Journal of Educational Measurement
Spring 2001, Vol. 38, No. 1, pp. 19–49

Evaluating Comparability in Computerized Adaptive Testing: Issues, Criteria and an Example

Tianyou Wang
ACT
Michael J. Kolen
The University of Iowa

When a computerized adaptive testing (CAT) version of a test co-exists with its paper-and-pencil (P&P) version, it is important for scores from the CAT version to be comparable to scores from its P&P version. The CAT version may require multiple item pools for test security reasons, and CAT scores based on alternate pools also need to be comparable to each other. In this paper, we review research literature on CAT comparability issues and synthesize issues specific to these two settings. A framework of criteria for evaluating comparability was developed that contains the following three categories of criteria: validity criterion, psychometric property/reliability criterion, and statistical assumption/test administration condition criterion. Methods for evaluating comparability under these criteria as well as various algorithms for improving comparability are described and discussed. Focusing on the psychometric property/reliability criterion, an example using an item pool of ACT Assessment Mathematics items is provided to demonstrate a process for developing comparable CAT versions and for evaluating comparability. This example illustrates how simulations can be used to improve comparability at the early stages of the development of a CAT. The effects of different specifications of practical constraints, such as content balancing and item exposure rate control, and the effects of using alternate item pools are examined. One interesting finding from this study is that a large part of incomparability may be due to the change from number-correct score-based scoring to IRT ability estimation-based scoring. In addition, changes in components of a CAT, such as exposure rate control, content balancing, test length, and item pool size were found to result in different levels of comparability in test scores.

Computerized adaptive testing (CAT) research has, in recent years, emphasized solving practical problems that arise when implementing CATs in large scale testing programs. In some testing programs, the CAT co-exists with a paper-and-pencil (P&P) conventional test and the scores based on the two modes are used interchangeably. In these programs, examinees' scores should be comparable so that no examinee receives an unfair advantage by taking the test in a particular mode. Comparability between the CAT version and the existing P&P version of the test is an important issue in developing the test. A number of studies (e.g., Davey & Thomas, 1996; Eignor, 1993; Eignor & Schaeffer, 1995; Eignor, Way, & Amoss, 1994; Eignor, Stocking, Way, & Steffen, 1993; Lunz & Bergstrom, 1995; Mazzeo, Druesne, Raffeld, Checketts & Muhlstein, 1991; Mills & Stocking, 1996; Parshall & Kromrey, 1993; Schaeffer, Reese, Steffen, McKinley & Mills, 1993; Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995; Segall, 1995; Segall & Carter, 1995;

and van de Vijver & Harsveld, 1994) have examined comparability of CATs and P&P tests.

Comparability issues may arise in two settings. The first setting is when a CAT version of a text coexists with its P&P conventional version. In this setting, the aim is to make the scores on both versions to be comparable to an acceptable level. The second setting is when two or more alternative CAT item pools are used. This setting includes using two or more distinct item pools or using one item pool that is updated. For this setting, the goal is to make the different CAT versions comparable to each other. When the CAT version is first introduced in a testing program and is not intended to replace the P&P version immediately the first setting may be of primary concern. As new CAT versions are developed the second setting will eventually become more important. These two settings may share some common comparability issues, but there also may be unique problems and challenges to be addressed within each setting.

The first purpose of the present paper is to discuss and integrate important comparability issues that are unique to each of these two different settings. Special attention has been paid to the need for achieving comparability, what the desired or acceptable level of comparability might be, what might contribute to the challenges and threats to comparability, and what ways comparability might be improved. The second purpose of the paper is to develop a framework of criteria for evaluating comparability by reviewing and synthesizing the literature in these areas. In this framework, important terms are defined, criteria for evaluating comparability are described, various CAT algorithms that have been proposed to improve comparability are examined, and additional ways to improve comparability are explored. Finally, the Mathematics test of the ACT Assessment is used to illustrate the applications of this framework and ways for assessing and improving comparability. The example is illustrative in that it highlights a small portion of the criteria discussed in the paper, demonstrates how to use simulation techniques to evaluate comparability at an early stage of CAT development, and does not take the additional efforts to improve comparability.

Comparability Issues Specific to CAT and P&P

There has been an informal discussion in the measurement community about whether it is a sound practice to build CAT tests that are comparable to their P&P counterparts. The main argument against this practice is that CATs are purported to be superior to P&P tests both in terms of their administrative features and their psychometric properties, and that to constrain them to be comparable to their P&P counterparts would diminish the advantages of having them. Also, they think CAT should be open to more innovative and interactive item formats to take full advantage of testing on new generations of computers. Some researchers believe that, given the different features of test design, such as the adaptive item selection in CAT and the mode of testing, it is even unrealistic to expect that comparability can be achieved even when comparability is desired. The arguments for this practice, however, are mainly pragmatic. So long as CAT versions co-exist with their P&P counterparts, which is often necessary for practical reasons, it would be extremely confusing and inconvenient for users of the test scores if two separate

reporting scales are adopted for the two modes. For many large scale testing programs the reporting score scale and the services relating to these scales have been so well established that any small change to the report score scale would be extremely risky for the testing program. As long as CAT scores are reported on the same scale as the P&P scores, comparability between these two types of scores must be established to an acceptable level. The remainder of this paper assumes that there is a need to achieve comparability between CAT and P&P scores. A commonly accepted level of comparability is that the CAT scores should be comparable to the P&P scores to the extent that two typical P&P test forms are comparable to each other.

Several major differences between CATs and their P&P potentially could contribute to incomparability. Kolen (1999-2000) summarized four aspects of such differences: (1) differences in test questions, (2) differences in test scoring, (3) differences in testing conditions, and (4) differences in examinee groups. His discussion provides general guidelines for those aspects of test development and administration to which CAT developers should pay attention to avoid serious problems with comparability.

The fact that different examinees are administered different items within a CAT version complicates the definition of test content, especially for educational tests. Should each examinee take the same number of items from each content category in the table of specifications for the test? Or, should the content be tailored to the examinee's skill level? Regardless of the answer to these questions, how can we be sure that the content of the CAT is comparable to the content of the P&P test? Answers to these questions are not easy but must be addressed in establishing comparability.

The mode of testing has the potential to affect the performance of examinees on tests and should be taken into account when addressing comparability of CATs and P&P tests. Mode effects can be most clearly studied when P&P tests are compared to computerized linear tests. Because both of these types of tests use the same scoring methods and same scoring procedures, score differences are directly attributable to mode effects. Mazzeo and Harvey (1988) reviewed studies that compared these two types of tests. Their review indicated that the computerized linear tests tended to be more difficult than their P&P versions. Primarily, these studies examined mean differences and correlational indices to assess the comparability of the two modes. The studies, in general, found that the constructs being measured in the two modes were similar for power tests but not for speeded tests. A review and meta-analysis by Mead and Drasgow (1993) resulted in similar conclusions.

The scoring methods used with CATs typically differ from those of P&P test because different sets of items are administered to different examinees in CAT. In P&P testing, all examinees administered a particular form are generally administered the same items. Also, CATs typically are scored based on item response theory IRT ability estimates, whereas raw scores are formed in P&P tests as a sum of the item scores. These differences in scaling and scoring make it difficult to directly compare raw scores. Stocking (1997) proposed a method of scoring for CAT which is based on the number-correct raw score instead of on the IRT ability estimates.

Test administration differences between CAT and P&P tests also can affect score comparability. For example, in CATs, examinees often are not able to change answers. Reading long passages on a computer screen can be a different experience than reading them on paper. Test security, susceptibility to test preparation, computer experience of examinees, testing conditions, timing of the test, and other differences in administration conditions can all affect comparability.

Because some of the differences described above are inevitable features of CAT, comparability cannot simply be assumed but should be carefully established and evaluated. Consideration of these threats to comparability in the CAT context will help to drive the development of the framework in the present paper.

A general approach to achieving comparability is through the design of the CAT tests. It is typically done through a series of simulation studies at the early stages and some real examinee studies at later stages (e.g., Eignor, 1993; Eignor & Schaeffer, 1995; Eignor, Stocking, Way, & Steffen, 1993; Schaeffer, Reese, Steffen, McKinley, & Mills, 1993; Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995). The simulation studies are useful in designing the CAT and examining the various technical aspects of the tests. The real sample studies give feedback to the CAT design and makes final adjustment to improve comparability. The real sample studies involve administering P&P tests and CATs to examinees. For example, Segall (1995) described a study in which examinees were randomly assigned to take either a P&P or a CAT test. The CAT scores were related to the P&P scale scores using equipercentile methods. Eignor and Schaeffer (1995) described studies in which examinees took both a P&P and a CAT version of the GRE General Test. They concluded that the scale scores produced by the P&P and the CAT were sufficiently similar to one another that no further adjustment was needed for the CAT on the Verbal and Quantitative scores. However, scores on the Analytic test required adjustment based on the results of the study.

Comparability Issues Specific to CATs Based on Alternate CAT Pools

In high stakes standardized P&P testing, multiple test forms are needed to meet test security requirements. Similarly, high stakes CATs cannot meet security challenges with only a single static item pool. Three approaches can be taken to address this concern. One is to update and refresh the old item pool periodically (Stocking, 1988). A second approach is to build alternate item pools and replace the old item pool with a new item pool, or to allow an item pool to be randomly chosen from multiple item pools before CAT testing, as suggested by Stocking (1994). A third approach is simultaneously to use multiple item pools and rotate them among testing sites (Way, 1997). With any of these approaches the comparability issues need to be resolved to ensure fairness for examinees who happen to receive CAT administrations based on different item pools or different versions of the same pool. As in the case of comparability between the CAT and P&P versions, a commonly accepted level of comparability for alternate CAT pools is that the CAT scores should be comparable to each other to the extent that two typical P&P test forms are comparable to each other.

The approach of updating or refreshing an existing item pool seems more appropriate for small volume low stakes CAT programs, where item exposure

would not be a very big concern. When item pools are replaced or updated it is important for the psychometric properties of the resulting scores to be the same for the old and new pools. These properties can be studied through simulation and by randomly assigning examinees to be administered items from the old and new pools. Item pool updating is sometimes accomplished using on-line calibration procedures. In Stocking's (1988) procedure, the anchor items are used to rescale the item parameter estimates of the seed items from the initial on-line calibration. She also examined three methods of selecting candidate items and three methods of identifying suitable new items. Levine, Thomasson, and Williams (1991) studied several item pool replacement strategies and concluded that the replacement process did not affect the examinee marginal score distribution too much but did affect conditional moments when large replacement or random replacement were used. More research, however, needs to be devoted to the issues of on-line calibration and item pool updating.

The second and third approaches, both of which use multiple item pools, would be more viable for large-scale standardized CAT programs because the high testing volume would soon make a single item pool so overused that merely updating the pool would not be sufficient to solve the problem. It is reasonable to hypothesize that achieving comparability among CATs based on alternate item pools would be an easier task than achieving comparability between CAT and P&P versions of the same test. However, the complicated on-going feature of CAT has posed daunting challenges that would likely require years of new experiences to find satisfactory solutions.

A basic approach to achieving comparability for CAT based on alternate pools is to create item pools parallel to each other so that CAT scores based on them will automatically have comparable properties without further adjustment, because such adjustment would not be feasible in a realistic testing setting. Thus, the critical issue is how to create parallel item pools, which requires understanding the defining characteristics of item pools that most affect comparability of the CAT scores. An ultimate criterion for pool parallelism is achieving a satisfactory level of comparability among CAT scores based on these pools. There are a number of factors that may affect pool parallelism and CAT score comparability. Consideration of these factors often requires balancing comparability and test security requirements. These factors include (a) item pool size, (b) the number of pools that are used simultaneously, (c) item pool overlap rate, (d) item pool assembly procedures and constraints used to assemble the pools, (e) item pool rotation scheme and criteria to retire an item pool, (f) rules for item reuse, (g) item pretesting schemes and item calibration procedures and procedures to control scale drift, and (h) CAT item selection procedures to meet content specification, to control item exposure rates and test overlapping rates, and so forth. These factors interact with each other and affect comparability and other properties of the resulting CAT test scores in a complex way.

A general characteristic of the current literature relating to these issues is that they primarily emphasize test security rather than comparability. Although it may be understandable that test security has been a driving force behind complicated item pool management strategies, the subsequent comparability issues must be

tackled in a careful and systematic manner. Given the complex nature of the present day CAT test designs and item pool management strategies, comparability among CATs based on alternate pools cannot simply be assumed but requires extensive studies and must be monitored in CAT operational administrations. Following is a brief discussion of literature relating to these factors and how they might affect comparability.

Stocking (1994) considered extensively the issue of item pool size, the number of pools, and the use of overlapping pools. A basic goal for developing item pools is to contain sufficient quantities of items of different aspects to support the CAT administrations with a desired level of properties. Stocking (1994) proposed a simulation-based method in determining the sufficient pool size to meet the content and statistical requirement of the CAT test. Based on her simulation results, she concluded as a rule of thumb that a pool size of six to eight typical linear forms should be used when the length of the CAT tests is about one-half that of the linear forms. Stocking did not consider the comparability requirement when considering sufficient pool size, but her approach can be adapted to decide on the sufficient pool size not only to maintain those properties she considered, but also to achieve desired level of comparability among CAT based on alternate item pools. It is reasonable to assume that the larger the pool size, the more likely it is that comparability can be achieved. The question is how large is large enough.

To generate multiple item pools requires the availability of a large quantity of items, referred to as an item "vat" by some researchers (Way, 1997; Way, Steffen, & Anderson, 1998). From this item vat, item pools can be assembled using some automated assembly procedures similar to those test form assembly procedures (van der Linden, 1998). Stocking and Swanson (1996) proposed a heuristic for constructing overlapping pools based on their weighted deviation model which was formerly used to construct conventional test forms and select items in CAT administration to meet different constraints (Stocking & Swanson, 1993; Swanson & Stocking, 1993). This heuristic is flexible in combining different dimensions of constraints such as content specification and statistical considerations. Other algorithms used for item pool assembly that can also accommodate a large set of constraints include linear programming techniques (van der Linden, 1998). The issues of how parallelism of item pools affects comparability of CAT scores has not received much attention in the literature. There is a need to identify the most influential characteristics of item pools that operationally defines the concept of parallelism of pools and to study what level of parallelism is needed in order to achieve a desired level of comparability of CAT scores.

Way (1997) considered schemes for continuously generating new pools and rotating these pools along with a process of pretesting and calibration new items. The calibrated new items are replenished into the item vat from which new item pools are generated. These schemes are used to support a continuous process of CAT administration without compromising test security. The sufficiency of the item vat itself in terms of both its quantity and quality is critical to support such a process. Way, Steffen, and Anderson (1998) used system dynamic methods to monitor and predict the long term trend of the item vat given a complicated set of rules for temporarily and permanently retiring items based on their usage.

Item pretesting and on-line calibration pose yet another enormous and unique challenge for maintaining long-term comparability for an on-going CAT administration process. Proper item pretesting and calibration may be more crucial for CAT than for conventional P&P testing because CAT relies heavily on the availability of high quality item parameter estimates. In P&P testing there is usually an equating stage after the test forms are constructed based on item pretesting statistics, which lessens the burden of having very accurate pretest data. For that reason, item pretesting schemes in P&P testing are often designed to have the least amount of interference with the operational part of the test, such as including a separate section of pretesting items. In CAT, however, calibrated item parameters based on pretest data are used operationally without another equating stage. Any systematic distortion in the item statistics is likely to cause the ability scale to drift away from its original metric, and thus undermine comparability. Another unique feature of CAT is its continuous nature. Although continuous testing lends flexibility for designing creative pretesting schemes, it also poses challenges because it may be hard to control the pretesting examinee groups and it may give less time for item review. It is also possible that, when not promptly discovered, poorly functioning items might disrupt examinee performance on operational items. It is sometimes proposed that items should be pretested with a small number of examinees in a P&P mode before being pretested on-line.

Several methods have been proposed for on-line calibration (Wainer & Mislevy, 1990; Stocking 1988a; Levine & Krass, 1998). The relative advantages and applicability of methods are still largely unknown and need to be further studied along with the issue of scale drift. Bock, Muraki, and Pfeiffenberger (1988) found drift for item location parameters for the College Board Physics Achievement Test but not for the discrimination parameters. They proposed statistical procedures to estimate and compensate for the scale drift. Stocking (1988b) compared two on-line calibration procedures (see previous descriptions) for scale drift. She found that linking through examinee ability estimates from the operational items had some negative effects. She also found that scale drift was in opposite direction for the two methods. With method A, slopes tended to be underestimated. With method B, slopes were better estimated but were also overestimated. Stocking also made some suggestions to improve the scale drift problem. Further research is needed to find the most effective procedure for detecting scale drift and to make appropriate adjustments.

Item selection rules in a CAT test design are the engine for controlling the properties of CAT tests and are also essential for achieving desirable level of comparability. Item exposure rates were traditionally used as the primary index for monitoring item security; various algorithms have been proposed to control the maximum item exposure rates with or without conditioning on ability levels (Sympson & Hetter, 1985; Stocking & Lewis, 1995a; Stocking & Lewis, 1995b; Davey & Nering, 1998). Way (1997) emphasized that absolute item exposure frequency is also important to monitor. In addition, test overlap rates, which consider all possible pair-wise item overlap rates between examinees, also should be controlled to ensure test security. A recent paper by Chen, Ankenmann, and Spray (1999) found that when other conditions (such as item pool size and test

length) are fixed, high variability of the individual item exposure rates can lead to high test overlap rates. In other words, equalizing item exposure rate in a pool can reduce average test overlap rate. The impact of the CAT item selection algorithms on comparability should be studied in combination with various item pool management strategies.

Criteria for Evaluating Comparability

A synthesis of the criteria used by previous researchers in evaluating comparability led us to consider three general categories of criteria: The first category is the validity criterion, which requires the CAT and P&P versions to measure the same construct; the second is the psychometric property/reliability criterion, which requires the CAT to have the same psychometric properties (e.g., reliability) as its P&P version; the third is referred to as the statistical assumption/test administration condition criterion, and includes whether or not the assumptions used to establish comparability actually hold and whether or not the operational test condition matches the comparability study testing condition. These criteria are described and discussed in the following sections in terms of their definition, their significance, their relationships to each other, how to evaluate comparability based on these criteria, and how to improve comparability under specific criteria.

The Validity Criterion

Traditionally, three aspects of validity evidence have been proposed to validate test scores. They are content-related evidence, criterion-related evidence, and construct-related evidence. More recently, validity has been considered to be a unitary concept with construct validity being the core of all aspects of validity evidence. Messick (1993) also added the consequential facet to the traditional evidential facet of the test score interpretation and use. Based on the theoretical concepts and the special context of CAT comparability issues, our discussion on the validity criterion for comparability will be divided into three parts: (1) content specifications, (2) dimensionality, and (3) relationship across modes with other variables, including subgroup differences. The three aspects are discussed separately below.

Content Specifications. Content specifications are vital, particularly in educational achievement testing. In P&P standardized testing, alternative test forms are usually constructed according to a detailed table of content specifications. Recent research and implementation of CAT tests has also incorporated such a table of specifications (e.g., Eignor & Schaeffer, 1995; Eignor, 1993; Eignor, Stocking, Way & Steffen, 1993). In realistic testing settings, considerations should also be given to less explicit specification such as balancing keys, choosing passage topics and balancing references to gender, ethnicity and other background subject. Two algorithms have been proposed to achieve this task in CAT. One algorithm was proposed by Kingsbury and Zara (1991) and gives first priority to content specifications in item selection. This algorithm is equivalent to partitioning the item pool according to different content areas and selecting the item with maximum information from each partition in a spiral fashion. The advantage of this algorithm is that it

ensures that a strict table of specifications is imposed. The disadvantage is that it sacrifices measurement precision because the balancing of the content is imposed at every stage of item selection.

Stocking and Swanson (1993) proposed a heuristic algorithm that treats the content specifications as a target rather than as a strict requirement. In our implementation of this weighted deviation algorithm an item's contribution is a weighted sum of its item information and its capacity to meet each of the content specifications. Items are selected based on their contribution values. The weights assigned to the content constraints are based on the importance of the constraints. Because information and content deviation are not on the same scale, the weights assigned to information need to be determined through simulations. The main advantage of this algorithm is that a large number of constraints besides content specifications can be incorporated in this weighted deviation algorithm. Another advantage is that the weights that determine the priority placed on content specifications and on measurement precision can be assigned flexibly. The disadvantage of this algorithm is the uncertainty in balancing the content specifications.

Van der Linden and Reese (1998) adopted the linear programming techniques in CAT item selection to maximize test information subject to a number of constraints. Their algorithm guarantees that given a feasible solution to the linear programming problem, all the constraints will always be met. Van der Linden (1998) also gave a very good introduction to other test assembly algorithms that can potentially be applied to CAT item selection.

Dimensionality. The dimensionality aspect of comparability of CATs refers to whether the CAT measures the same construct dimensions as the P&P version. This criterion is closely related to the content specification criterion. Balancing content is important for maintaining the same construct dimensions. Although content validity is important for ensuring validity in educational measurements, the fact that with CAT different examinees receive different sets of items can cause complications. For some subject areas, such as mathematics, the content areas represent a hierarchy of difficulty levels because some content areas may be the prerequisite for learning the other content areas. CATs purport to target the measurement at examinees' ability levels and thus, by nature, possibly should not be restricted to have exactly the same content specifications as the conventional tests which target all ability levels with a single form. Moreover, as Davey and Thomas (1996) pointed out, because CAT often uses θ estimates as a basis for score transformation, the contribution of different content areas to the ability estimates might not be in the same proportion as in conventional testing, even if the same table of specifications is imposed.

The crucial issue behind this dilemma is the issue of dimensionality. If the item pool is strictly unidimensional, then there is no need to balance the content specifications, at least from a construct validity point of view. But, in reality, different content areas within a test often represent different but highly correlated dimensions. When a unidimensional IRT model is used for calibrating the items for a test these correlated dimensions are represented in a composite dimension. Reckase (1979) and Folk and Green (1989) demonstrated that the composite

dimension may not represent all the sub dimensions in a balanced way, and that the imbalance could be worsened when a simple maximum information item selection algorithm is used. Thus, an item selection algorithm that incorporates content specifications might help to recover the intended dimensionality representation in the reported scores. Evaluation of comparability in terms of dimensionality between the CAT and P&P tests can help to assess the appropriate priority or weights that are given to the content specification or to the measurement precision.

There are three categories of methodologies that can be used to assess dimensionality in CAT. The first category is IRT-based methodology, the second category is structural equation modeling-based techniques, and the third category is factor analytical methods. Different IRT-based techniques have been proposed to assess dimensionality in fixed-item testing settings (e.g., Stout et al., 1996). Assessing dimensionality directly for CATs has yet to be researched. Structural equation methods and techniques can be used to assess the structural relationship among tests. Cudeck (1985) provided an example of comparison of the structural relationship of a CAT battery and a conventional battery for the ASVAB test. His results showed that the CAT-ASVAB maintains a comparable construct structure as its conventional versions. Henly, Klebe, McBride, and Cudeck (1989) used a factor analytical model and multitrait-multimethod procedure and confirmed the structural comparability of the CAT version of the Differential Aptitude Test (DAT) to its P&P version.

Alternative approaches have been proposed and studied to enhance comparability in terms of dimensionality representation. Davey and Thomas (1996) proposed a criterion that forces the proportion of information contributed by each content area to be the same for both tests. How effective this algorithm is in achieving greater comparability needs empirical investigation. Segall (1996) and Luecht (1996) explored the use of multidimensional IRT models as the psychometric foundations for CAT. This approach could allow for better control of the dimensionality of the CATs and thus enhance comparability in this respect.

Relationship Across Modes and With Other Variables, Including Subgroup Differences. A useful way of assessing comparability is to evaluate the relationship between the CAT & P&P versions of the same test. This relationship can be assessed by simply examining the correlations adjusted for attenuation, or by more sophisticated techniques such as structural equation modeling. Another important criterion for the comparability in terms of validity is whether a CAT test battery maintains the same structural relationship with other tests in the battery and the same predictive and concurrent validity with other related measures. Cudeck (1985) and Henly, Klebe, McBride, and Cudeck (1989) also investigated these relationships in validating the CAT versions of the ASVAB and DAT.

Another aspect of this criterion pertains to subgroup differences. Subgroup differences based on CAT scores should be compared to those of the conventional versions. Kolen (1999-2000) pointed out that test scores from the different modes may be comparable for one group but not for another. For example, the CAT version might be comparable to its P&P version for examinees with considerable computer experience, but not for examinees with little experience. The study design

should take into account individual difference variables such as test anxiety, computer anxiety, computer experience, and so forth (e.g., Wise, Roos, Plake, & Nebelsick-Gullett, 1994; Vispoel, Rocklin, & Wang, 1994). Gender- and ethnicity-related differences should also be studied. Eignor and Schaeffer (1995) examined this issue in their two comparability studies involving GRE CAT and National Council Licensure Examinations (NCLEX for registered nurses and practical/vocational nurses) and found some differences for various ethnic groups. Segall (1995) also examined gender and ethnic group differences with the CAT-ASVAB and found that female examinees were slightly disadvantaged with the CAT version on certain subtests and Black examinees were slightly advantaged with the CAT version on certain subtests.

The Psychometric Property/Reliability Criterion

The psychometric property/reliability criterion refers to whether the two versions of the test have the same psychometric properties such as conditional standard errors of measurement (CSEM) and reliability. This criterion can be divided into two components: conditional properties and overall properties. The conditional properties are based on the equity criterion defined by Lord (1980). The overall properties are the same score distribution criterion and the same overall reliability criterion. The consideration of psychometric properties in the present paper is based on scale scores, which are the scores reported to examinees. For example, with the ACT Assessment Mathematics test, scale scores are integers that range from 1 to 36. Number-correct scores or estimated IRT abilities (θ) would be converted to scale scores. Some authors (Kolen, Hanson, & Brennan, 1992; Kolen, Zeng, & Hanson, 1996; Wang, Kolen, & Harris, 2000) have emphasized examining psychometric properties of scale scores rather than of raw scores.

The Equity Criterion. The equity criterion requires that the conditional scale score distributions be the same for two test forms conditioned at any ability level. In practice, equity is often assessed at two less demanding levels: first-order equity (also called weak equity), which requires that an individual at any ability level be expected to earn the same score on both forms, and second-order equity (also called equal precision), which requires that examinees at a given ability level be measured with the same precision on the two test forms.

More explicitly, first order equity can be expressed as

$$E(s_1|\theta) = E(s_2|\theta), \text{ for all } \theta, \tag{1}$$

where E refers to expected value of scale scores over examinees of a given ability, θ, and s_1 is the scale score on Test 1 and s_2 is the scale score on Test 2. Second order equity can be expressed as

$$\sigma(s_1|\theta) = \sigma(s_2|\theta), \text{ for all } \theta, \tag{2}$$

where σ is the standard deviation of scale scores over examinees of a given ability, θ (conditional standard error of measurement, CSEM).

First order equity might be achieved to some extent by eliminating the bias in the ability estimates. The maximum likelihood estimate (MLE) in CAT is relatively unbiased over a range of ability for well-designed item pools (Wang, 1995). But the MLE has relatively a large standard error and could also be biased if the item pool lacks a sufficient number of items of extreme difficulty levels. Bayesian estimates such as the expected a posteriori (EAP) estimates are usually seriously biased toward the prior mean if a standard normal prior is given. The example in this paper uses relatively unbiased EAP estimates with a flat beta prior distribution (Wang, 1997; Wang, Hanson, & Lau, 1999). However, eliminating the bias in θ estimates will not automatically guarantee first order equity of the CAT and P&P versions on the reported score scale. This criterion must be evaluated empirically using simulation methods.

The second order equity criterion requires a CAT to have the same conditional standard error of measurement (CSEM) as its conventional counterpart. This criterion can be satisfied by varying the termination rules, or changing the characteristics of the item pool so that the resulting information curve will approximate that of the conventional test. The Stocking and Swanson (1993) algorithm can be used to achieve in part the target information curve by varying the weights in the weighted deviation model.

One additional related equity criterion applies for tests with a passing score or cut scores. This criterion is called *equal probabilities of achieving passing scores.* That is, given passing score s_c, examinees of a given true level on the construct should have the same probability on Test 1 of meeting or exceeding score s_c as they do on Test 2, so that

$$Pr(s_1 \geq s_c|\theta) = Pr(s_2 \geq s_c|\theta), \text{ for all } \theta. \tag{3}$$

This property should hold for all passing scores, s_c ($c = 1, \ldots, C$), that are used with the test.

The Overall Criterion. This criterion has two components: the same score distribution criterion and the same reliability criterion. The former requires that the marginal distribution of the reported scores derived from the two test versions be the same for the examinee population of interest. This criterion requires, for example, that the two score distributions have the same means and standard deviations. Mazzeo and Harvey (1988), Mead and Drasgow (1993), and Bergstrom (1992) provided literature reviews and a meta-analysis on the equivalency of the test scores derived from the two versions. Their major concerns were with mean differences. Also, for tests with a passing score or cut score, it is required that the proportion of examinees above the passing score will be the same for the two tests or modes.

The same reliability criterion requires that both test versions have the same overall reliability as defined in classical test theory. In theory, if the conditional equity criterion is met perfectly, the overall criterion will be met automatically. Because in practice the equity criterion is never achieved perfectly, this overall criterion provides a valuable overall check on the psychometric aspect of the comparability.

The Statistical Assumption/Test Administration Criterion

The test administration criterion refers to whether or not the assumptions that are used to establish comparability actually hold. The assumptions may be required for the data collection design or for the statistical analyses of the test scores. For instance, if a random groups design is used to collect data, then any violation of random assignment may cause score incomparability. If a single group design is used, then it is important to separate the order effect from other effects unless it is safe to assume no practice effect.

The IRT assumptions that are required with the IRT analyses in CAT are also important aspects of the criterion. Among the most important assumptions are unidimensionality and local independence. Data-model fit statistics provide a good way of assessing whether the assumptions hold (Hambleton & Swaminathan, 1985).

The test administrative criterion involves the equivalence of the two modes of item presentation. Kolen (1999-2000) described five aspects of possible differences in the presentation of test items in the two modes as follows: (a) ease of reading lengthy passages, (b) ease of reviewing or changing answers to previous questions, (c) speed in taking the test, and the effects of time limits on test speededness, (d) clarity of figures and diagrams, and (e) responding on a keyboard and/or mouse versus responding on an answer sheet. The effects of some of these aspects have been studied. Spray, Ackerman, Reckase, and Carlson (1989) studied item presentation mode effect on item characteristics. Combined with previous research, they concluded that, if there is sufficient equivalency in item-taking flexibility, such as review and change of answers, and it is easy to move from item to item, then the mode difference of CAT and P&P produces no real difference in item characteristics. But if the test-taking flexibility was different, then the items might function differently. Some aspects of the differences such as speededness and ease of reading lengthy passages, and clarity of figures and diagrams have not been thoroughly researched. Hetter, Segall, and Bloxom's (1994) more recent study supported the interchange in the use of item parameter estimates calibrated from either P&P or computer administered data for CAT uses. Another factor that may affect the functioning of the items is item location. Kingston and Dorans (1984) studied the item location effect and concluded that some item types are more susceptible to item location effects than other types. In practice, item parameter estimates are often calibrated based on P&P test data and used for CAT administration. How computer administrations affect the item parameter values is thus important to discover and control. Once the comparability study is carried out and comparability carefully established, it is important to ensure the operational testing conditions be as close to those in the comparability study as possible. Otherwise, new studies are needed to re-establish comparability.

Other issues under this category include test security, test preparation, and test disclosure. These issues are important to ensure the fairness and validity of the test. Here, we will only discuss one important issue that is related to test security: item exposure rate control. There are several algorithms for controlling item exposure rate in CAT. One algorithm implemented by some early researchers used a simple randomization method, that is, at each stage a set of optimal items was selected

instead of a single item, and then an item was drawn randomly from this set for actual administration. The Sympson and Hetter (1985) algorithm uses a conditional probability of administering an item given that an item is selected as candidate to control the marginal (actual) exposure rate. The conditional probability, which is referred to as the exposure rate control parameter, needs to be estimated for each item through a series of simulations. The Sympson and Hetter algorithm is statistically elegant and easy to implement. One disadvantage of this algorithm is that there is no guarantee that a certain number of items can be administered to an examinee unless there are at least that number of items that have exposure control parameters equal to one, which means that an item will definitely be administered if it is selected. Stocking and Lewis (1995a) remodeled this procedure with a multinomial distribution and introduced an adjustment to the cell probabilities to ensure that there are always enough items to be administered. Both of these algorithms control the exposure rate for a population of examinees. An item with an exposure rate under some level, say 20%, may still have a high exposure rate for a group of examinees with similar ability level. To overcome this problem, Stocking and Lewis (1995b) again proposed a conditional multinomial algorithm to control item exposure rate conditioned at all ability levels. Davey and Parshall (1996) proposed a variation of the Sympson and Hetter algorithm to control item exposure rate conditioned on the administration of other items. But this algorithm is difficult to implement in practice.

There is a general restriction in controlling the item exposure rate, that is, the upper limit of the exposure rate multiplied by the item pool size should be greater or equal to the CAT test length. This means that the exposure rate can not be controlled to an arbitrarily low level for a given item pool.

An Example Using ACT Mathematics Items

In the early stages of building a CAT that is comparable to its P&P version, psychometric analyses can be conducted that help to set various design components, such as test length and exposure control, for the CAT. This example illustrates how computer simulation techniques might be used to design the features of this CAT test, and a series of comparisons between the CAT and P&P version are carried out. The comparison focuses on the psychometric properties/reliability criterion, but content validity and test security criteria are also addressed. In addition, this section also illustrates how various decisions that are made in defining the CAT can affect the comparability of the scores on the CAT to the scores on the P&P test. A simplified situation is chosen here, in which a pool of test questions that already exist in P&P forms are used to develop a pool of items for a CAT pool. Only fairly simple content-balancing and exposure control procedures are used. In addition, to examine the effects of pool size, the pools were randomly divided in half for convenience. Much more extensive procedures would likely be necessary if an operational CAT ACT Mathematics test were to be developed to be comparable to a P&P ACT Mathematics test.

Methodology

Comparability between the CAT version and its conventional version can be studied with simulation methods and real data methods. Simulation methods usually are used at the pilot and developmental stage that sets the basic design of the CAT version. Real data studies provide final checks and adjustments on the CAT version that have been developed and tested with simulations. These approaches are exemplified by the development and implementation of the GRE General test (Eignor, Stocking, Way, & Steffen, 1993; Eignor & Schaeffer, 1995). Simulation studies typically are used heavily to assess the psychometric properties and content validity, particularly the conditional properties of comparability, whereas real sample studies are primarily used to assess the overall reliability and other validity aspects of comparability. This illustrative study used the simulation method.

Data. ACT Assessment Mathematics forms were used a basis for the simulation. Each P&P form contains 60 multiple-choice items. The P&P forms are scored based on the number-correct raw scores which are converted to the scale scores that range from 1 to 36. Response data for seven P&P forms of the ACT Assessment Mathematics test collected using a random groups design were calibrated using the calibration program EM1 (Zeng, 1995). This program uses the EM/MML algorithm in estimating the item parameters. Altogether 420 items were calibrated to construct a CAT pool. The descriptive statistics are presented in Table 1. The items belong to six different content categories of the table of specifications. The six content categories are Pre-Algebra, Elementary Algebra, Plane Geometry, Coordinate Geometry, Intermediate Algebra, and Trigonometry.

Raw-to-Scale Score Conversions. To make the comparisons more meaningful, scores were translated to the ACT Assessment score scale. Raw-to-scale score conversions exist from the process used to equate the P&P forms of the ACT Assessment. To compute comparable ACT scale scores from a CAT test, it is necessary to have a process in place for converting IRT abilities to ACT scale scores. The procedure that was chosen was to conduct this mapping using a process closely related to IRT true score equating (Kolen & Brennan, 1995). One of the seven forms was chosen as the base form. The number-correct raw-to-scale score conversion for this form was used.

Assuming that the IRT model holds, the true number-correct score on the base form associated with a given θ can be found from the test characteristics curve as follows:

$$\tau(\theta) = \sum_{g=1}^{k} P_g(\theta), \tag{4}$$

where k is the number of items and $P_g(\theta)$ is the probability of correctly answering item g given θ. The true scale score was defined as

$$\tau_s(\theta) = s[\tau(\theta)], \tag{5}$$

TABLE 1
Descriptive Statistics for the Item Parameter Estimates in the Item Pool
a. For the 420-item pool

Parameter	a	b	c
Mean	0.965	0.183	0.150
Median	0.955	0.261	0.152
Standard Deviation	0.289	0.966	0.047
Variance	0.083	0.933	0.002
Kurtosis	−0.334	−0.293	−0.275
Skewness	0.166	−0.361	0.050
Minimum	0.296	−3.099	0.031
Maximum	1.933	2.582	0.282
Count	420	420	420

b. For pool a

Parameter	a	b	c
Mean	0.963	0.181	0.149
Median	0.949	0.306	0.153
Standard Deviation	0.303	0.986	0.045
Variance	0.092	0.973	0.002
Kurtosis	−0.326	−0.087	−0.031
Skewness	0.177	−0.490	0.085
Minimum	0.324	−3.099	0.041
Maximum	1.933	2.582	0.282
Count	210	210	210

c. For pool b

Parameter	a	b	c
Mean	0.966	0.184	0.151
Median	0.958	0.246	0.151
Standard Deviation	0.275	0.947	0.048
Variance	0.076	0.897	0.002
Kurtosis	−0.378	−0.524	−0.459
Skewness	0.155	−0.216	0.013
Minimum	0.296	−2.185	0.031
Maximum	1.681	2.199	0.256
Count	210	210	210

where s is the raw-to-scale score conversion for number-correct scores. Because true number-correct scores from Equation 4 are non-integer scores, linear interpolation was used to find the conversion using Equation 5.

Through these equations, number-correct true scores, true scale scores, and θ are all mapped into one another. That is, if θ is known, then through Equations 4 and 5 the associated number-correct true scores and number-correct true scale scores on a

particular test form are also known. In the present study, estimated IRT abilities were converted to ACT scale scores by substituting the estimated abilities for θ in Equation 5.

Simulation Procedures. A CAT simulation system was programmed in the *C* language and was used in this study. The content specifications for the P&P version were taken as the goal and were weighted against item information in the CAT item selection algorithm. Based on the research literature on CAT and experience from other CAT testing programs (e.g., the GRE and SAT programs, see Eignor, Stocking, Way, & Steffen, 1993), it was expected that the CAT test would need to be about half as long as the conventional form to achieve a similar level of precision. So a fixed test length of 30 items was tentatively used at the developmental stage. Simulations were carried out conditionally on 17 points on the θ scale from –3.2 to 3.2 in increments of .4. A total of 400 replications were simulated at each of the θ points for each simulation condition.

For the CAT, the EAP estimates, which use a beta prior to reduce bias (Wang, 1997), were used for both the provisional estimates and the final estimates of θ. The θ estimates were transformed to the number-correct true scores on the base form and scale scores using the procedures already described and associated with Equations 4 and 5.

For simulating the P&P test scores for comparison, the base form was used. The simulation of this test was the same as that of the CAT test except that it always had the same 60 items in their original sequence. The number-correct raw scores were converted to scale scores using the conversion table for the base form.

The simulation procedure was conducted in several steps. First, a pure CAT test without content balancing and item exposure rate control was simulated and compared to the conventional version. Second, the Stocking and Swanson (1993) algorithm with the weighted deviation model was used to select items based on content specifications of six content categories. The weights between information and content balancing were changed to achieve a satisfactory level of content balancing while still giving information as much weight as possible. Note that in a more realistic setting, more detailed content specifications would be required and content experts would be called in to review the resulting CAT tests from simulation. In this illustrative example, we only used the simplest content specifications. Third, item exposure rate control was introduced with the Sympson and Hetter (1985) algorithm. A series of simulations were carried out to estimate the exposure rate control parameter for each item. The goal was to control the exposure rate at under 15%; that is, no item should be administered to more than 15% of the examinee population. Fourth, test length was varied from 30 items to 20 items to examine the effect of this reduced test length. Fifth, the item exposure rate was varied from 15% to 10% to examine the effect of more strict exposure rate control. Finally, the 420-item pool was divided into two randomly equivalent item pools, each having 210 items, to examine the effect of the reduced item pool size and to assess the comparability between these randomly equivalent CAT pools. The simulation conditions are summarized in Table 2.

TABLE 2
Simulation Conditions

Simulation Condition	Administration	Scoring	Test Length	Content-Balancing	Exposure Control	Pool Size
P&P NC based	P&P (Fixed)	Number-Correct	60	N/A*	N/A	N/A
Pure CAT	Adaptive	EAP θ	30	None	None	420
cat30i15e	Adaptive	EAP θ	30	Yes	15%	420
cat20i15e	Adaptive	EAP θ	20	Yes	15%	420
cat30il0e	Adaptive	EAP θ	30	Yes	10%	420
pool a	Adaptive	EAP θ	30	None	None	210
pool b	Adaptive	EAP θ	30	None	None	210

*N/A indicates not applicable

Four aspects of comparability were examined that pertain to the psychometric properties/reliability criterion, all based on the simulated data. First, for each replication of the procedure, CAT and P&P EAP θ estimates ($\hat{\theta}$) were calculated. Over the 400 replications conditional on θ, the $bias_\theta$, standard error (se_θ), and root mean-squared error ($rmse_\theta$) were calculated as follows:

$$bias_\theta = \sum(\hat{\theta} - \theta)/400, \tag{6}$$

$$se_\theta = \sqrt{\sum(\hat{\theta} - \bar{\hat{\theta}})^2/400}, \text{ and} \tag{7}$$

$$rmse_\theta = \sqrt{\sum(\hat{\theta} - \theta)^2/400}, \tag{8}$$

where all of the summations are over the 400 replications and $\bar{\hat{\theta}}$ is the mean ability estimate over the 400 replications. For each condition, each of these indices were calculated for the P&P and CAT estimated abilities.

Second, statistics were calculated to evaluate first and second-order equity of scale scores. For each condition of the simulation, number-correct scores on the P&P test, the EAP θ estimates on the P&P test, and EAP θ estimates on the CAT were converted to scale scores (s). Over the 400 replications and conditional on θ, the mean scale score for the number-correct P&P tests, designated $\bar{s}_{\theta P\&P}$, was calculated. The mean scale score for each of the other estimation methods (e.g., EAP θ estimates on the P&P test) over the 400 replications was also calculated at each θ, and designated as $\bar{s}_{\theta\, method}$. The quantity,

$$Sdiff_\theta = \bar{s}_{\theta\, method} - \bar{s}_{\theta\, P\&P}, \tag{9}$$

was calculated at each θ for each method and was used as an index of first-order equity. If first-order equity holds, the value of this index would be approximately zero at each θ.

Conditional on θ, the standard deviation of the scale scores for each of the methods (including the number-correct P&P method), which represents the condi-

tional scale score standard error of measurement for that method, was calculated as

$$CSEMSS_\theta = \sqrt{\sum(s_\theta - \bar{s}_\theta)^2/400}, \quad (10)$$

where the summation is over the 400 replications and $\bar{s}_\theta$ is the mean scale score over the 400 replications for a particular method, conditional on θ. If second order-equity holds, then at each θ the value of $CSEMSS_\theta$ would equal the value of $CSEMSS_\theta$ for the P&P number-correct score method.

Third, based on standard normal population distribution of θ, the frequency distribution and the cumulative distribution of the scale scores were compared across methods. Finally, proportions of examinees in the standard normal population distribution that exceed certain cut scores were examined and compared.

Results

The results are divided into sections. The first two sections provide the results of content balancing and item exposure rate control. The third section compares the CATs and P&P tests on the θ metric. The fourth section examines first and second order equity for scale scores. The fifth section compares scale score distributions. The sixth section focuses on comparisons among the CAT variations. The seventh section focuses on the two half-size item pools. The eighth section examines some practical consequences of the differences among different methods by comparing the proportions of examinees in the population that exceed certain cut scores.

Content Balancing. Table 3 presents the results from implementing the Stocking and Swanson algorithm for content balancing. The goal for each of the six content categories for a 30-item test according to the proportional content representation in the table of specifications for the P&P version is 7, 5, 4.5, 4.5, 7, and 2 items, respectively. With the weights chosen, the resulting mean numbers of items administered from each content category meet the goal quite well. The results show that the Stocking and Swanson (1993) algorithm works well in this one-dimensional content balancing task.

Item Exposure Rate Control. Table 4 contains results from using the Sympson and Hetter (1985) algorithm for item exposure rate control. The goal was to limit the item exposure rate under 15% for all the items. A series of simulations was carried out to find the item selection rate and to estimate the item exposure control parameter for each item. The results show that the goal was reached quite well. The exposure rate for all the items are under 16%. It can be seen from Table 4 that before item exposure control, more than half of the items in the pool were never administered; after exposure control, that number dropped to 113 items.

Comparison Among the CATs and P&P Test on the θ Metric. Figure 1 provides plots of the indice on the θ-metric that were defined in Equations 6 through 8. Figures 1a and 1d present results for the $bias_\theta$ statistic. Two figures were needed to plot the results for all of the methods and conditions. Note that the cat30i15e condition is repeated in the two plots to aid in making comparisons across plots.

TABLE 3
The content Proportions Conditioned on Each Theta Point
a. Before Content Balancing (pure CAT)

Content(target)	1(7)		2(5)		3(4.5)		4(4.5)		5(7)		6(2)	
Theta	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
−3.2	13.97	1.48	6.33	0.62	2.75	1.05	1.01	0.21	5.94	1.13	0.00	0.00
−2.8	13.74	1.81	6.51	0.69	2.39	1.10	0.95	0.27	6.41	1.41	0.00	0.00
−2.4	13.46	2.04	6.61	0.78	2.20	1.26	0.88	0.36	6.84	1.41	0.00	0.00
−2.0	11.97	2.29	6.61	0.88	3.04	1.77	0.53	0.51	7.86	1.38	0.00	0.00
−1.6	9.80	2.37	6.55	1.03	4.60	1.64	0.17	0.39	8.87	1.30	0.00	0.00
−1.2	6.97	2.37	7.58	1.80	5.75	0.80	0.09	0.32	9.61	1.01	0.00	0.00
−0.8	4.27	1.68	9.56	1.59	5.77	0.70	0.52	0.78	9.88	1.17	0.01	0.09
−0.4	3.40	0.84	8.65	1.27	4.49	1.06	2.09	1.07	11.26	1.22	0.11	0.31
0.0	2.27	0.86	7.44	0.96	3.46	0.76	3.23	0.95	13.28	1.78	0.33	0.49
0.4	1.04	0.61	6.22	0.98	4.01	0.82	2.90	1.04	14.91	1.92	0.92	1.07
0.8	1.73	0.92	5.71	0.81	3.92	0.82	4.91	1.59	10.20	3.10	3.55	1.60
1.2	3.04	0.95	5.48	0.68	4.22	1.37	6.75	0.64	5.10	2.54	5.41	0.90
1.6	3.08	0.63	4.75	1.18	6.68	1.45	6.85	0.54	2.87	1.34	5.76	0.87
2.0	2.70	0.55	3.02	1.10	8.09	0.54	6.47	0.78	3.08	1.08	6.64	0.98
2.4	2.29	0.51	3.10	0.66	8.31	0.51	6.02	0.56	4.43	0.84	5.84	0.96
2.8	2.19	0.50	3.36	0.58	8.28	0.48	6.08	0.53	4.96	0.29	5.12	0.57
3.2	2.04	0.63	3.60	0.73	8.33	0.51	6.13	0.64	5.01	0.17	4.89	0.41

b. After Content Balancing (CAT30i15e)

Content(target)	1(7)		2(5)		3(4.5)		4(4.5)		5(7)		6(2)	
Theta	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
−3.2	7.00	0.00	5.00	0.00	5.00	0.00	5.00	0.00	7.00	0.00	1.00	0.00
−2.8	7.00	0.00	5.00	0.00	5.00	0.00	5.00	0.00	7.00	0.00	1.00	0.00
−2.4	7.00	0.00	5.00	0.00	5.00	0.00	5.00	0.07	7.00	0.00	1.00	0.07
−2.0	7.00	0.00	5.00	0.00	5.00	0.00	5.00	0.07	7.00	0.00	1.00	0.07
−1.6	7.00	0.00	5.00	0.07	5.00	0.00	4.97	0.16	7.00	0.00	1.02	0.15
−1.2	7.00	0.00	5.01	0.12	4.99	0.09	4.93	0.26	7.00	0.07	1.05	0.23
−0.8	7.00	0.05	5.05	0.23	4.97	0.16	4.82	0.40	7.01	0.10	1.15	0.35
−0.4	6.92	0.27	5.05	0.22	4.95	0.23	4.71	0.47	7.01	0.10	1.36	0.48
0.0	6.55	0.50	5.02	0.19	4.89	0.34	4.82	0.40	7.04	0.21	1.68	0.47
0.4	6.21	0.40	4.98	0.22	4.84	0.40	4.97	0.17	7.05	0.30	1.95	0.28
0.8	6.12	0.33	4.94	0.29	4.85	0.40	5.00	0.11	6.86	0.36	2.23	0.46
1.2	6.12	0.32	4.95	0.26	4.95	0.27	5.02	0.14	6.26	0.44	2.71	0.53
1.6	6.33	0.47	4.96	0.23	5.06	0.26	5.01	0.10	6.04	0.18	2.60	0.53
2.0	6.30	0.46	4.98	0.18	5.16	0.38	5.00	0.00	6.20	0.40	2.35	0.48
2.4	6.46	0.50	5.00	0.00	5.08	0.27	5.00	0.00	6.28	0.45	2.19	0.39
2.8	6.62	0.49	5.00	0.00	5.02	0.13	5.00	0.00	6.32	0.47	2.04	0.21
3.2	6.68	0.47	5.00	0.00	5.00	0.00	5.00	0.00	6.31	0.46	2.01	0.09

The $bias_\theta$ statistic is mainly less than .1 over the range $-2 < \theta < 2$ for all of the methods, suggesting little bias. This statistic was close to −.1 around $\theta = 1$ on the CATs based on pool a, and pool b, which might occur if there were relatively few highly discriminating items near this level. Bias is larger at more extreme values of θ, which is characteristic of EAP θ estimates when there are few highly discriminating items that have difficulty values around these θ-levels.

TABLE 4
The Item counts within Exposure Rate Ranges (for CAT30i15e)

Exposure Rate Range	Counts	Cumulative Counts
a. Before Exposure Control		
(.50 to 1.00)	6	6
(.30 to .50)	35	41
(.15 to .30)	34	75
(.00 to .15)	102	177
Never Administered	243	420
b. After Exposure Control		
(.15 to .16)	9	9
(.10 to .15)	180	189
(.05 to .10)	34	223
(.00 to .05)	84	307
Never Administered	113	420

Figures 1b and 1e present the se_θ statistic and Figures 1c and 1f present the $rmse_\theta$ statistic. Focus is on $rmse_\theta$ because it is an overall index. As can be seen, the pure CAT has the smallest index value at all values of θ. In Figure 1c, the cat30i15e has smaller index values than the cat20i15e and the cat30i10e, which is as should be expected when test length is shortened (from 30 to 20 items) or exposure control is made more stringent (from 15% to 10%). As expected in figure 1f, the pool a and pool b statistic values are larger than those for the pure cat at all θ values, which is as expected when pool size decreases. The statistic values for pools a and b are also larger than those for the P&P test at all score values. The differences among the different CAT pools is discussed further in a later portion of the results section.

First and Second Order Equity for Scale Scores. The first and second order equity indexes for scale scores are plotted in Figure 2. Figures 2a and 2d show values of the $Sdiff_\theta$ statistic, which indexes the mean scale score differences between the derived scale scores and the number-correct score-based scale score on the P&P version. Statistic values of zero are expected if first-order equity holds. Figure 2a shows that there was a one-half to one point mean scale score difference between these scores and the number-correct based P&P scores at θ values in the range of 0 to 3.

The $CSEMSS_\theta$ statistics for scale scores are plotted in the middle portion of Figure 2. The same indexes are plotted against the conditional mean scale scores in the bottom portion of Figure 2. Note that the vertical axis values are the same in both the middle and bottom plots. The horizontal axis values differ and are just monotonic functions of one another, which are related through Equation 5. Two sets of plots are displayed so that the scale score standard errors can be related to both the θ scale and scale score scale. Outside of the middle score ranges, $CSEMSS_\theta$ is larger for the P&P number correct based method than for the other methods. This finding suggests that none of the CAT-based methods produce scores with $CSEMSS_\theta$ values that are close to those for P&P number correct based method.

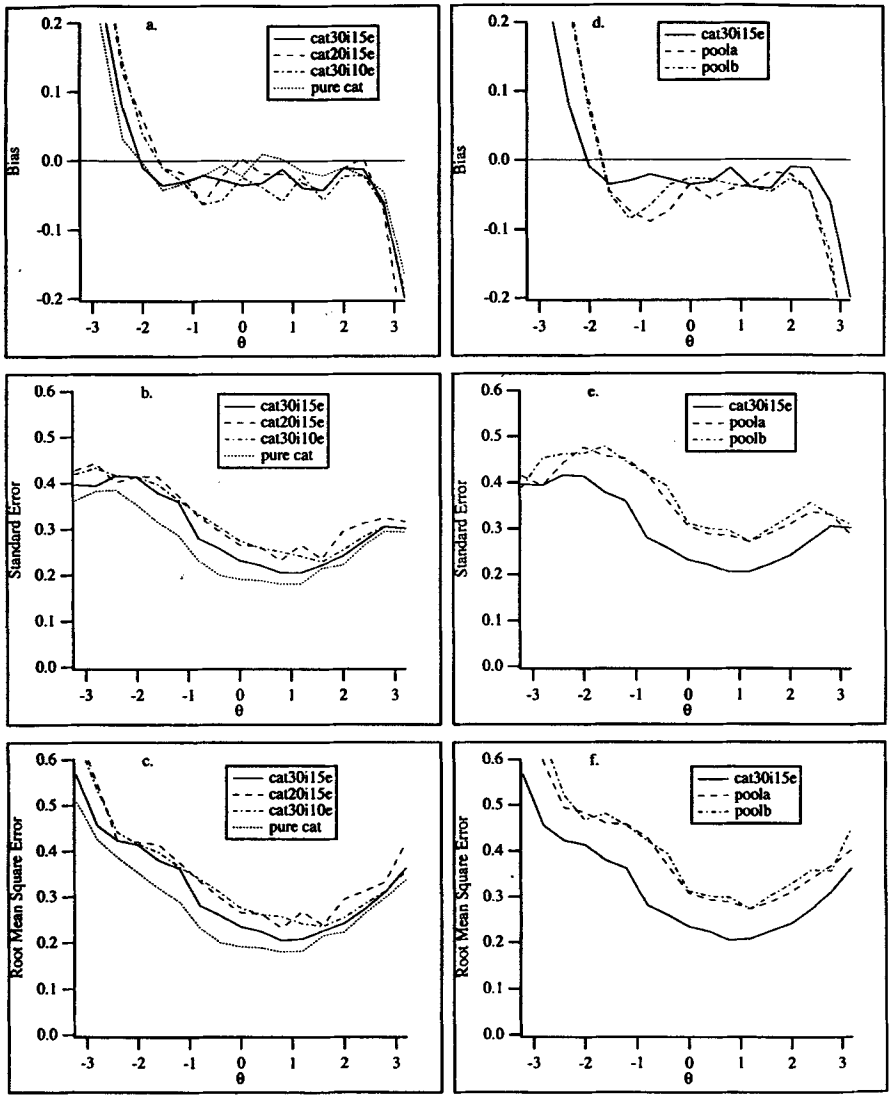

FIGURE 1. *Bias, SE, and RMSE of the estimated* θ's

Overall, the results provided in Figure 2 suggest that both first and second order equity, as compared to the current P&P number correct based scale scores, has not been achieved very closely by any of the methods. To some extent, this result is expected because the CAT and P&P scores used very different scoring methods. How to minimize the differences of $CSEMSS_\theta$ at the two ends of the scale remains a question to be investigated. Possible approaches to this problem include using a variable test length termination rule rather than fixed test length rule, and using different estimation methods such as maximum likelihood estimation. However, it

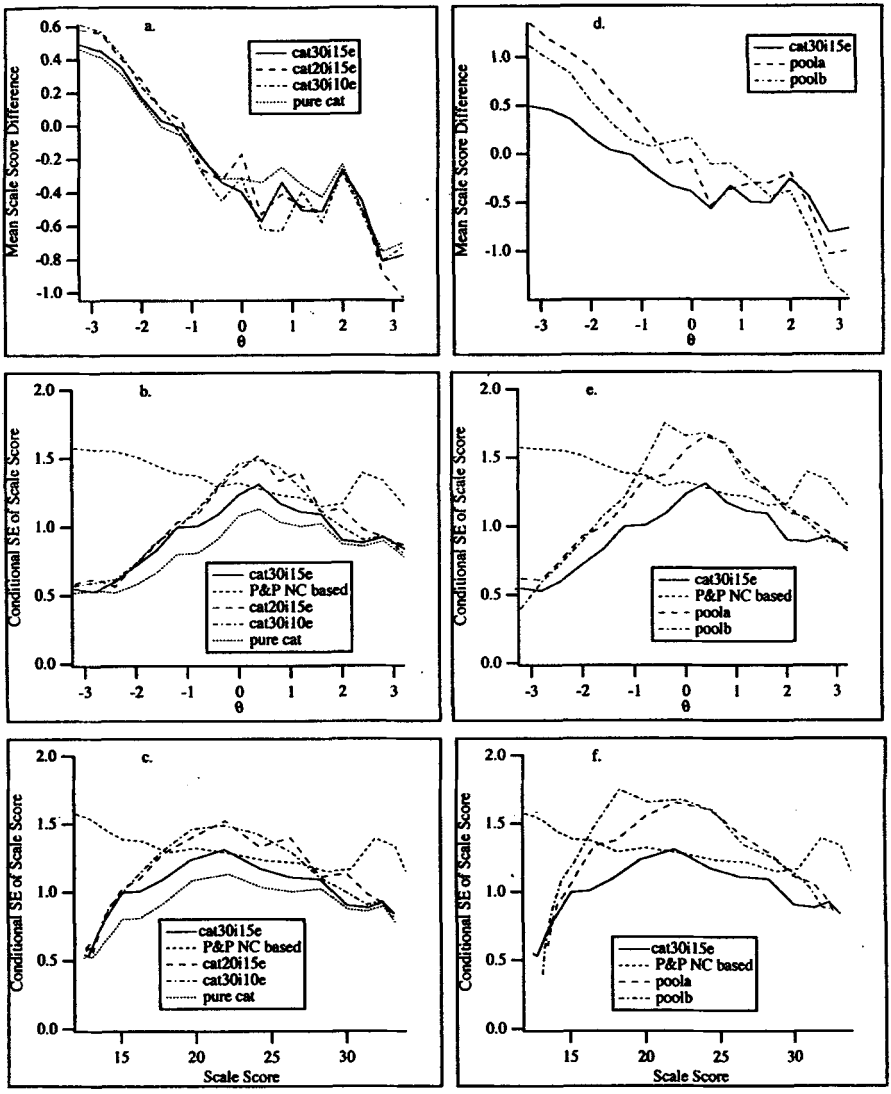

FIGURE 2. *The conditional mean and SE of scale scores*

is not expected that these approaches would solve this problem completely. In real testing situations, practitioners will have to relate the CAT and P&P scores without fully satisfying the second order equity criterion.

Comparison of Scale Score Distributions. The scale score distribution and cumulative distribution for a population of examinees with a standard normal θ distribution are plotted in Figure 3. Figures 3a and 3b show that the cumulative scale score distributions for all CATs were quite similar to one another, but they all differ from those for the P&P number-correct-based scale scores. This finding indicates that

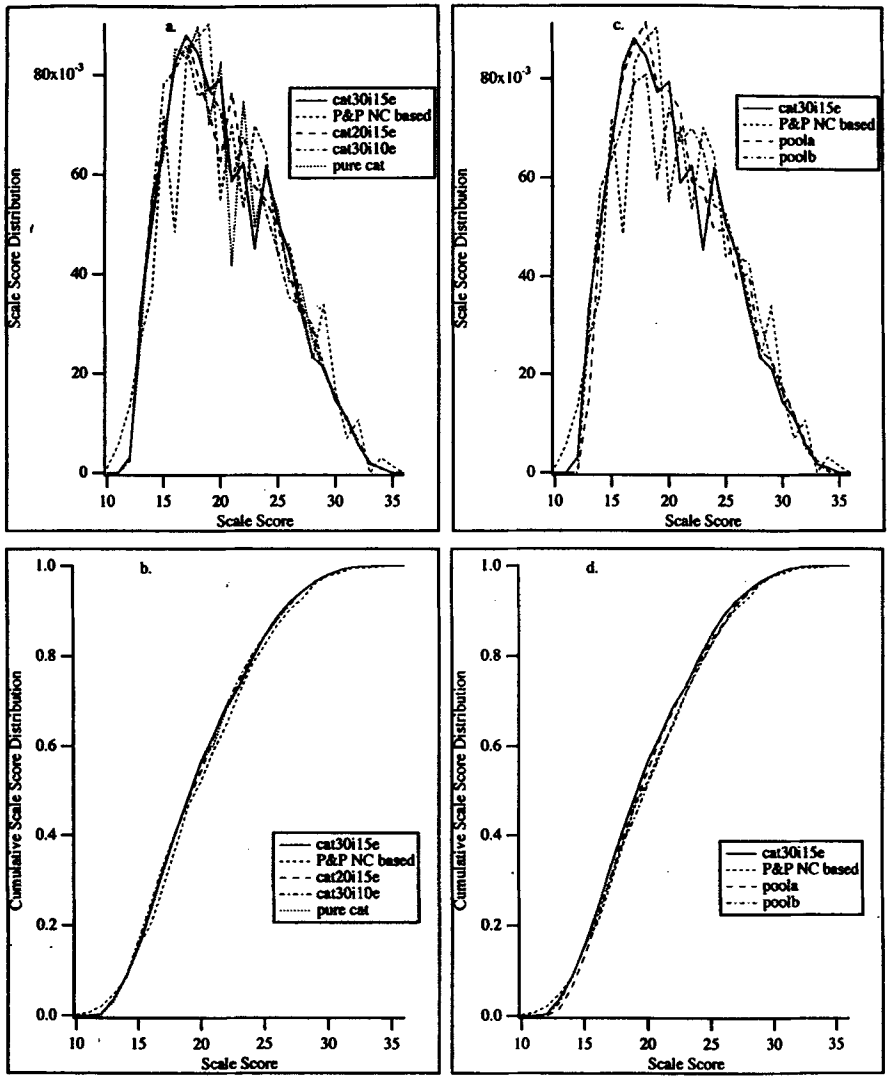

FIGURE 3. *The distributions and cumulative distributions of the scale scores*

the difference in scoring methods might contribute to the major difference in scale score distributions. The magnitude of these differences suggest that an equipercentile adjustment would still be needed if it were desired to make the scale score distributions nearly equal.

Variations of CATs. The figures give results for various CATs, including two different variations of the CAT test design, with one changing the test length from 30 items to 20 items, and the other changing the exposure rate limit from 15% to 10%. Results from these comparisons are plotted in the left parts of Figures 1

through 3. As mentioned earlier, Figures 1a, 1b, and 1c show that the two variations consistently increased $rmse_\theta$ of the standard CAT by a small amount. These two variations actually produced very similar errors. The $rmse_\theta$ of these two variations were slightly larger than that of the P&P version in the middle part of the θ scale, but were similar at other parts of the scale. Figures 2a shows that the two variations did not affect $Sdiff_\theta$ (first order equity) much, but increased the $CSEMSS_\theta$ in the middle part of the scale scores. The difference in $CSEMSS_\theta$ between these two variations and the P&P number-correct based scores might be greater than what could be accepted in practice. This result indicates that for this item pool, a test length of 30 items and item exposure rate upper limit of 15% is needed to make the CAT comparable to the P&P version. If exposure rate is to be controlled at a lower value, a larger item pool is needed, or the test length might need to be increased. Figure 3b shows that the cumulative scale score distributions for all CAT variations were quite similar to each other, although they all differ slightly from the P&P scores.

Comparison of Two Half-Size Pools. The half-size pools were constructed by randomly selecting items from the full pool. The results are plotted in the right side of Figures 1 through 3. Figures 1d, 1e, and 1f show that reducing the item pool size by half substantially increases estimation error. Figures 2d, 2e, and 2f show that reducing pool size by half also had a negative effect on comparability under the first and second equity criteria. The right side of Figure 3 shows a similar pattern to the left side of Figure 3, except that the cumulative scale score distributions of the pool b was very close to that of the P&P number-correct-based scale scores. One particularly interesting finding is that there appear to be some differences between the conditional means and standard errors for pools a and b as shown in Figure 2e and 2f. The characteristics of the two pools appear to lead to scale scores in the pool a that do not closely achieve the first and second order equity property with scale scores from pool b. There are also differences in scale score distributions for the two pools as is shown in Figure 3c and 3d. A possible explanation of these differences is that pool b has a much narrower range of difficulty than pool b. Thus, it could be argued that a further score adjustment would be needed before reporting scale scores from these two pools.

Proportions Exceeding Cut Scores. One way to evaluate some of the practical consequences of using different procedures is to compare the proportions of examinees who would exceed important cut scores on a test. Table 5 contains the proportions of examinees in the population with a standard normal distribution of θ that are above certain cut scores on the ACT score scale. These cut scores were chosen because they are relevant for certain important decisions. For example, a scale score of 18 is used in deciding whether National Collegiate Athletic Association (NCAA) athletes are eligible to participate in college sports as freshmen, and a score of 20 is near the mean for college-bound students.

The proportions in the table suggest that there could be a 2% to 4% difference in percentages exceeding cut scores between the CAT version and P&P number correct based versions for scale scores of 18 and 20, but less than 1% at scores of 15 and 30. One interesting finding is that there were larger differences in propor-

TABLE 5
Proportions of Examinees in the Population that Were At or Above Certain Cut Scores

Versions \ Cut Scores	15	18	20	30
a. The Proportions that were at or above cut scores.				
P&P NC based	0.9155	0.7122	0.5340	0.0385
CAT 30i 15e	0.9133	0.6758	0.5143	0.0341
CAT 30i 10e	0.9111	0.6656	0.5124	0.0341
CAT 20i 15e	0.9101	0.6734	0.5206	0.0352
CAT pool a	0.9362	0.7012	0.5319	0.0388
CAT pool b	0.9134	0.6993	0.5587	0.0353
b. The Differences in Proportions with P&P Number-Correct Based Scale Scores.				
P&P NC based	0.0000	0.0000	0.0000	0.0000
CAT 30i 15e	−0.0022	−0.0364	−0.0198	−0.0044
CAT 30i 10e	−0.0044	−0.0466	−0.0216	−0.0044
CAT 20i 15e	−0.0054	−0.0388	−0.0134	−0.0033
CAT pool a	0.0207	−0.0111	−0.0021	0.0003
CAT pool b	−0.0021	−0.0129	0.0247	−0.0033

tions for pools a and b. For example, 53.19% exceeded the cut score of 20 with pool a whereas 55.87% exceeded a 20 with pool b. Differences of this magnitude might have practical consequences in test use.

Table 6 contains the overall means and standard deviations (SDs) of scale scores for various CAT and P&P conditions. The means shows the maximum difference among these conditions to be about .4, which is a significant mean difference. The SDs show that the P&P score have larger SDs. These results again indicates there is sizable incomparability among these various conditions.

Discussion and Conclusions

The ACT Mathematics example illustrates how the psychometric characteristics of a CAT and a P&P version of a test, as well as alternate CAT pools, could be compared and evaluated through simulation procedures This example was intended to illustrate an approach that can be used at the early stages of CAT development. This psychometric evaluation was presented in the context of a theoretical framework that can be used to evaluate comparability issues with CATs. In CAT development, real data comparability studies and validation studies also should be considered, as is indicated in the theoretical framework that was provided.

The ACT Mathematics example illustrated conditions that influence the comparability of scores on CATs and P&P tests. The results clearly suggest a lack of comparability between any of the CATs that were simulated and the P&P tests. This lack of comparability was reflected in differences in expected scale score, scale score error variability, and score distributions. With considerable effort, the CAT length, pool size, and other CAT parameters might be able to be adjusted so that the resulting scale scores from the CAT and P&P tests would be sufficiently comparable for them to be used interchangeably.

TABLE 6
Overall Means and Standard Deviations of Scale Scores

Versions	Mean	Standard Deviation
P&P NC based	20.656	4.802
Pure CAT	20.405	4.625
CAT 30i 15e	20.334	4.609
CAT 30i 10e	20.288	4.637
CAT 20i 15e	20.374	4.634
CAT pool a	20.574	4.549
CAT pool b	20.668	4.651

Reviewers of an earlier revision of this paper made the strong case that by focusing on making the CAT and P&P scale scores comparable, many of the advantages of a CAT are lost. For example, one advantage of a CAT is that for a given amount of testing time, a CAT can be made to be more reliable than a P&P test. We agree that this increase in reliability is a significant potential benefit for a CAT and that our approach negates this benefit. (Note that even if the CAT and P&P tests are constructed to have the same reliability, the use of CAT could be beneficial by leading to shorter testing time while maintaining the same level of reliability as the P&P test.) However, if scores on a P&P and CAT test are to be used interchangeably, then making a CAT more reliable than the P&P test can result in some inequities. For example, a score of 18 on the ACT Assessment is used as a cut score for collegiate sports eligibility by the NCAA. In Table 5, approximately 71% of the examinees earn P&P number-correct based scale scores of 18 or higher. However, approximately 68% of the examinees earn CAT based scale scores above 18. If the CAT test were used, it appears that approximately 3% fewer examinees would be eligible. In addition, many individual examinees would be more likely to be eligible on the P&P test than on the CAT. Clearly, the P&P and CAT scores are not strictly interchangeable in this case. Therefore, it seems to us that if scores on a CAT and P&P test are to be used interchangeably in high stakes testing situations like the ACT Assessment, then efforts must be made to ensure that the scores on the two modes are comparable.

An alternative approach would be to develop a CAT with the intent that the P&P version would be abandoned in favor of the CAT. In this case, the CAT could be developed to maximize the benefits of the CAT and there would be no need to ensure comparability with the P&P test.

The example also illustrates how changes in components of a CAT can lead to scale scores from one configuration of components not being comparable to scores from another configuration of components. For example, if a CAT that had a pool of 200 items were changed to a CAT that had a pool of 400 items, then the scale scores might become more reliable, but it is likely that they would not be interchangeable with the scale scores based on the 200-item pool. The results suggest that scores on a CAT might not be comparable if any significant changes are made, which could include changes in the composition of item pools (e.g., choosing more

reliable items), changes in content balancing strategies, changes in exposure control parameters, and changes in pool size. That is, all of the components that go into an adaptive test, can affect the properties of the resulting scale scores, and when any of these components are changed, the resulting scale scores might not be comparable to scale scores earned prior to these changes.

Overall, the framework that was provided, as well as the example, is intended to help guide the development of CATs that are comparable to P&P tests and to produce comparable scores from alternate CAT pools. Developing comparable scores is clearly a complex undertaking. As CAT continues to increase in its popularity in large scale and high stakes standardized testing applications, the search for sound design and procedures for ensuring comparability is essential.

References

Bergstrom, B. (1992, April). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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.

Chen, S-Y, Ankenmann, R. D., & Spray, J. A. (1999). *Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing.* (ACT Research Report Series, 99–5). Iowa City, IA: ACT, Inc.

Cudeck, R. (1985). A structural comparison of P&P and adaptive versions of the ASVAB. *Multivariate Behavioral Research, 20*, 305-322.

Davey, T., & Nering, M. (1998, September). *Controlling item exposure and maintaining item security.* Paper presented at the colloquium, Computer-Based Testing: Building the Foundation for Future Assessments, Philadelphia, PA.

Davey, T., & Thomas, L. (1996, April). *Constructing adaptive tests to parallel P&P programs.* A paper presented at the annual meeting of the American Educational Research Association, New York.

Davey, T., & Parshall, C. G. (1996, April). *New algorithm for item selection and exposure control with computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, April, San Francisco.

Eignor, D. R. (1993). Deriving comparable scores for computer adaptive and P&P tests: An example using the SAT. (ETS Research Report, RR-93-5.) Princeton, NJ.: Educational Testing Service.

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

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computerized adaptive test design through simulation* (RR-93-56). Princeton, NJ.: Educational Testing Service.

Eignor, D. R., Way, W. D., & Amoss, K. E. (1994, April). *Establishing the comparability of the NCLE using CAT with traditional NCLEX examinations.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13*, 373-389.

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

Henly, S. J., Klebe, K. J., McBride, J. R. & Cudeck, R. (1989). Adaptive and P&P versions of the DAT: The first complete test battery comparison. *Applied Psychological Measurement, 13*, 363-371.

Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1994). A comparison of item calibration media in computerized adaptive testing. *Applied Psychological Measurement, 18*, 197-204.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*, 241-261.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147-154.

Kolen, M. J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment, 6*, 73-96.

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

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*, 285-307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional CSEM of the scale scores for scale scores using IRT. *Journal of Educational Measurement, 33*, 129–140.

Levine, M. L., & Krass, I. A. (1999, April). *Formula score and direct optimization algorithms in CAT ASVAB on-line calibration.*

Levine, M. V., Thomasson, G. L., & Williams, B. A. (1991). *Effects of replacing items on test equating* (Research Report 91-1). Model-Based Measurement Laboratory, University of Illinois at Champaign-Urbana.

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

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*, 389-404.

Lunz, M. E., & Bergstrom, B. A. (1995, April). *Equating computerized adaptive certification examinations: The Board of Registry series of studies.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Muhlstein, A. (1991). *Comparability of computer and P&P scores for two CLEP general examinations* (College Board Report 91-5). New York: College Entrance Examination Board.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and P&P educational and psychological tests. A review of the literature* (College Board Report 88-8). New York: College Entrance Examination Board.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and P&P cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458.

Messick, M. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement.* Phoenix, AZ: The Oryx Press.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*, 287-304.

Parshall, C. G., & Kromrey, J. D. (1993, April). *Computer testing versus P&P testing: An analysis of examinee characteristics associated with mode effect.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

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

Schaeffer, G., Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993). *Field test of a computer-based GRE general test.* (Research Report 93-07). Princeton, NJ: Educational Testing Service.

Schaeffer, G., Steffen, M., Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computer adaptive GRE general test.* (Research Report 95-20). Princeton, NJ: Educational Testing Service.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354.

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

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

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of medium of item presentation on examinee performance and item characteristics. *Jounral of Educational Measurement, 26*, 261-271.

Stocking M. L. (1988a). *Some considerations in maintaining adaptive test item pools.* (Research Report RR-88-33-ONR). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1988b). Scale drift in on-line calibration. (ETS Research Report, RR-88-28-ONR). Princeton, NJ: Educational Testing Service.

Stocking M. L. (1997). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21*, 365-389.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools.* (Research Report 94-5). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Stocking, M. L., & Swanson, L. (1996). Optimal design of item pools for computerized adaptive testing (ETS Research Report, RR-96-34). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995a). *A new method of controlling item exposure in computerized adaptive testing* (Research Report RR-95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995b). *Controlling item exposure conditioned on ability in computerized adaptive testing* (Research Report RR-95-24). Princeton, NJ: Educational Testing Service.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensional assessment. *Applied Psychological Measurement, 20*, 331-354.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151–166.

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

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

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

van de Vijver, F.J.R., & Harsveld, M. (1994). The incomplete equivalence of the P&P and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79*, 852-859.

Vispoel, W. P., Rocklin, T. & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computer-adaptive, and self-adaptive testing. *Applied Measurement in Education, 7*, 53-79.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In Wainer, H. (Ed.), *Computer adaptive testing: A primer* (Chapter 4, pp. 65–102). Hillside, NJ: Lawrence Erlbaum.

Wang, T. (1995). The precision of ability estimation methods in computerized adaptive testing. Unpublished doctoral dissertation. University of Iowa: Iowa City, Iowa.

Wang, T. (1997, March). *Essentially unbiased EAP estimates in computerized adaptive testing*. A paper in preparation. Paper will be presented at the annual meeting of the American Educational Research Association, Chicago.

Wang, T., Hanson, B. A., & Lau, A. C. (1999). Reducing bias in CAT trait estimation: a comparison of approaches. *Applied Psychological Measurement, 23*, 263-278.

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

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37*, 141–162.

Way, W. D. (1997, March). *Protecting the integrity of computerized testing item pools*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Way, W. D., Steffen, M., & Anderson, G. S. (September, 1998). *Developing, maintaining, and renewing the item inventory to support computer-based testing*. Paper presented at the colloquium, Computer-Based Testing: Building the Foundation for Future Assessments, Philadelphia, PA.

Wise, S. L., Roos, L. L., Plake, B. S., & Nebelsick-Gullet, L. J. (1994). The relationship between examinee anxiety and preference for self-adapted testing. *Applied Measurement in Education, 7*, 81-91.

Zeng, L. (1995). *EM1 for Windows. Version 2.0.* Iowa City, IA: author.

Authors

TIANYOU WANG is Principal Research Associate at ACT, Inc., PO Box 168, Iowa City, IA 52243; wang@act.org.

MICHAEL J. KOLEN is a professor at the University of Iowa, 224C Lindquist Center, Iowa City, IA 52242; michael-kolen@uiowa.edu.