Design of a Computer-Adaptive Test to Measure English Literacy and Numeracy in the Singapore Workforce: Considerations, Benefits, and Implications

Jared Jacobsen
Senior Research Consultant
CASAS

Richard Ackermann
TOPSpro and CASAS eTests Manager
Team Code

Jane Egüez
Director, Program Development
CASAS

Debalina Ganguli
Director, Research and Analysis
CASAS

Patricia Rickard
President
CASAS

Linda Taylor
Director, Assessment Development
CASAS

Authors' Note

The authors declare no potential conflicts of interest with respect to the authorship and/or publication of this article.

The authors of this study disclose the receipt of financial support from the Singapore Workforce Development Agency.

The authors would like to thank Russell W. Smith, Ph.D., Senior Psychometrician, Alpine Testing Solutions, Inc., Dr. Soo-Yin Hee, Director Singapore Workforce Development Agency, and Christopher Yip, Deputy Director Singapore Workforce Development Agency for their valuable assistance in reviewing portions of this paper.

Design of a Computer-Adaptive Test to Measure English Literacy and Numeracy in the Singapore Workforce: Considerations, Benefits, and Implications

Abstract

A computer adaptive test CAT) is a delivery methodology that serves the larger goals of the assessment system in which it is embedded. A thorough analysis of the assessment system for which a CAT is being designed is critical to ensure that the delivery platform is appropriate and addresses all relevant complexities. As such, a CAT engine must be designed to conform to the validity and reliability of the overall system. This design takes the form of adherence to the assessment goals and objectives of the adaptive assessment system. When the assessment is adapted for use in another country, consideration must be given to any necessary revisions including content differences. This article addresses these considerations while drawing, in part, on the process followed in the development of the CAT delivery system designed to test English language workplace skills for the Singapore Workforce Development Agency. Topics include item creation and selection, calibration of the item pool, analysis and testing of the psychometric properties, and reporting and interpretation of scores. The characteristics and benefits of the CAT delivery system are detailed as well as implications for testing programs considering the use of a CAT delivery system.

Background

In 2004 the Singapore Workforce Development Agency (WDA) conducted an international search for assistance in providing the assessment foundation for its Employability Skills (ES) framework. The purpose of ES is to enhance the overall skills standards of Singapore workers at three levels: occupational, supervisory, and managerial. The framework is part of the Singapore Workforce Skills Qualifications System (WSQ), a nationally agreed upon competency-based skills development system in which skills acquisition, training, and assessment is based on industry-approved competency standards. The ES framework comprises two main series: the Workplace Skills (WPS) series and the Workplace Literacy and Numeracy (WPLN) series.

Singapore selected CASAS – Comprehensive Adult Student Assessment Systems – from the field of international organizations to develop the assessment foundation to test performance in the mastery of the English language for the ES. For this project, CASAS developed assessments for five different modalities. The Workplace Literacy component comprises the reading, listening, speaking, and writing modalities. The fifth modality is workplace numeracy. Together these assessments are referred to as the Workplace Literacy and Numeracy Assessments (WPLN).

Originally, assessments were delivered via traditional fixed paper-and-pencil test forms. However, after a short period of administering the assessments, CASAS and the Singapore WDA determined that this was not the most efficient delivery mode due to the number of examinees being served, the need for very quick turnaround time of test results, and the inability to leverage other benefits associated with CAT administration. These benefits are discussed in this article. The CAT delivery mode was then chosen for three modalities – reading, listening, and workplace numeracy. This paper focuses on the modalities delivered through CAT. The assessments are referred to as the WPLN CAT.

Purpose and Objective of the Assessments

The purpose of the CASAS assessments that comprise the WPLN CAT is to assess the level of competency in English literacy and numeracy of the Singapore workforce in a life skills context. The reading domain of the WPLN CAT measures examinees' ability to read and to be understood in a work environment. The listening domain measures examinees' ability to listen and to understand in a work environment. The numeracy domain of the WPLN CAT measures examinees' ability to use basic and fundamental arithmetic skills to identify, locate, act upon, interpret or communicate a problem; provide information about mathematical ideas including quantity and number, dimension and shape, pattern and relationship and change; represent information about mathematical ideas using objects and pictures, numbers and symbols, diagrams and maps, graphs and tables and text.

Individuals who seek training are administered a WPLN CAT to determine their English literacy and numeracy ability as indicated by a CASAS scale score. The scale score is indexed to achievement levels that allow participants to be placed into an appropriate level of training where effective learning can take place. The WPLN CAT is designed for pre- and post-testing and it equally serves to monitor learning progress.

The WPLN CAT reports an examinee's proficiency in English literacy and numeracy on a level scale of one to eight, which encompasses a continuum of skills from pre-beginning through proficient use of language and numeracy. The assessments are designed to assess and appropriately place examinees across the full range of the eight WPLN functioning levels in each content domain. This continuum provides descriptions of examinees' general job-related ability in reading, listening, writing, and speaking for each proficiency level.

Based on an examinee's performance, a statement of attainment (SOA) is issued for each content modality. The SOA is a nationally recognized qualification indicating an individual's ability or competence in a particular area. More than 1,500 employers and training institutions in the tourism, food and beverage, retail, healthcare, logistics, manufacturing and security sectors recognize the WPLN credentials. Achievement of an SOA, via a score on the assessments, means that the examinee is certified by the Singapore WDA as being able to perform a task appropriate to a corresponding occupational level. The partnership between employers and the WDA is imperative to the sustained administration of the WPLN CAT and the delivery and use of results.

In addition, examinee performance on the WPLN CAT qualifies individuals for additional training designed to continue to develop skills that will be recognized by employers.

Because of the multiple purposes of the CAT, for many examinees there is a need to both classify the examinee into a performance level and give a precise score measuring performance. This represents an important consideration and challenge – the scoring of an examinee's performance must be equivalent and also appropriate for classification into SOA levels. This dual purpose influences the amount of precision desired for the performance scores. Special attention must be given to standard errors (SE) in the "tails" – the highest and lowest performing examinees on the scale. One approach is to first focus on the tails of the scale. By first achieving the desired precision on the tails, the middle-level score points should also be within the desired degree of precision.

CAT Design

The WPLN CAT is conducted in a one-and-a-half hour time period. It draws items from the item pool and is designed to be used by the Singapore WDA to assess examinees in an efficient period of time while providing a score that is reliable for placing them into an appropriate training level, issuing Statements of Attainment (SOA), and measuring learning gains over time.

Item Pool

To operate a WPLN CAT, it is necessary to have a significant pool of calibrated items from which to draw (Wainer et al., 2000). The procedure for drawing and calibrating items depends on IRT and the specific model selected, in this case the Rasch model. While there are no specific guidelines for the number of items that should comprise the pool, a number of factors were considered to determine the appropriate size and continued expansion of the item pool. What is essential for any pool of items is that they span the full range of trait levels in the population served by the assessments. Each trait level must include items covering all content areas. All items are chosen based on the competency and content standard coverage necessary to assess the target population for the purposes of the WPLN.

The total number of items in the pool must be sufficient to supply information throughout a testing session over time. A main consideration is the determination of the acceptable level of item exposure. The acceptable level is based, in part, on the stakes of the test. A high stakes test can be thought of as a test that has direct and major consequences or is the basis of a significant decision. Typically a high stakes test is a test that has the potential to deny an examinee access to education or employment. For the WPLN CAT, examinees are classified into levels, and decisions are made based on these classifications. This use of the test is considered in the determination of acceptable levels of item exposure. Item exposure should be reviewed for all items in the pool. Those items identified as having higher exposure rates should be focused on in an examination of the performance of the item pool, via analyses such as an item parameter drift analyses, and when planning new item development. The integrity of the CAT is dependent upon the item parameters remaining stable. There are various options for controlling for item exposure; these include algorithms such as creating a table where all items of equal difficulty are exposed randomly. Another option is retiring items that have a relatively high exposure rate for a specific time period and then reintroducing the items to the item pool.

The items in the pool must have characteristics that provide adequate information at the proficiency levels that are of greatest interest to the stakeholder. In the case of the WPLN CAT there must be sufficient items to measure performance equally at all score points because assessments are used to measure performance and progress along the entire scale. The standard error of measurement (SEM) at each score point must be low enough to give the desired level of precision. Likewise, the SEM at each cut point must be low enough to allow score points to reliably classify examinees into different performance levels.

Along with the monitoring of items to analyze how often each item appears on a testing event, during the development of the WPLN CAT it was important to communicate with the Singapore WDA regarding the estimated demand for the assessments. This was necessary to estimate future absolute item exposure. In addition, estimates were made for the replacement, revision, and retirement of items over the course of the life of the WPLN CAT. Based on these estimates, to allow for the continuous replenishment of the item pool, an item development plan was proposed. This was fundamental to promote the sustainability of the WPLN CAT.

In the original development of the item pool for the WPLN CAT, many existing items from the item pool were determined to be appropriate for the WPLN assessment without any changes. Following a review by subject matter experts, other items deemed inappropriate for the Singapore context were removed from the item pool. A third category of items were flagged for modifications. The majority of those items modified to fit into the Singapore context needed minor changes, such as personal and place names and vocabulary differences between American

English and British English. Edits were made in such cases and a Singapore version of the item was created. The modified versions of the items were then added as new entries into the master item pool. Prospective items were dropped from consideration where the proposed modifications would affect the viability of the item, change its difficulty, or have a potential impact on the examinee's answer selection. Some numeracy items involving U.S. units of measure were altered to metric measure as suggested, but most were dropped from consideration, as changes would have affected the math calculations involved.

IRT - Rasch Model

The Rasch model was selected for the computer administered assessments because of its model fit and scoring advantages. The Rasch model is governed only by the difficulty of the item and the ability of the person located on the same continuum (Rasch, 1980). Although the Rasch model appears on the surface merely as a reduction of the 2PL and 3PL models, it has some features that are not shared by those models. One is that raw score is a sufficient statistic, which means that if one examinee answers more items correctly than another, the scale score for the first examinee will be higher than the second if both examinees take the same set of items. The Rasch model specifies a uniform discrimination and zero left asymptote in order to sustain sufficiency of simple, unweighted raw score. Discrimination does not influence standard error. These characteristics are important given the scoring system used for the Singapore WDA and the multiple uses of the test scores. Test scores need to be explained to different parties, examinees, educators, and employers, with as little complexity as possible. This is facilitated by the Rasch model.

With 2PL and 3PL models the raw score may not be a sufficient statistic because their item characteristic curves are allowed to cross.

The CAT Mechanism

The adaptive mechanics of a CAT is illustrated in Figure 1. The figure represents a single examinee responding to test items. The first item is drawn randomly from the middle of the ability range, in this case an item with a Rasch Unit (RIT) value of 202 on the CASAS scale. The examinee answered the first item correctly as shown by the blue dot. Consequently, the second item will be more difficult than the first, which the examinee also answers correctly. The third item is finally too difficult for the examinee, thus the CAT responds by drawing a fourth item at a lower ability level. This adaption to the examinee tends to smooth out fluctuations seen at the beginning of the test as the true ability level of the examinee is reached. At the same time, the standard error (SE) is recomputed after each item is administered. The test begins with a high SE because there is little information at that point regarding the interaction between examinee and test. As the test progresses the SE steadily decreases until either the SE threshold or the maximum number of items is reached. When examinees reach either of these thresholds, they receive a score which corresponds to a Statement of Attainment (SOA) Level.

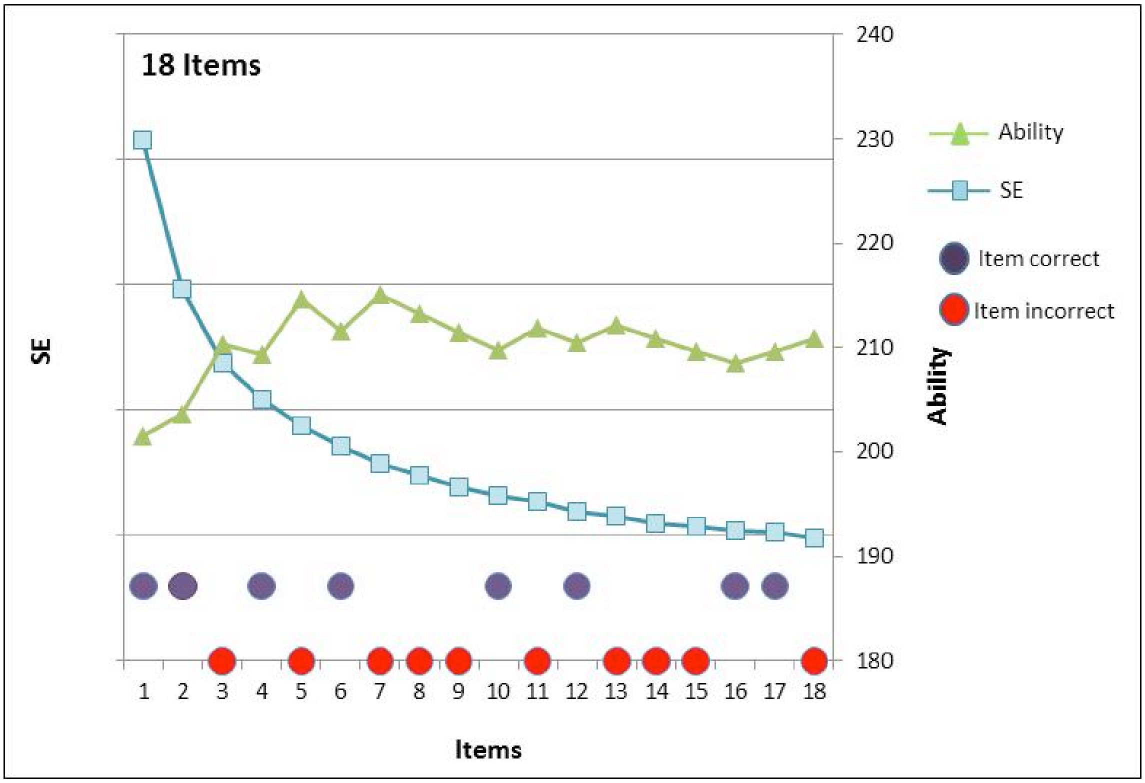

Figure 1. Graphical Representation of the CAT Mechanism

Unidimensionality and the CAT

Fundamental to all IRT models is the notion that a test measures a single, underlying construct. This is referred to as unidimensionality. The assumption is that the items in a test are homogenous and are measuring a single trait. As with other measurement models, unidimensionality is necessary for CAT delivery because the software must dynamically assign items on the same scale that are either easier or more difficult based on examinee ability level. The score it provides must be valid even though two examinees may be administered a different set of items. Multidimensionality confounds interpretations of any scores, but particularly a CAT because it introduces ambiguity about the implication of a correct or an incorrect answer.

Start Rule

The implementation of an appropriate start rule is a fundamental part of the design and plays a large role in how the design supports the classification consistency of examinees.

When examinees take their first test, a precise measure of their ability level is not known. Therefore, the administration of the first item is defined arbitrarily with respect to individual examinee ability. Under these circumstances, consideration must be given regarding the choice of an appropriate start rule. Generally, it is optimal to start with an item of middle difficulty with respect to the pool of items, which presumably has been calibrated to serve a particular population. A counter argument is that examinees of low ability may be frustrated by an item of middle difficulty that is already too hard for them. Psychometrically, it is acceptable to begin with either a low or middle difficulty item, but a potential drawback of starting with a low-level item is that middle and high performing examinees may need to see more items before they reach their reliability threshold. Because scores in the Singapore population were found to have a normal ability distribution on the CASAS scale, somewhat skewed toward the high end, the first item is drawn from the middle of the range. To limit exposure of the first item, that item is drawn randomly from a group of items with individual item difficulties within a ten-point scale score range. One option is a dichotomous variable measuring an examinee's previous years of education. Examinees would receive the initial item from one of two groups of items based on previous years of education. When an examinee returns to take a post-test, the first item is determined by the ability level achieved on the pretest.

Scoring Method

The scoring method is already determined by selecting the Rasch model. The algorithm is enhanced by maximum likelihood estimation (MLE). However, MLE has a disadvantage in that it cannot provide an estimate if the examinee provides all correct or all incorrect answers. In practice, this situation is seldom encountered. It is also aided by the stopping rule which includes a maximum number of items. It is important that all scores on a scale are interpretable and that there are not scores, particularly in the tails of the scale that do not yield interpretable results.

Item Selection Rule

After selecting the first item and estimating a score, the CAT engine must have a design for continuing the test. There are three features of item selection that might be accounted for in the CAT algorithm, the first of which is essential: psychometric characteristics, item exposure, and content balancing. Content balancing may be implemented both in terms of the control of content standards or competencies and in terms of item type or task areas. Item exposure is addressed with regard to the first item as described under start rule and can be addressed in a variety of ways as described in the *Item Pool* section. The item selection rule is described by the maximum likelihood estimation, wherein the CAT algorithm for this program uses maximum information as the psychometric selection criterion. Each item is chosen to maximize information about the examinee's current ability level.

Stop Rule

The rules for terminating a CAT differ depending on whether the CAT is designed for equiprecision or classification. The CAT designed for the WPLN is constructed for both equiprecision and classification so that a precise score can be provided, learning gains can be measured, and the ability to accurately classify into performance levels is offered. Two stopping rules are employed, a standard error of measurement (SEM) threshold, and a maximum number of items. The implied tradeoff of SEM to test length is intended because of the stipulation that the test be as short as possible while retaining an adequate level of accuracy. In the case of an equiprecise CAT which also serves to classify examinees, a consistent SEM at all score points is desired. With this in mind, the goal is for the item pool to retain a relatively equal distribution of items across performance levels. Although a higher percentage of examinees are performing at a mid-level ability, special care needs to be taken to assure that sufficient items are at the low and high end of the ability range. With this mind, more items may still be needed at the mid-level ability ranges due to the larger number of examinees functioning at these levels and, consequently, higher item exposure.

Benefits of CAT for Singapore WDA

There are a variety of characteristics that make the CAT appropriate and advantageous for assessing ability in English and Math proficiency under the WPLN.

Shorter tests

Fewer items are needed to measure an examinee's performance level compared to the traditional fixed paper-and-pencil test forms. This, and the resulting reduction in administration time, produces a better test-taking experience for the examinee. This feature was particularly attractive for the WPLN because of the volume of examinees, staffing considerations, and the administration of tests at remote locations. In many cases examinees are being assessed during work time and it was especially important that test-taking time be economized.

Equiprecision of test scores

CAT delivery in the WPLN results in comparable standard errors of measurement (SEM) at scale score points when analyzed against scores achieved by examinees who are administered a fixed form paper-and pencil (PPT) or linear computer-based test (CBT). When designing the CAT to achieve a desired level of precision, the test developer must consider the purpose of the test. As previously mentioned, a consistent level of precision is necessary for testing ability in the WPLN across all score points including decision points used to place examinees into functioning levels.

Timely reporting of test results

Because of the multiple purposes of the assessments, including the awarding of Statements of Attainments (SOAs) that may be provided to employers and guide the placement into training programs, it is important to provide examinees with scores in as timely a manner as possible. The CAT delivery was deemed the most efficient solution to maximize the timeliness of score reporting.

Increased security

The random exposure of items is a security benefit for the WPLN. Fixed test forms, in which groups of students receive the same exact items in the same exact order, could result in security issues. In addition, experience has shown that the CAT allows for improved accountability in terms of test materials. CASAS continues to focus on ways to leverage the use of technology to provide the appropriate levels of test security of the WPLN CAT. One of the most significant ways to leverage this technology is in the efficient retirement and replacement of items that have been over-exposed or had their security compromised. When delivered via a CAT, items can be retired and replaced in an efficient and economical manner. When delivered via a fixed form paper-and-pencil mode, the process of removing items can be logistically difficult and require an extensive recall of test forms.

Administration to all ability levels

As mentioned, the scores in the target population of examinees using WPLN were determined to have a normal distribution of scale scores, somewhat skewed toward the high end when compared to other populations using the CASAS tests, but consisting of examinees across the entire spectrum of ability levels. This distribution is regularly examined by the continual monitoring of score distribution. Therefore, it is very important to have a delivery system that could measure performance across the entire range of ability levels. CAT delivery met this requirement without requiring the development of a large number of fixed test forms to assess examinees across a wide range of ability levels. In addition, the CAT delivery system was deemed the most efficient in assessing examinees across the full ability spectrum without the additional step of having to administer an appraisal to determine which fixed test form was appropriate to begin the assessment process. This, along with the appropriate fit of the Rasch model which operates under the assumption that a difficult item is more difficult for all ability levels and a less difficult item is less difficult for all ability levels, leads to a model and delivery mode appropriate for assessing all ability levels.

Field testing of new items

The replenishment of items in the item pool is crucial to maintaining the integrity of test scores. It was initially determined that CAT delivery would most efficiently facilitate the field testing of new items. Items can be embedded relatively easily into a testing administration without the need of creating or revising fixed test forms. Also facilitated is the collection and analysis of item statistics and the eventual incorporation of new test items into the item pool.

Given these efficiencies, challenges were also encountered during the process of incorporating field test items into the normal delivery of the WPLN CAT. A challenge was a sparse matrix and the resulting inability to anchor the calibration. Therefore it was necessary to imbed a significant number of linking items in addition to the actual field test items. This led to a test length beyond what was planned which made the field test process more time-consuming for examinees. It is important to find a balance between effectively field-testing the appropriate number of field-test items yet not overburdening examinees with field-test items so as to potentially cause examinee fatigue or lengthen the testing event beyond the time allotted. In the case of the WPLN CAT examinee fatigue was not deemed an issue due to test length. Comparative studies can be conducted to measure test taking time and performance for examinees who are tested with embedded field-test items compared to those who are not. In addition, linking items can be chosen so they are part of the testing event and in the calculation of the examinee's scale score. One option is to design a type of hybrid system in which the field-tested items are in a fixed testlet form and all examinees receive both this testlet and the adaptive portion of the assessment. The total score is based on all items from the testlet and the adaptive portion. The benefit and purpose to the testlet not being adaptive is that the low-level examinees still receive some more difficult items and high-level examinees still receive some less difficult items.

Promote use of technology

The administration of assessments on computers, both at the WDA testing center and at remote locations, allows examinees to use technology that they will encounter in the workplace. The Singapore WDA viewed the necessary investment in technology as an additional benefit to implementing CAT as it promoted the expanded and efficient use of technology. Technology, in terms of necessary investment outlays, must be determined and communicated at the outset of the project. In addition, vendors must also remain current on the constant technological advances and potential upgrades and collaborate with the Singapore WDA regarding the appropriate implementation of new technology as necessary to further enhance the WPLN CAT.

Simulations

Using a CAT facilitates the use of simulations to aid in the design and maintenance of the assessment. For example, simulations can be set up to test how the assessment performs when different criteria are used to determine, the start rule, stop rule, item selection, and other CAT characteristics. This will help plan the initial size of the item pool and item pool expansions needed to achieve desired precision levels. Running the simulations with a wide range of hypothetical criteria will illustrate the resulting trade-offs that occur due to design and policy decisions, for example the level of precision that will be achieved at a range of test lengths. In addition, CAT properties such as item exposure can be simulated which will help in planning item exposure tolerance or in the analysis of the cost of various security levels.

Psychometric Process

Appropriate analyses and research studies must be conducted to provide evidence of the psychometric properties of the assessment system.

During the field-testing process, and on a continual basis, classical item statistics are reviewed for all items in the item pool that comprise the WPLN assessments. Because many of these items were originally placed on fixed paper-and-pencil forms, item statistics needed to be analyzed and continually monitored with respect to their performance when delivered via the CAT in the WPLN.

Demographic Characteristics of Study Population

Analyses and research studies are conducted taking into consideration the demographics of the target population, in this case examinees tested in the WPLN. When an assessment is designed for a new target population the analysis of the demographic characteristics of this population is necessary even if it is deemed that the purpose of the assessment and the skills measured are unchanged.

Fairness and Sensitivity Review

All items are reviewed for fairness and sensitivity. A diverse panel of subject matter experts (SMEs) reviews all items based on the *ETS Fairness Review Guidelines* (Educational Testing Service, 2008). Also the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and *Educational Measurement* 4th edition (Brennan, 2006) are used in developing the fairness and sensitivity review policy. When items are used in a new system, culture, or country, the fairness and sensitivity review must include processes to ensure that all items are appropriate for the country or culture.

All items are then analyzed for potential bias using a Differential Item Functioning (DIF) analysis. Special attention is given to any items that were cited by the SMEs who reviewed items on the fairness and sensitivity panel. CASAS uses the Mantel-Haenszel statistic to compare the performance on an item of a "focal" group to that of a "reference" group matched in overall ability (Dorans & Holland, 1993; Holland & Thayer, 1988). For the items being used in the WPLN CAT, the DIF analyses focus on analyzing the consistency of performance among the focal and reference groups in gender, age, and native language categories. Analyses with respect to native language are important because of the high immigrant population in Singapore. From the outset systems must be in place to collect all relevant data necessary to conduct the appropriate DIF analyses for all population subgroups.

Standard Setting

CASAS regularly evaluates the "decision" or "cut" points that define the WPLN CAT Proficiency Levels. Separate performance level standard setting studies are conducted for each skill area or "modality" – reading, numeracy, and listening. The cut scores and scale are reviewed for consistency with the reporting and analytical guidelines and standards established in the *ETS Standards for Quality and Fairness* (ETS, 2002).

The appropriate standard setting method must be chosen. CASAS has typically validated cut-scores using the Bookmark standard setting procedure (Mitzel, Lewis, Patz, & Green, 2001). The implementation of the Bookmark method follows the general guidelines outlined in *A Primer on Setting Cut Scores on Tests of Educational Achievement* (Zieky & Perie, 2004). Other independent evaluations by instructors or other persons who interpret examinee scores should be conducted to continually validate the classifications of examinees.

During the item development and design process a main consideration is that appropriate coverage must be achieved across all performance levels being measured and at all decision points. Subsequent item development, to ensure the continual validity and reliability of the assessment, is focused on the continual need to maintain and enhance the coverage needs to ensure reliable measurement across all performance levels and at all decision points.

Test Information Function

To examine the information being provided by the WPLN CAT assessments, the Test Information Function (TIF), a sum of all the item information functions within the test, has been a useful tool in measuring the reliability of a test. This function reports the amount of information being provided by the test and over what ability range. It is a useful tool to examine and evaluate the degree to which the test is discriminating and providing sufficient information across the appropriate ability range.

Item Exposure Information

The exposure rate is able to be continually monitored. CASAS examines the exposure rate as expressed as the percentage of test administrations that each item has appeared on. This, along with data on the total number of tests administered, provides current information on item exposure. This information is used to analyze the security or maintenance of the assessment. If facilitates decisions on the retiring of items and future item development.

Concurrent Calibration Correlation

To examine the reliability of performance on the WPLN CAT items, correlations are calculated between Calibrated RITs and the Rasch Item Measure computed from the concurrent calibration of the WPLN CAT item vectors. This provides evidence of the consistency of item difficulty across the item difficulty measures and, as a result, the reliability of the measurement.

Score Distribution Information

The distribution of test scores is monitored across test modalities, among demographic sub-groups, over specific time periods, and among different testing sites. The percentage of examinees that are placed into each of the WPLN proficiency levels is calculated and analyzed.

CAT Functioning

Characteristics of the CAT delivery system are continually monitored across all items and assessments in the WPLN assessment system. The mean ability estimate and mean standard error of measurement for examinees taking the WPLN Assessments are calculated at each score point – from item one to the maximum number of test items.

The percentage of WPLN Assessments that reach desired SEM is calculated compared to the percentage of assessments that reach the item maximum. An example is shown graphically in Figure 2.

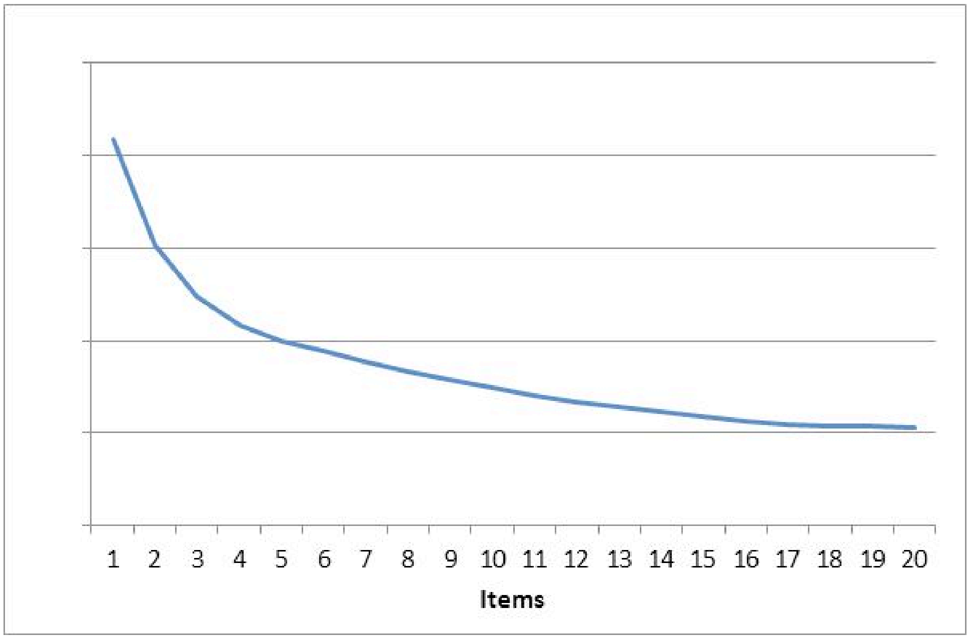

Figure 2. Example Graphical Representation of Mean Standard Error of Measurement

The fit of the Rasch model should be continually analyzed. For example, as previously mentioned, the measurement and determination of unidimensionality is necessary. To examine unidimensionality information, the total variance accounted for by the Rasch measures, the Rasch person measures and the item measures are compared to the total variance of the empirical observations. This then allows the analyses of the unexplained variance in the observations. To summarize, the total percent of the empirical variance and the modeled variance in the observations that is accounted for by the common unidimensional person and item measures is calculated. After extracting the variance due to the measures, standardized residual analysis is conducted to evaluate if there is any significant unexplained variance that could be accounted for by different principal component contracts. Based on the results, some items are investigated further to evaluate multidimensionality.

Construct Validity Evidence – WPL Reading and WPN Correlation

An analysis of the relationship between assessments in different modalities, for example, the WPL Reading and Workplace Numeracy (WPN), is conducted to examine construct validity. Data on examinees who take both a reading and a numeracy assessment is analyzed. Correlation results show that while there is a relationship between the two scores, we would expect higher ability examinees to score higher on both tests and lower ability examinees to score lower on both tests, there is evidence that the reading and numeracy items are not measuring the same construct.

Criterion-Related Validity

Criterion-related validity evidence supports the extent to which the results of an assessment correlate with a current or future event. Another way to think of criterion-related evidence is to consider the extent to which the examinees' performance on the given task may be generalized to other relevant activities (Rafilson, 1991).

To examine criterion related evidence, the average CASAS scale score obtained by examinees on the WPLN assessments is compared to their education level using a variety of measures. These measures include highest level of education achieved (i.e. primary, lower secondary, upper secondary, diploma, Masters/Ph.D.) and A Level, N Level, O Level which are academically oriented qualifications used in Singapore. In addition it is beneficial to research other assessments intended for the population of interest and design comparative studies to examine the relationship of examinee performance on two different assessments designed for the same purpose. It is important to note that these studies can be challenging to coordinate and conduct. They tend to be relatively expensive and time consuming. An appropriate number of examinees must be convened and be administered two assessments in a short time period to eliminate learning effects and yield statistically reliable results. There are various considerations in choosing an appropriate test to administer for comparative purposes. This test must have the same purpose, be measuring similar skills, and be designed for the same population. The psychometric properties of this assessment should be verifiable to ensure that this assessment is appropriate to use for comparative analysis. There are other logistic considerations, such as the availability of trained personnel qualified to administer and score the assessment for comparative purposes.

Descriptive and Reliability Summary Statistics

Appropriate descriptive and summary statistics are analyzed on a continual basis to examine evidence that the tests continue to provide valid and reliable measures. These summary statistics will differ from those analyzed from traditional fixed paper-and-pencil assessments. Examples of the statistics include mean starting and final ability measures for examinees, mean starting and final standard errors for examinee score points, raw score to measure, item reliability, persons reliability, and empirical reliability.

Parallel Test Administrations

CASAS also conducts studies to examine evidence of score comparability and reliability across parallel test administrations. Each set of traditional, fixed paper-and-pencil parallel forms consisting of items from the CASAS item pool are administered to examinees on separate testing events. Careful attention is paid to the research design which includes limiting extraneous factors that could influence examinee scores between the two testing events such as instruction. Careful selection of participating examinees is conducted to help ensure that a proper sample of motivated examinees is used.

Although the previously cited studies of parallel form reliability provide supporting evidence, the most relevant analyses measure the reliability of repeated test administrations in the environment and delivery mode in which the test is being used. For the WPLN CAT it is necessary to examine the comparability of scores across multiple tests administered to the same examinee. Given that the CAT classifies examinees into specific performance levels, of particular importance is classification consistency. Classification consistency must be thoroughly examined both in terms of how the CAT design supports the consistency and through the use of empirical evidence which indicated appropriate levels of classification consistency.

Additional studies that can be undertaken to provide evidence of reliability in repeated test administrations include the administration of the assessment twice with mutually exclusive item sets. Because the WPLN CAT is designed to predict performance at the time of the testing event and before significant additional training or instruction has taken place, the retest should be administered at the same testing event or as soon after as possible.

Implications for Practitioners

A variety of implications for entities considering the development and implementation of a CAT delivery system have been presented in this article. The CAT design must take into account the logistics of the testing environment to ensure proper test length, and test taking time, while still achieving reliable results with an appropriate standard error of measurement (SEM) based on purpose and use of the test. Given the design requirements, the item pool must be of sufficient size taking into account the desired score range and content coverage across all performance levels measured. The blueprint for item development must incorporate appropriate strategies for bolstering the item pool as needed.

The purpose of the test must be clearly stated and agreed upon to ensure both the test and the delivery system are effective and efficient. The assessment system designed for Singapore needed to have a developmental perspective as opposed to just a one-time “snapshot” of a student’s ability at a given point in time. The assessment system was designed to assess the development of student understanding of particular concepts and skills over time – with items serving as a framework for the assessment and to make measurement possible. The assessments being based on a learning progression framework, with both equiprecision and classification necessary across a range of examinee ability levels, has implications across various procedures of test development. These include the size of the item pool, the degree of item coverage across all ability levels and content areas, the amount of precision desired and the acceptable SEMs, and appropriate test length including the starting and stopping rules.

The ability, desire, and commitment to invest in the initial purchase and maintenance of the necessary technology to deliver the assessment must be present. While the advantages of a CAT may be very apparent the initial design and implementation is time consuming. This makes the commitment to the analysis and examination of all relevant psychometric properties, while meeting the needs for timeliness in implementation and enhancement, a joint challenge requiring the cooperation of the test developer and the entity using the CAT.

Methods to effectively leverage and promote all benefits of this investment in technology should be pursued to maximize the return on the investment. These include benefits in the field testing of new items and conducting simulations to assist in the development, maintenance, and enhancement process. Enhancements must be considered over time to take advantage of technological advancements. For example, in the Singapore WPLN program the movement from WAN to Internet-based delivery. It is important to emphasize that while a CAT delivery system may have a more diverse variety of potential enhancements, all enhancements must take into account the ramifications and how they affect the psychometric properties of the assessment. For example, studies have shown that changes in components of a CAT, such as exposure rate control, content balancing, test length, and item pool size result in different levels of comparability in test scores.

Extensive analysis and examination of the psychometric properties of the assessment with a focus on the model fit is imperative. When the test items are used in a new mode of delivery, the item and assessment performance must be thoroughly reviewed using a population with demographic characteristics that are representative of the new target population. Additional research is imperative to provide evidence that the content and constructs remain valid as used within the system. This is necessary even when the assessment has the same purpose and measures the same skills. Additional reviews for fairness and sensitivity are important to ensure that items continue to use appropriate terminology, avoid unnecessarily controversial material, and represent diversity in their depictions of persons. Subject matter experts (SMEs) must reflect the demographics and cultural characteristics of the target testing population.

In many cases criterion based evidence may be of especially high importance. This criterion evidence will compare the scores and functioning levels on the CAT to other measures of educational functioning that have been used to determine an examinee's ability. It is important that the test developer analyze the current educational levels of the target examinee population and how these educational levels are currently measured and reported. This is essential to the comparative analyses between the newly created educational levels, or standard setting, and existing measures. It is also imperative in the process of effectively training those who are interpreting the scores and are accustomed to interpreting scores from the other measurements used to determine an examinee's ability level. In the case of the Singapore WPLN program it was important that the new assessments provide organizations an additional means to determine ability level. Traditionally, after the completion of primary education students are administered the Primary School Leaving Exam (PSLE) to determine the level of performance. Based on the results of the PSLE students are placed on different secondary education tracks which include the *Normal* track which is a four-year course leading up to the N-Level exam or the *Special* or *Express* which are also four-year courses which lead up to the O-Level exam. The *Normal* track also has the possibility of a fifth year followed by an O=Level exam (Singapore Ministry of Education, 2011). Previously many organizations required that candidates had an N or O Level of education. However, statistics showed that there are about half-a-million adult workers who had missed out on education that would allow them to reach the N or O level of education. The WPLN provides an alternative qualification for these adults to demonstrate their ability levels and qualify for training to gain additional skills that will be recognized by employers. Independent analyses can also provide valuable validity evidence. In the case of the Singapore WPLN program, a valuable reference is the *National and International Benchmarking of WDA Workplace Literacy and Numeracy Qualifications.* (UK NARIC, 2008).

The scoring rubric must be appropriate for the system into which the CAT is being implemented and standard setting must be properly determined for the system. It is necessary that validation studies are conducted on a continual basis for all established decision points. To supplement standard setting studies such as the Bookmark method or the Angoff method (Angoff, 1971), it is valuable to use methods based on judgments regarding individual examinees. When results are used by other entities, such as in the Singapore WPLN program, this should include feedback from these entities as part of the continual validation process. It is also valuable to examine decision points and examinee placement through the independent judgment of instructors using surveys to capture instructor feedback. Evidence of classification consistency at decision points must be thoroughly presented both in how the design of the CAT promotes this consistency and empirical evidence which demonstrates the desired degree of classification consistency.

Analysis of the intricacies of the delivery system is necessary to address appropriate security measures, individual testing needs (such as accommodations), and reporting requirements. The desired level of security must be determined and measures must be established to quantify this level to continually evaluate the level of security. The implications of different levels of security, in terms of the development, maintenance, and enhancement cost, should be analyzed.

The appropriate reporting of scores, to both examinees and other parties that may use and interpret test scores, must be developed and implemented. In the case where different parties with different purposes interpret scores, for example educators and potential employers, it is important to customize the delivery and training on how to effectively and efficiently interpret these scores. Any potential misinterpretations or improper usage of scores must be initially identified and mitigated. This includes interpretations by educators that use results to guide further instruction and employers that use results to estimate ability to perform a job function. There must not be scores in the assessment that yield uninterpretable results.

Providing feedback to educators within the CAT framework involves new challenges because of the dynamic and adaptive nature of the test administration. Because not all examinees receive the same items on a given test report there are new considerations in the development of performance reports, at a class or group level, for educators. If a program is accustomed to administering fixed form paper-and-pencil tests it will usually have easy access to reports that detail which items, and therefore which competencies, the class performed well or poorly on as a whole. With CAT this reporting becomes more difficult because all examinees do not receive the same set of items and, depending on the design, competencies. This can make determining the specific areas where more preparation is needed, at a class level, more difficult. It is critical that this issue is addressed when the assessment design is based on a developmental or learning progression perspective in which results will guide future instruction and training. As with any assessment used by educators to guide instruction, it is crucial that these educators interpret results, see the benefit in these results, and are trained effectively to use these results. With this in mind, instructionally relevant and actionable feedback that is provided to instructors from CAT results can be qualified as necessary. If class level information is provided in content categories, how many examinees saw items in these categories and how examinees many missed items in these categories may also be reported to help instructors make instruction and curriculum decisions.

References

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Angoff, W.H. (1971). Scales, Norms and Equivalent Scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600) Washington, D.C.: American Council on Education.

Brennan, R. L. (2006). *Educational Measurement* (4th ed.). Westport, CT: American Council on Education and Praeger.

Dorans, N. J. & Holland, P.W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P.W. Holland and H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Educational Testing Service. (2008). *ETS Fairness Review Guidelines.* Princeton, N.J.: Author.

Educational Testing Service. (2002). *ETS Standards for Quality and Fairness.* Princeton, N.J.: Author.

Holland P. W. & Thayer, D.T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer, and H. I. Brown (Eds.), *Test Validity* (pp. 129-145). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.). *Setting performance standards: Concepts, methods and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates.

Rafilson, F. (1991). The case for validity generalization. *Practical Assessment, Research & Evaluation*, 2(13). Washington, D.C.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedagogiske Institut, 1960. Reprint, Chicago: University of Chicago Press.

Singapore Ministry of Education (2011). *Primary School Leaving Exam*. Retrieved from http://www.moe.gov.sg/

UK NARIC (2008). *National and International Benchmarking of WDA Workplace Literacy and Numeracy Qualifications*. Retrieved from http://app2.wda.gov.sg/data/imgcont/936/Full%20Naric%20Report%20-%20last%20updated%20200409.pdf

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., Thissen, D. (2000). *Computer Adaptive Testing: A Primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Zieky, M. & Perie, M. (2006). *A Primer on Setting Cut Scores on Tests of Educational Achievement*. Princeton, N.J.: Author.