceptual issues in equating performance assessments: Lessons from writing assessment. *Journal of Research and Development in Education, 29,* 81–88.

Grabe, W. (1997, March). *Developments in reading research and their implications for computer-adaptive reading assessment.* Paper presented at the 19th Language Testing Research Colloquium Conference, Orlando FL.

Green, B. F. (1995). Comparability of scores from performance assessments. *Educational Measurement: Issues and Practice, 14 (4),* 13–15.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144–149.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 59–78). Washington DC: National Center for Education Statistics.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York: Macmillan.

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

Hedges, L. V., & Vevea, J. L. (1997). *A study of equating in NAEP.* Palo Alto CA: NAEP Validity Studies Panel, American Institutes for Research.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62,* 331–347.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (Technical Report No. 89-84). Princeton NJ: Educational Testing Service.

Huynh, H., & Ferrara, S. (1994). A comparison of equal percentile and partial credit equatings for performance-based assessments composed of free-response items. *Journal of Educational Measurement, 31,* 125–141.

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

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

Lee, Y. (1998). *Examining the suitability of an IRT-based testlet approach to the construction and analysis of passage-based item sets in an EFL reading comprehension test in the Korean high school contexts.* Unpublished doctoral dissertation, Pennsylvania State University, University Park, *Dissertation Abstracts International, 59/08-A,* 2893.

Li, Y. H. (1997). *An evaluation of multidimensional IRT equating methods by assessing the accuracy of transforming parameters onto a target test metric.* Unpublished doctoral dissertation. University of Maryland, College Park, *Dissertation Abstracts International, 58/11-A,* 4246.

Linacre, M. (1988). *Many-faceted Rasch measurement.* Chicago: MESA Press.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83–102.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453–461.

Loyd, B., Engelhard, G., & Crocker, L. (1996). Achieving form-to-form comparability: Fundamental issues and proposed strategies for equating performance assessments for teachers. *Educational Assessment, 3,* 99–110.

Messick, S. (1996). Validity of performance assessment. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington DC: National Center for Education Statistics.

Miller, M. D. & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice, 12 (2),* 9–15.

Mislevy, R. J. (1992a). *Linking educational assessments: Concepts, issues, methods, and prospects (Policy Issue Perspective).* Princeton NJ: Educational Testing Service, Policy Information Center.

Mislevy, R. J. (1992b). Scaling procedures. In E. G. Johnson & N. L. Allen (Eds.), *The NAEP 1990 technical report* (Report No. 21-TR-20, pp. 199–213). Washington DC: National Center for Education Statistics.

Mislevy, R. J. (1995). Test theory and language learning assessment. *Language Testing, 12,* 341–369.

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

Muraki, E. (1993, April). *Variations of polytomous item response models: Raters' effect model, IF model, and trend model.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta GA.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement, 36,* 217–232.

Muraki, E., & Bock, R. D. (1998). *PARSCALE (Version 3.5): IRT item analysis and test scoring for rating-scale data* [Computer program]. Lincolnwood IL: Scientific Software.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73–90.

Muraki, E., & Chang, H. (1994). *Horizontal and vertical test equating methods based on the generalized partial credit model.* (ETS Internal Report). Princeton NJ: Educational Testing Service.

Muraki, E., & Engelhard, G. (1989, April). *Examining differential item functioning with BIMAIN.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Muraki, E., & Hombo, C. (1999). *Application of a multiple-group generalized partial credit model to NAEP linking procedures.* Unpublished manuscript.

Myford, C., Marr, D. B., & Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the Test of Written English* (TOEFL Research Report No. 52). Princeton NJ: Educational Testing Service.

Nitko, A. J. (1996). *Educational assessment of students* (2nd ed.). Englewood Cliffs NJ: Prentice-Hall.

Oosterhof, A. (1994). *Classroom applications of educational measurement* (2nd ed.). New York: Macmillan.

Payne, D. A. (1997). *Applied educational measurement.* Belmont CA: Wadsworth.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17.*

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Belmont CA: Wadsworth.

Stahl, J. A., Lunz, M. E., & Wright, B. D. (1991, April). *Equating examinations that include judges (multiple facets).* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago.

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale NJ: Erlbaum.

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

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75,* 200–214.

Wright, B. D., & Linacre, J. M. (1992). *BIGSTEPS Rasch analysis* [Computer program]. Chicago: MESA Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local dependence.

Journal of Educational Measurement, 30, 187–213.

Yen, W. M., & Ferrara, S. (1997). The Maryland school performance assessment program: Performance assessment with psychometric quality suitable for high-stakes usage. *Educational and Psychological Measurement, 57,* 60–84.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer program]. Lincolnwood IL: Scientific Software.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP Reading Proficiency. *Educational Measurement: Issues and Practice, 10 (3),* 10–16.

Author's Address

Send requests for reprints or further information to Eiji Muraki, Rosedale Road, Educational Testing Service, Princeton NJ 08541, U.S.A. Email: emuraki@ets.org.