

Running Head: STANDARD-SETTING IN CAT

Standard-setting Issues in Computerized-Adaptive Testing

Matthew M. Gushta

Centre for Research in Applied Measurement and Evaluation

Paper Prepared for Presentation at the Annual Conference of the Canadian Society for
Studies in Education, Halifax, Nova Scotia, May 30th, 2003

Abstract

The progression and evolution of the use of computers in education has led to the development of computerized adaptive tests (CATs). CATs administer subsequent items tailored to an examinee's ability estimate (see Wainer, 2000, for further explanation), resulting in a greater amount of information provided at the item level and a lower estimate of standard error, in comparison to paper-and-pencil, linear form examinations. Adaptive testing, however, introduces complexities to the process of standard-setting and, especially, establishing cut-scores. An examination of The Criteria for Defensibility (Berk, 1986) reveals the need for serious evaluation of such procedures within CAT environments. Three of the most salient issues are: (1) precision versus equity, (2) randomly parallel forms, (3) medium effects, (4) item pool versus subsets, and (5) analysis.

Introduction

The progression and evolution of the use of computers in education has led to the development of computerized adaptive tests (CATs). Most often used in large-scale assessments, CATs offer flexibility of location and scheduling of test administrations as well as allowing for presentation of more dynamic item types than possible in a paper-and-pencil based test (P&P) setting. Instantaneously calculating an estimate of an examinee's ability given their performance on a specific item, CATs administer subsequent items tailored to that examinee's ability estimate (Wainer, 2000). CATs yield a greater amount of information at the item level and a lower estimate of standard error of measurement in comparison to linear format examinations. CATs, however, introduce complexities to the process of standard-setting. An examination of *The Criteria for Defensibility* (Berk, 1986) and the *Criteria for Evaluating a Performance Standard-Setting Study* (Hambleton, 2001) reveal the need for serious reconsideration of the procedure for setting standards, especially in regards to the establishment of cut-scores on CATs. Kane (2001) draws the distinction between *performance standards* and *cutscores* (or cut-scores) by stating that the former describes "a level of performance in terms of what examinees at that level know and can do" (p. 55) and the latter is a point on the score scale "associated with each performance standard" (p. 55). The use of a CAT system changes the mode and method by which examinees' results are gathered and calculated, while the definition of examinee performance is not necessarily altered. This paper specifically seeks to highlight points for consideration in setting cut-scores in a CAT environment.

Before issues associated with standard-setting for CATs can be discussed, it is important that the CAT procedure itself be understood. Wainer (2000) provides a comprehensive overview and synthesis of many issues in the computerized adaptive testing domain. For the purposes of this paper only a brief discussion of these concepts will be presented. The following are five of the most salient issues to be considered in establishing cut-scores for CATs:

- (1) Precision versus equity: using Item Response Theory (IRT), CATs offer the ability to gather maximal information throughout examinations using termination criteria while traditional paper-based methods use definite end points (Crocker & Algina, 1986; Hambleton, Swaminathan, & Rogers, 1991);
- (2) Randomly parallel forms: CAT administration yields unique, randomly parallel, forms for each examinee, which must be considered by panellists while attempting to set standards and establish cut-scores;
- (3) Medium effects: it is possible that CATs offer a novel testing situation for examinees, resulting in confounding of performance estimates, due to factors such as computer experience or test anxiety;
- (4) Item pools and subsets: item banks in CATs are often extremely large, raising the question of the appropriateness in using a subset of items to set cut-scores;
- (5) Analysis procedures: one of the greatest problems associated with standard-setting in CATs is that of scale conversion and interpretation.

This paper will detail each of the above points in an effort to illuminate issues associated with setting cut-scores for CATs. Through an example of the CAT process and resulting performance estimates, it will become evident that the existing methods for establishing cut-scores are not appropriate for CATs. Three documented methods with the specific purpose of setting cut-scores in a CAT framework are described. These procedures do not represent all possible standard-setting methods applicable to CATs, only those presented in the literature at the time of writing. Pitoniak and Sireci (1999) further review standard-setting methods applicable for, but not specific to, CATs. Any procedures used in setting CAT cut-scores must be updated with additional considerations regarding computerization and adaptive testing and in regards to criteria established for the evaluation of standard-setting procedures.

Introduction to Computerized Adaptive Testing

Computerized adaptive testing directly incorporates item response theory (IRT) in the estimation of examinee ability and the selection of test items (Hambleton et al., 1991). Under this theory, it is assumed that performance on an assessment is reflective of examinee ability or a latent trait, θ , underlying test performance. This ability is normally distributed across the score scale. In CATs subsequent items are selected based on previous estimates of examinee ability obtained from earlier items such that the information provided is maximized for the associated ability or performance estimate. This process continues until specified termination criteria are satisfied. This process can be seen in Figure 1. Within CATs, examinees' ability or proficiency scores

are typically calculated according to the IRT three-parameter logistic (3PL) model (Hambleton et al., 1991)

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}},$$

where $P_i(\theta)$ is the probability that an examinee with ability θ answers item i correctly, a_i , is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the pseudo-chance-level parameter that represents the probability of an examinee with an extremely low ability answering the item correctly, D is a scaling factor equal to 1.7, and e is a transcendental number whose value (correct to three decimal places) is 2.718. Ability estimates within an IRT framework are normally distributed about a mean ability score of 0 with a standard deviation of 1.0.

Starting Point

The most popular method for determining examinee starting points in CATs is to select initial items based on maximal information provided at $\theta = 0.0$. Given that an examinee's ability score may be unstable for the first few items administered, a burn-in period is often established before arriving at a stable ability estimate the results of these calculations are then discarded (Thissen & Mislevy, 2000). A modification of the starting point procedure involves the use of a Bayesian technique for initial ability estimation by which theta is estimated based on a known ability distribution. This distribution is based on previous examinee information regarding population variables.

Selecting Subsequent Items

Once a response to the initial item has been calculated, one of two item selection algorithms is enacted by the CAT program: maximal information selection or Bayesian

selection. Both algorithms can be thought to be expressed in terms of θ , a provisional estimate based on preceding items, where an “examinee’s next item [is selected] to be informative in the neighbourhood of a current proficiency estimate based on the responses to previously administered items” (Thissen & Mislevy, 2000, p. 105).

The maximal information selection procedure selects items based on the provision of maximal information at the level of the current ability estimate. Item information functions as defined within IRT are the “information” provided by an item at a given ability level, with higher values being associated with highly-discriminating items whose difficulty parameter closely approximates the theta level of interest (Hambleton et al., 1991). Calculation of item information functions is operationalized through

$$I_i(\hat{\theta}_i) = \frac{[P_i'(\hat{\theta}_i)]^2}{\{P_i(\hat{\theta}_i)[1 - P_i(\hat{\theta}_i)]\}},$$

where $I_i(\hat{\theta}_i)$ is the “information” provided by item i at θ , $P_i'(\theta)$ is the derivative of $P_i(\theta)$ with respect to θ , and $P_i(\theta)$ is the item response function.

The Bayesian procedure (Owens, 1969; Owens, 1975; Wainer & Mislevy, 2000) uses the posterior distribution of θ , determined after n preceding items, to select subsequent items providing a maximum amount of information given $p(\theta|S_n)$, where S_n represents all of the information available about an examinee given n items.

In actual CAT scenarios, the maximal item information selection procedure is most often utilized as the ensuing estimates of ability are influenced less by a previously assumed population distribution (Thissen & Mislevy, 2000). A table containing a matrix of items, information, and ability levels is stored within the CAT software and from this

Info Table the algorithm selects the next item providing the most information at the current ability estimate to the examinee (Thissen & Mislevy, 2000).

Termination Criteria

It is in selecting a method for termination that CATs reveal the most promise, offering test administrators the opportunity to specify any combination of measurement precision and/or administrative aspects of testing, such as test length or administration time as criteria. Flexibility in the definition of termination criteria allows test developers a greater degree of control over both the practical and methodological/psychometric considerations in test creation as instruments can be produced unique to each test and administrative scenario.

Measurement Precision Termination Criteria

The use of measurement criteria in defining termination rules means that specific attention is paid to the standard error (SE) associated with examinees' estimates of ability; in CATs this is referred to as the *measurement precision* termination procedure. Given that item selection in CAT is typically performed according to item information, the calculation of SE is:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}},$$

where $SE(\hat{\theta})$ is the standard error associated with a given ability estimate, $\hat{\theta}$, and $I(\theta)$ is the test information function: $I(\theta) = \sum_{i=1}^n I_i(\hat{\theta})$. A stopping rule based on measurement precision gives the ability to ensure equality of precision across examinees and

administrations, as compared to other methods of testing which are based on a linear, fixed form examination resulting in variable SE across examinees.

Administrative Termination Criteria

The latter termination criteria, test length and time, more closely resemble practices in traditional, linear format examination administration. Specifically, *n items termination criterion* refers to the procedure in which a CAT is terminated once an examinee has completed a predetermined number of items, while the criterion of time simply terminates the examination session once a predetermined amount of time has elapsed.

The administrative termination criteria can each be used as the sole criterion in the termination of CATs or in conjunction with the measurement precision criterion, creating an exam that meets both psychometric and practical considerations. Hybrid construction of CATs is typical as practical constraints often do not permit every examinee to reach a pre-specified level of measurement precision within a practical administration: some examinees may require an extremely large number of items to reach a certain SE and this is simply not feasible as good test administration practice.

An Example

To illustrate the computer adaptive testing algorithm as compared to the traditional linear format, an example of a test administration using both adaptive and linear item selection procedures is presented. Across both procedures, examinee responses were simulated according to the stopping rules (1) *measurement precision*

termination or (2) *n items termination*. For the *measurement precision termination* criterion example, the stopping rule was set at $SE = 0.3$; the *n items termination* criterion set the stopping rule at 25 items.

CAT responses were simulated according to the 3-parameter logistic model (3-PL) using the POSTSIM software package (Assessment Systems Corporation, 1999). Descriptive statistics for both the simulated examinees and the simulated item parameters can be found in Table 1. To illustrate the differences between adaptive- and linear-format testing procedures, ten examinees were selected randomly from the simulation and matched on ability across termination conditions. For the *measurement precision termination* condition items were selected consecutively beginning with the first item in the pool; standard errors of each ability estimate given administration of the first 25 items in the item pool were calculated for the *n items termination* condition. The results of the *measurement precision termination* criteria administration and the *n items termination* criteria administration simulations can be seen in Table 2.

Examination of the results of the *measurement precision termination* administration for the adaptive-format ($\bar{X}_n = 55.40, \hat{\sigma}_n = 9.51$) and the linear format ($\bar{X}_n = 104.10, \hat{\sigma}_n = 3.78$) reveal a significant difference between the number of items administered before termination ($SE=0.3$), $t(9) = 10.79, p < .05$. Similar results can be seen when comparing the standard error achieved in the *n items termination* condition, with the adaptive-format ($\overline{SE} = 0.40, \hat{\sigma}_{SE} = 0.02$) and linear-format ($\overline{SE} = 0.62, \hat{\sigma}_{SE} = 0.04$) yielding significantly different standard errors, $t(9) = 24.64, p < .05$. From these results we can see that test scores derived from within an adaptive framework require a shift in the interpretation paradigm; scores for examinees deemed near-identical under a classical,

or raw score testing and analysis framework are significantly different in an adaptive framework. In the following sections, issues arising in establishing cut-scores as a result of this shift are discussed.

Issues in Establishing Cut-Scores

In a recent study by Wang and Kolen (2001), issues associated with the comparability of paper-and-pencil based (P&P) test administrations and CATs were examined. Issues of comparability and equivalency between CATs and P&Ps may be used to illustrate those that arise in standard-setting procedures when trying to establish cut-scores. When considered as a paradigm shift, paper to computer administration and linear to adaptive format, issues associated with establishing cut-scores in a CAT environment become apparent.

“Criteria for Evaluating Comparability” (p. 26) are detailed by Wang and Kolen (2001) and are correspondent to the five issues prevalent in setting cut-scores considered in this paper: 1) precision versus equity, 2) randomly parallel forms, 3) medium effects, 4) item evaluation/selection, and 5) raw score versus ability score analysis.

Precision versus Equity

As previously established, flexibility offered within a CAT framework allows test developers and administrators to specify any combination of termination rules as would be possible in a practical testing environment and as acceptable with regards to measurement precision. These are advantages not offered by the traditional, linear-form

P&Ps. The ability to pre-specify a value for standard error of measurement is appealing but, in actuality, creates an entirely different testing experience for the examinee, which requires a significant amount of deliberation given that “everyone should have an equal opportunity to display proficiency on the test” (Wainer, Dorans, Green, Mislevy, Steinberg, & Thissen, 2000, p. 254). Adaptive testing and computerized testing have the potential to introduce confounding effects on performance, given that the method or mode of administration may be associated with a greater degree of novelty and/or anxiety for the examinee.

Lord (1980) and Wang and Kolen (2001) discuss the reliability criterion of equity in comparing test results – conditional scale score distributions must be the same across test forms. *First order equity* is defined as

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

where E refers to the 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 (Wang & Kolen, 2001, p.29). An individual, at any given ability level, should be able to earn the same scale score across forms. The criteria of *second order equity* states that measurement precision must be the same across forms for all examinees at a given ability level

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

where σ is the standard deviation of scale scores over examinees of a given ability, θ (Wang & Kolen, 2001, p.29). These two notions may be summarized by the criterion of *Equal probabilities of achieving passing [or cut] scores*

$$P(s_1 \geq s_c | \theta) = P(s_2 \geq s_c | \theta), \text{ for all } \theta,$$

where P is the probability of passing and s_c is the passing score. The above criteria can empirically be assessed through the calculation of the mean difference between mode scale scores

$$Sdiff_{\theta} = \bar{s}_{\theta CAT} - \bar{s}_{\theta P\&P} \quad (\text{Wang \& Kolen, 2001, p.36}),$$

hypothesized to be zero, $Sdiff_{\theta}$ is the difference in scale scores at any given ability, θ , and \bar{s}_{θ} is the mean scale score at θ for the mode of administration; second order equity is ascertained through calculation of conditional standard error of measurement of scale score (CSEMSS) as given by the formula:

$$CSEMSS_{\theta} = \sqrt{[\Sigma(s_{\theta} - s_{(\text{mean})\theta})^2]} \quad (\text{Wang \& Kolen, 2001, p.37}).$$

Given such stringent requirements for the criteria of equivalency, it can be seen that the issue of precision versus equality is much more complex than simply establishing score stability with respect to mode or method of administration.

Using the American ACT Mathematics exam, a large-scale assessment instrument that is administered adaptively to exiting high school students, Wang and Kolen's (2001) equity analyses indicated that neither first- nor second-order equity were satisfied. Simulating response data given previously administered item parameters, comparisons showed non-equivalence between CATs and P&Ps. Differences in both true score and scale scores ($Sdiff$) were found as well as differences in true and scale score error variability (CSEMSS). Specifically, Wang and Kolen (2001) state that "71% of the examinees earn[ed] P&P number-correct based scale scores of 18 or higher. However, approximately 68% of the examinees earn[ed] CAT based scale scores above 18" (p. 45). The importance of such inequity becomes apparent when such a difference is seen to affect 204 examinees within this simulation. "A score of 18 on the ACT

assessment is used as a cut score for collegiate sports eligibility by the NCAA” (Wang & Kolen, 2001, p. 45).

As can be seen from the above example, administration of a CAT cannot be considered equivalent to a P&P and the results have serious implications when examining the results of setting a cut score across both forms. So it remains that equity and precision must be considered when evaluating CATs; while computerization and adaptive form tests offer a great number of possibilities they also introduce complex considerations, beginning with the possibility of an entirely different psychometric foundation resulting from satisfaction or sacrifice of measurement precision or equity criteria requirements.

Randomly Parallel Forms

Item pool size recommendations suggest that CATs have at least six- to eight-times the number of items required for a linear test form of twice the size of the average CAT administered (i.e., a CAT of average 50 items requires a pool of approximately 600 items) (Stocking, 1994). With such a large repository of items to draw from and innumerable possibilities for presentation of these items within a test administration given the above stopping rules, it can be seen that it is possible for each and every examinee to be administered an entirely unique, or randomly parallel, test. Depending on how strictly content specifications are adhered to, if at all, this poses a number of problems for setting cut-scores in this framework – many documented procedures are not applicable. Certain test-centred procedures have been adapted for use in a CAT setting (see ‘Methods for Standard-setting and Establishing Cut-scores in CATs’ later in this paper). Examinee-centred methods, however, are inappropriate for use in setting

cut-scores on CATs as it can be extremely difficult for judges to compare examinee results given that each exam could be entirely different and performance evaluation (ability estimation) for CATs is statistically intensive, as opposed to lending itself to visual inspection as in the case of raw score-based analyses (Jaeger, 1989).

For the purpose of conceiving of CATs as randomly parallel forms, an analogous example may be seen in practice in Alberta primary and secondary schools. Alberta Learning administers province-wide achievement tests at the grade 3, 6, and 9 levels for the purpose of providing program information to administrators (Alberta Learning, 2003; Alberta Learning, personal communication, May 13, 2003). Each year, these exams are developed comprising approximately one-half unique items and one-half repeated items from the year previous (Table 3). Drawn from parallel item pools year to year, these exams are randomly parallel across administrations. Setting cut-scores for the *Acceptable Standard* and *Standard of Excellence* for these exams involves an equipercentile equating process (Kolen & Brennan, 1995) by which cut-scores for the current year are scaled to the previous year. Table 4 shows the results of the 2001 Grade 6 science achievement test. Given the requirement that 85% of examinees are required to meet the *Acceptable Standard* and 15% are to achieve the *Standard of Excellence*, it can be seen that the 2001 exam was easier than the 2000 exam. As a result of this, the cut-scores for the two levels were shifted: from 26 to 27 for the *Acceptable Standard*, from 42 to 43 for the *Standard of Excellence*.

As seen in the above example, establishing methods for setting cut-scores in a CAT framework benefits from procedures already in use by more traditional testing systems when considered as randomly parallel forms. While more traditional testing

paradigms consider the test period as a single test administration, CATs require that the form administered to each examinee be considered as a single administration as each test administered is unique. By first establishing the reference form and cut-score to which CAT performance may be compared, establishing a cut-score is then facilitated by the application of documented equating procedures used across test systems.

Medium Effects

Medium effects refer to the introduction of other sources of error in test performance attributable to the mode and/or medium of administration, which must be accounted for in the adaptation or consideration of CATs (Wainer & Mislevy, 2000). The design of a computerized adaptive test must consider such factors as 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 as well as examinee variables such as test and computer anxiety (Kolen, 1999-2000).

Although strong correlations have been reported between computerized and paper-based instruments (Mead & Drasgow, 1993), this is an instance of rank misidentification. Divgi and Stoloff (1986) revealed systematic differences between item characteristic curves (ICCs) across media despite strong cross-modal correlations. These results illustrate the conclusion that “differences between P&P and CAT ICCs are often systematic and substantial, indicating the existence of a significant medium of administration effect on item response curves” (Divgi & Stoloff, 1986, p.6).

Evidence of medium effects has implications for Wang and Kolen's (2001) "Statistical Assumption/Test Administration Criterion" when the comparability of CATs and P&Ps is examined. The evidence presented above shows that simple statistical comparisons, such as correlation, cannot reveal differences between the two modes of administration – differences that are present, violating the assumption of equivalence across test administration modes. Non-equivalency requires that panel members, while setting cut-scores, be made aware of the possible effects resulting from computerization and a shift from a linear to adaptive format. The establishment of cut-scores and the determination of performance standards in a CAT framework must, at the very least, account for the possible introduction of performance and measurement error due to medium effects.

Item Pools and Subsets

Given the enormity of the item pools associated with CAT, it may not be feasible to consider the evaluation of each item by each judge – that could very easily result in 500 judgements made by each judge, a cognitively and practically overwhelming prospect. Should the entire pool not be evaluated, procedures for establishing cut-scores must then establish an acceptable method in which only some of the items will be used.

One method available for establishing cut score using a sub-pool is through the calculation of an Item Pool Score:

$$E(IPS) = \sum P_j(\theta),$$

where IPS is the Item Pool Score (Dorans, 2000). This calculation provides the expected Item Pool Score using only a meaningful subset of items. This method is appropriate given that subsets have been found to be generalizable to the whole item

pool (Dillon & Walsh, 1998). As this is an unavoidable practical issue associated with CAT, it is further addressed in the “Methods for Standard-setting and Establishing Cut-scores in CATs” later in this paper.

Raw Score versus Ability Score Analysis

Based on modern measurement theory, understanding of CAT results requires that the standard-setting judges have, at minimum, an appreciation for the distribution of ability scores as θ and for the calculations involved in deriving those values. Without knowledge of IRT, panellists would not be able to evaluate or appreciate examinee performance within such a testing framework especially given that in a perfectly functioning adaptive test “every test taker will answer about half of the items correctly” (Zieky, 2001). A comparison of raw, or total, score and ability score analysis procedures was found to yield 2 –4 % differences in proportions of examinees exceeding cut-scores of relatively lower frequencies, while cut-scores of higher frequencies (closer to the mean) evidenced less than a 1% difference (Wang & Kolen, 2001). These results further illustrate the inappropriateness of any type of raw score or percentage mastery calculations used in establishing cut-scores (Dorans, 2000). Additional training, while expensive and time-consuming, is strongly advocated for panellists who are setting standards for CATs (Sireci & Clauser, 2001).

Reconsideration of Berk (1986) and Hambleton's (2001) Criteria

As can be seen through illustration of the issues presented above, computerization and the use of an adaptive testing system introduces additional complexities to the process of setting cut-scores in educational assessment. In comparing, contrasting, and assessing the capability and suitability of each procedure that might be used in this process, utilization of CATs requires that these additional issues be weighed and incorporated to ensure proper and thorough evaluation of the procedure. Specifically, these issues must be incorporated into Berk's (1986) "Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests" and Hambleton's (2001) "Criteria for Evaluating a Performance Standard-Setting Study."

Berk's (1986) guide was written for the purpose of applying the criteria to 38 methods that were reviewed. At the time of writing, CAT was not as popular or widely used as it is now, with many large-scale exams beginning to implement CATs; following are revisions proposed for the "Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests" in evaluating standard-setting procedures applied to computerized adaptive tests. The "Criteria for Defensibility" (p. 140) refer to points of evaluation that establish the "Technical Adequacy" (p. 140) and "Practicability" of the standard-setting procedure. Evaluation of these points require satisfaction of certain psychometric and statistical standards such as would be defensible to experts; sensitivity to the instruction or training of the examinees; and the procedure must be simple in its implementation, calculation, and interpretation. Within these criteria, it is strongly suggested that standard-setting methods and the setting of cut-scores not appear "statistically magical" (p. 144) to the panellists or the general public. Adjustment

of the criteria to account for CATs require that further attention be given to the context and measurement theory upon which the test is based; CATs offer many examinees, and panellists, a novel testing environment, through computerization and its basis in IRT, as a result, examinees' performance and standard-setting judges' evaluations must account for such effects. One possibility for satisfaction of Berk's (1986) criteria lies in training of examinees and judges with regards to the CAT mechanism, reducing medium effects for students, and familiarizing panellists with the calculations and interpretations of ability scores.

More recently, Hambleton (2001) introduced the "Criteria for Evaluating a Performance Standard-Setting Study," which are preceded by "Steps for Setting Performance Standards on Educational Assessments" that outline good practice at all points of a standard-setting procedure. It is in these initial steps that issues relevant to CAT are evident, beginning with the definition of *content standards* as "What examinees are expected to know and be able to do" (Hambleton, 2001, p. 91) and *performance standards* as communicating how well examinees are expected to perform in relation to the content standards (Linn & Herman, 1997). Hambleton (2001) suggests training panellists to use the method, including practice in providing ratings, as effective training must include "taking the test under standard or near standard conditions" and a review of "the item pool on which the standards will be set" (p. 99). Requiring that panellists review the test under adaptive, not linear, testing conditions, a full review of the item pool is likely not feasible for the purpose of standard-setting. This again exposes the issue of determining a method for sampling CAT item pools for use in setting standards. Further to understanding content and performance standards, panellists must also be

informed of possible confounds to examinee performance including medium effects, resulting from a novel testing procedure, as well as the possibility of examinees' computer anxiety. The above definitions and issues require an understanding of both performance within a CAT framework as well as the relation between the test and examinees' ability as previously established in discussion of Berk's (1986) criteria.

The "Criteria for Evaluating a Performance Standard-Setting Study" (Hambleton, 2001, p. 108) details a set of 20 questions that provide a framework for evaluating the quality of a standard-setting study or may be used to guide the setting of performance standards. Of the 20 questions that should be addressed, three have direct implications when CATs are examined: "Were sufficient resources allocated to carry out the study properly?" (p. 109), "Were panellists administered the educational assessment or at least part of it?" (p. 111), and "Were panellists given the opportunity to 'ground' their ratings with performance data and how was the data used?" (p. 112). As would be expected from computerization, CATs can both simplify and introduce complexity to the test administration and standard-setting process. Data collection and handling can be streamlined, however, computerization also requires additional expertise and resources (i.e. hardware and software) – it must be determined if the costs outweigh the benefits for each individual case. In addressing the second question, it must then be asked if panellists gain a full appreciation for CAT through a single exposure or through multiple exposures. A recommendation of many standard-setting studies is that panellists complete the exam themselves or in the mindset of a minimally competent candidate (MCC), but given the functioning of CATs it is possible, and more than likely, that an MCC may be administered randomly parallel forms across multiple administrations. This

is true even if it were possible to control the ability estimate of the MCC. The question must then be “How many administrations are required for panellists to fully appreciate the examination experience?” Finally, Hambleton’s (2001) question/criterion regarding the grounding of ratings in performance data must be carefully considered as this procedure relies heavily on the appropriate training of panellists in the interpretation of IRT results and thorough understanding of latent trait theory. Raw score data is easily interpreted and can be visually inspected and used in affecting judgments, ability estimates, however, require greater competency in statistics and measurement theory or at least training in its interpretation.

Methods for Standard-Setting and Establishing Cut-scores in CATs

Standard-setting studies incorporating the issues addressed above have yet to be documented in the literature let alone put into practice. Three methods specific to CATs have been suggested and are for the most part simply modified Angoff procedures. As untested and undocumented procedures, the following are currently recommendations and require further research to determine efficacy and suitability for setting standards in CATs.

The first of these methods is the *Wainer Method* (Sireci & Clauser, 2001; Sireci, 2002; Walter Way, personal communication, 1994, as cited in Zieky, 2001). This method is a simple application of the Angoff procedure in which judges perform as MCCs, analogous to indicating in binary format the possibility of answering correctly. The resulting θ estimate could then be set as the cut-score. A variation of this has also been proposed in which judges would use likert ratings of the possibility an MCC would

answer correctly (Sireci, 2002). As in the strict version of the Wainer Method, the sum of these ratings could be used to establish the cut-score.

The *Exemplar Test Form* (Martha Stocking, personal communication, 1992, as cited in Zieky, 2001) requires that the CAT in question utilize content specifications in the design of the item pool and in item administration. Using these specifications, items for evaluation in a standard-setting study are selected and used to create a linear format. Prior calibration of items and adherence to content specifications then allows this method to be considered similar to an actual CAT administration without requiring an actual adaptive or computerized setting; judges may then focus on rating items instead of the mode and medium of administration.

The final standard-setting procedure proposed is the *Direct Consensus Method* (Sireci, 2002), a modification of the Ebel (1991) procedure. Under this method, items are first grouped according to the content specifications prior to examination by the judges. The cut-score is then set according to the sum of the proportion of items from each area that an MCC could be expected to answer correctly. This judgement is based on a holistic overview of each content category and therefore does not require the selection of items sub-pools for the purpose of reviewing.

A possible fourth, undocumented, method for setting cut-scores in a CAT framework could be informed by a procedure such as the previously mentioned Alberta Learning Achievement Test equating method (Alberta Learning, personal communication, May 13, 2003). While it is possible, and very likely, that each CAT delivered could be entirely unique, it is also possible that exams could be constructed to contain a minimal number of common items (possibly in the burn-in period of the

administration). Given the administration of parallel and common items, cut-scores could be set through the incorporation of an equating procedure across forms.

Discussion and Conclusion

Through an introduction to computerized adaptive testing and the typical algorithm associated with the administration of a CAT, additional complexities have been illustrated in comparison to setting standards and cut-scores for traditional, linear, paper-based, educational assessments. Fundamental differences beginning with the theory upon which the test functions through to cognitive differences associated with the test delivery introduce additional considerations to be accounted for. While not exhaustive, the issues discussed throughout this paper provide a foundation that may be used in the implementation of and further research into standard setting for CATs.

Precision, equity, parallelism, medium effects, item selection throughout test progression, and scoring and scaling are issues that add complexity to any method use in establishing cut-scores in a computerized and adaptive testing environment. The *Wainer*, *Direct Consensus*, and the *Exemplar Test Form* methods are all attempts to account for the differences in testing introduced by a computerized, adaptive administration, but as first attempts, these procedures have flaws. Examination of equating processes already in practice may serve to facilitate the conceptualization of setting cut-scores in CATs, with each CAT form constructed considered as a single administration. With the proliferation of computer-based testing, there is a great opportunity and need for research in this area. As the technology and its implementation progresses ahead of research in this area, emphasis must be placed on the importance

of panellist training in regards to latent trait theory and item response theory, from which appropriate performance standards and the corresponding cut-scores may be established for CATs.

References

- Alberta Learning, Government of Alberta (2003). *Description of the Achievement Testing Program*. Retrieved May 15, 2003, from http://www.learning.gov.ab.ca/k_12/testing/results_2002/ach/ach_descrip.asp
- Assessment Systems Corporation (1999). POSTSIM [Computer program]. St. Paul, MN: Author.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56(1), 137-172.
- Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Orlando, FL: Harcourt Brace Jovanovich.
- Dillon, G. F., & Walsh, W. P. (1998). *Standard setting judges' perceptions on the use of performance data to guide their decisions*: Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Divgi, D. R., & Stoloff, P. H. (1986). *Effect of the medium of administration on ASVAB item response curves*. Alexandria, VA: Center for Naval Analysis.
- Dorans, N. J. (2000). Scaling and equating. H. Wainer (Ed.), *Computer Adaptive Testing: A Primer* (2nd ed., pp. 135-158). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of Educational Measurement* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 89-117). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.
- Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485-514). Washington, DC: American Council on Education.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- 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 practice*. New York, NY: Springer-Verlag.
- 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.

Linn, R.L., & Herman, J.L. (1997). *A policymaker's guide to standards-led assessment*.

Denver, CO: The Educational Commission of the States.

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

Hillsdale, NJ: Lawrence Erlbaum Associates.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114(3), 449-458.

Owen, R.J. (1969). *A Bayesian approach to tailored testing*. Princeton, NJ: Educational Testing Service.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351-356.

Pitoniak, M. J., & Sireci, S. G. (1999). *A literature review of contemporary standard-setting methods appropriate for computerized-adaptive tests*. Laboratory of Psychometric and Evaluative Research, Report No. 369. School of Education, University of Massachusetts, Amherst, MA.

Sireci, S. G. (2002). *Balancing efficiency and validity when setting standards on computer-based tests*. Paper presented at the annual meeting of the Association of Test Publishers, Carlsbad, CA.

Sireci, S. G., & Clauser, B. E. (2001). Practical issues in setting standards on computerized adaptive tests. G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 355-369). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. H. Wainer (Ed.), *Computer Adaptive Testing: A Primer* (2nd ed., pp. 101-133). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H. (2000). *Computer Adaptive Testing: A Primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). Future challenges. H. Wainer (Ed.), *Computer Adaptive Testing: A Primer* (2nd ed., pp. 231-270). 2000: Lawrence Erlbaum Associates, Inc.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. H. Wainer (Ed.), *Computer Adaptive Testing: A Primer* (2nd ed., pp. 61-100). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement*, 38(1), 19-49.

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 19-51). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Appendix

Table 1: Descriptive Statistics for Response Simulation

Table 2: Termination Criteria Test Results

Table 3: Alberta Learning Grade 6 Science Achievement Test Results on 2000 and
2001 Items

Table 4: Standards Achieved on the Alberta Learning Grade 6 Science Achievement
Test 2001

Figure 1: CAT Algorithm Flowchart

Table 1:

Descriptive Statistics for Response Simulation

Value	n	Mean	Standard Deviation
Examinee Ability (θ)	1000	0	1
a-parameter	200	0.6299	0.1299
b-parameter	200	-0.1413	1.2454
c-parameter	200	0.2495	0.0182

Table 2:

Termination Criteria Test Results

Examinee	Theta	n Items Termination			Measurement Precision Termination			
		n	SE		n		SE	
			Linear	Adaptive	Linear	Adaptive	Linear	Adaptive
1	1.81	25	0.6929	0.4291	114	72	0.2987	0.2996
2	-0.49	25	0.5852	0.3993	101	47	0.2999	0.3000
3	1.44	25	0.6435	0.4132	100	68	0.2989	0.2996
4	-0.92	25	0.6372	0.4342	122	46	0.2985	0.2991
5	1.63	25	0.6674	0.4286	106	66	0.2996	0.2992
6	-0.83	25	0.6236	0.3963	116	52	0.2999	0.2982
7	-0.68	25	0.6042	0.4237	109	52	0.2989	0.2993
8	0.34	25	0.5628	0.3654	84	52	0.2989	0.2977
9	-0.49	25	0.5852	0.3897	101	51	0.3000	0.2969
10	0.26	25	0.5612	0.3697	88	48	0.2934	0.2990
Mean	0.2070		0.6163*	0.4049*	104.1000*	55.4000*	0.2987	0.2989
Std. Dev.	1.0666		0.0443	0.0249	11.9578	9.5126	0.0019	0.0010

*: $p < .05$

Table 3:

Alberta Learning Grade 6 Science Achievement Test Results on 2000 and 2001 Items

	n	Average (2000)	Average (2001)
Common Items	20	15.0	15.4
Unique Items	30	20.1	21.3

Note. From Alberta Learning, personal communication, May 13, 2003. Adapted with permission.

Table 4:

Standards Achieved on the Alberta Learning Grade 6 Science Achievement Test 2001

Category	Maximum Possible Score	Cut-Score (2000)	Cut-Score (2001)	Expected Number	Actual Number	Expected Percent	Actual Percent
Acceptable Standard	50	26	27	33425	34540	85	87.8
Standard of Excellence	50	42	43	5899	10754	15	27.3

Note. From Alberta Learning, personal communication, May 13, 2003. Adapted with permission.

Figure 1: CAT Algorithm Flowchart

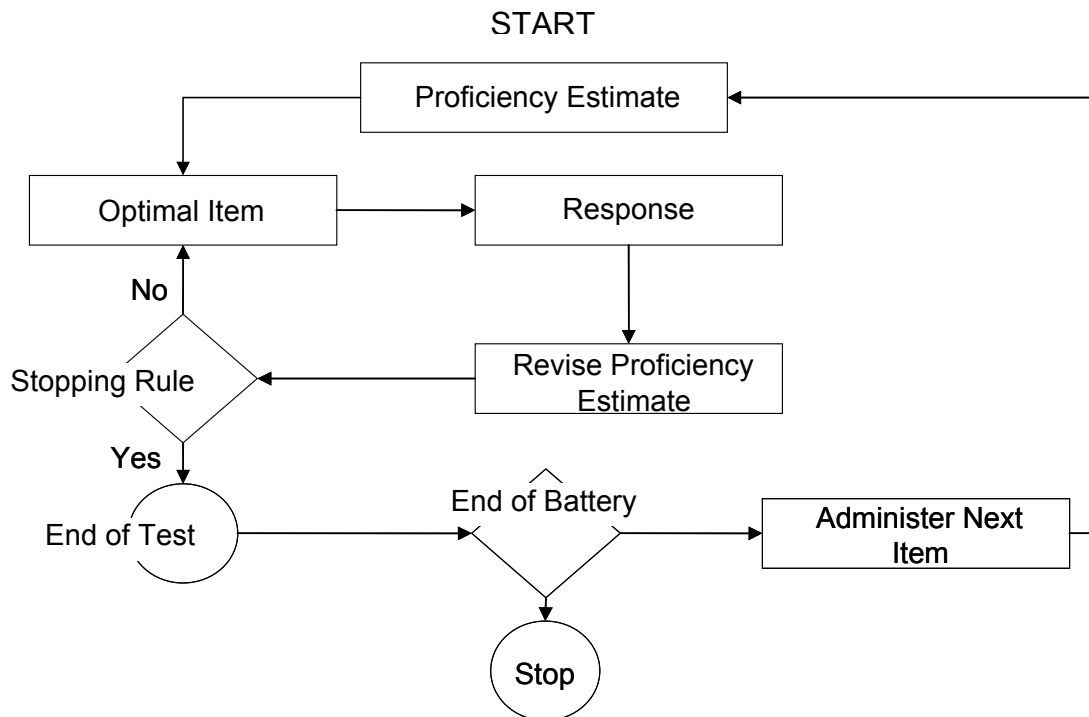


Figure 1: Item selection and ability estimation within a CAT framework.

Note. From "Testing algorithms," by D. Thissen & R.J. Mislevy, 2000, In H. Wainer (Ed.), *Computer Adaptive Testing: A Primer*. (p. 108). Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2000 by Lawrence Erlbaum Associates. Adapted with permission.