

Running Head: PARALLEL FORMS CONSTRUCTION

**Automated Test Assembly Procedures for Criterion-Referenced
Testing Using Optimization Heuristics****

Keith A. Boughton and Mark J. Gierl
Centre for Research in Applied Measurement and Evaluation
University of Alberta

Paper Presented at the Annual Meeting of the
American Educational Research Association (AERA)

New Orleans, Louisiana, USA

April 24-27, 2000

** This paper can also be downloaded from the Centre for Research in Applied Measurement and Evaluation (CRAME) website: <http://www.education.ualberta.ca/educ/psych/crame/>

Abstract

There is a growing interest in automated test assembly algorithms and optimization heuristics. Combined with recent advances in computer technology, computer-assisted optimization models and procedures are now able to build parallel forms that meet complex content constraints and statistical targets. The present study was designed to investigate the efficiency and effectiveness of the computerized-adaptive sequential testing (CAST) procedure, which used the normalized weighted absolute deviation heuristic, to build parallel forms for criterion-referenced achievement tests and medical licensure exams. A comparison of the parallel forms produced in each situation provided an idea of how well the testing objectives were met under highly constrained but realistic conditions. The constraints that were studied include: content area, cognitive level, test length, item exposure, statistical target, and number of parallel forms. The results from this study indicate that the CAST procedure successfully solved each assembly problem. Many testing organizations could use this approach to create multiple parallel forms for their criterion-referenced tests if relatively well developed items banks are available.

Automated Test Assembly Procedures for Criterion-Referenced Testing Using Optimization Heuristics

Computer adaptive testing (CAT) can improve the measurement process by reducing testing length, improving test administration standardization, increasing measurement precision, improving testing security, increasing the flexibility for examinees by allowing for testing on-demand, and scoring and reporting results immediately (Sands, Waters, & McBride, 1997; Wainer, 1990). Unfortunately, CAT also has many limitations, the most important of which is the loss of control in the assembly stage of test development by content specialists. As a result, Luecht and Nungester (1998) state that test quality is limited when CAT is used since this real-time automated procedure eliminates the item assembly and test review process formerly completed by content specialists. They also note that CAT requires a great deal of trust in statistical criteria and outcomes, often at the expense of content specifications and test cohesion (i.e., some of the art in test development is lost). To overcome this limitation, Luecht and Nungester created a test assembly procedure called computer-adaptive sequential testing (CAST) that draws from the strengths of CAT while still allowing for quality assurance across test forms. A sequential test, like an item-based computer adaptive test, could be used to reduce the number of items administered to an examinee without the loss of measurement precision thereby reducing testing time and cost. CAST also has the added benefit of quality control because it allows the test developer to create multiple test forms that can be reviewed by content specialists before test administration.

Computer-Adaptive Sequential Testing

With the CAST procedure, blocks of items called “modules” can be assembled and reviewed for extensive content and cognitive-level coverage while meeting multiple statistical targets. These modules then become part of a computer-adaptive process, in which modules with different difficulty levels can be administered to examinees depending on their ability level after the starter module is completed. All examinees are administered the same starter module and then move to the next module according to their current ability estimate . Figure 1 illustrates how the CAST modules are linked together by seven possible routes or “panels” that examinees may

take depending on their current ability estimate. For example, if an examinee does poorly on module A, then she/he will be administered module B in stage 2. If the examinee then does well on module B, then she/he will be administered module C in stage 3. The modules from left to right cover the same content and cognitive level constraints and only differ in difficulty level. A computer program called CASTISEL was developed by Luecht (1998a) and is an automated test assembly program that assumes that data from the three-parameter logistic item response theory model are available for the items in the bank. This program was designed to build computerized-adaptive sequential test forms that could meet up to 32 item response theory (IRT) test information targets as well as 2000 constraints.

Insert Figure 1 about here

Parallel Forms Construction

In addition to sequential testing, the CAST model can be used to create parallel test forms. The number of parallel test forms that can be created is only limited by the item bank. Within this CAST framework, a test developer can generate parallel forms by constraining the first stage of the CAST assembly process (see bold module at Stage 1 in Figure 1) and by using the same statistical target for all forms while controlling for many variables including content and cognitive-level coverage, test length, and item exposure. It was only recently that computerized test assembly algorithms could be used to simultaneously match both statistical (i.e., test information functions) and content related specifications across test forms. Samejima (1977) described parallel forms using identical test information functions as “weakly parallel forms”. Adema (1992) also addressed the issue of weakly parallel forms and noted that the methods used at that time could not meet all of the necessary complexities and constraints of the actual test construction process. In regards to these earlier optimization models, Wightman (1998) states that the linear programming models could not simultaneously build parallel test forms and therefore, as new forms were created, each subsequent form was less than parallel. Because of the advances in computer algorithms and technology, optimal test assembly procedures have seen a resurgence

in the literature (see special issue of *Applied Psychological Measurement*, 1998). However, to date, no one has used the CAST model for parallel forms construction.

van der Linden and Reese (1998) argued that with the recent developments in computer technology and the acceptance of item response theory for item bank calibration, computerized adaptive testing has become feasible. However, it is our contention that many testing organizations may not be ready or willing to move into a CAT framework because it does not allow for test form quality assurance and it relies heavily on a sorting algorithm. In addition, some testing organizations may not be ready to move into a computer-based administration, like CAST, in place of their current and familiar paper-and-pencil administrations. In these situations, testing organizations may want to use item banks and statistical targets but may not want to rely solely on a computer-adaptive sorting algorithm and/or computer-based administration. Thus, the need for automated assistance comes about as item banks grow in size making the test development process ever more labor intensive. One alternative, described in this paper, is to adopt some of the CAST properties to build parallel test forms with specific target test information functions and content constraints using Luecht and Nungester's (1998) procedure.

Purpose

This study was designed to investigate the efficiency and effectiveness of computer-assisted parallel forms construction using a CAST procedure for two different criterion-referenced testing situations. A comparison of the parallel forms produced in each situation should provide an idea of how well the testing objectives were met under constrained but realistic conditions. The constraints that were studied include content area, cognitive level, test length, item exposure, statistical target, and number of parallel forms. This paper will also address issues related to the optimization of item banks when the goal of test construction is to create parallel test forms. We begin with a description of the algorithm used to operationalize the CAST procedure for parallel forms construction, the normalized weighted absolute deviation heuristic.

The Normalized Weighted Absolute Deviation Heuristic

Luecht (1998b) introduced a variation of a greedy algorithm that could meet very complex test assembly problems found in large-scale testing situations. van der Linden & Adema (1998)

described the negative aspect of the greedy algorithm—namely, that the algorithm loads all of the best items onto the first test and, thus, each successive test will never be quite parallel. However, this characteristic of the algorithm has been altered by Luecht (1998b) to prevent the first-form item build up from occurring by including a randomization factor into the item selection process. As a result, each form has an equal opportunity to choose quality items. Luecht also notes that after a few items have been selected, the objective functions in the revised algorithm will usually diverge enough that the two forms will be searching for different items.

The algorithm is called a normalized weighted absolute deviation heuristic (NWADH) and it uses a series of locally optimal searches to build one or more parallel test forms from an item bank. The normalization procedure will allow for the selection of numerous objective functions to be met simultaneously. The absolute deviation is the absolute difference between the target test information function and the current function. The weighted aspect gives weight (or priority) to items within content areas that do not meet the minimum constraints. This characteristic forces less discriminating items within certain content areas to be chosen first and thus allows for items within content areas that exceed the minimum to make up the difference for those items that do not exceed the minimum. Heuristic describes the problem-solving technique used to choose the most appropriate solution at each stage in the test assembly process.

Quantitative Constraints

When a test is being assembled from an item bank, ideally, test developers want to maximize information but, in reality, they must control for test length. Thus, instead of creating a test composed of all items (and thus having total information that is the sum of all item information), one must now create a test that satisfies two criteria: a) it must be of a certain length and b) it must satisfy a target of test information. The following summary will only include the two constraints listed above in order to describe the algorithm for the reader.

The NWADH chooses items by dividing the total information of the items in the bank, T , by the number of items needed in the test, n . Thus, the first item the algorithm selects from the bank should have T/n information. In other words, T/n minus the information of the item under consideration should be close to zero. This difference, divided by the total amount of deviations of

all the items left in the item bank from our item-target, should also be close to zero. By dividing the difference between T/n minus the information of the item by the total amount of deviations for all of the remaining items in the bank, a value is obtained that describes the current item under selection in terms of the other possible items that could be chosen. This outcome will indicate whether or not our item is as good as the average of the remaining items in the bank and allows one to be confident that the best possible item will be chosen.

For each succeeding item, the algorithm, instead of looking at the target test information function, looks at the target test information function minus the total information of the items that have already been selected. With this difference, the algorithm then divides by the number of items that still need to be found (i.e., the total test length minus the number of items that have been already chosen). This value gives the target for what the next item should have. The same procedure is then followed for every other item needed in the test.

In using this approach, one may notice some odd characteristics: a) we are getting very small numbers and b) the sum of these numbers is meaningless. In order to put this sum on a meaningful metric, one solution is to index the value by test length. To do this, the algorithm takes our proportion of absolute item difference from the target to the sum of all absolute item differences-from-the-target and subtracts it from one. Therefore, instead of a meaningless range, we can now say that we are aiming for a value of 1. For example, a perfect item would equal the target item information function, giving a proportion of 0 divided by the total items left in the item bank. Obviously, 1 minus 0 is 1, a perfect score. Thus, the algorithm compares each item choice with all other possible choices while trying to use the best combination of items that meet the interim target information function. Next, we review some of the actual formulas in the algorithm.

The NWADH turns a minimization problem (i.e., the distance between the information function of an item and a target test information function) into a maximization problem. That is,

$$e_i = 1 - \frac{d_i}{\sum_{i \in R_{j-1}} d_i}, i \in R_{j-1},$$

where e_i is an arbitrary statistic denoting the goodness-of-fit for an item i and d_i is the index that defines how well the item information fits with our current item search value. The d_i value is then divided by the sum of all other possible item fits and is an assessment of how the fit of the item we are currently looking at compares to all other possible items. R_{j-1} is the remaining items in the bank that are indexed after the selected items have been removed. Further, d_i is given by

$$d_i = \left| \frac{T - \sum_{k=1}^I u_k x_k}{n - j + 1} - u_i \right| ; i \in R_{j-1},$$

where u_i is the information of the item that we are currently inspecting, T denotes the target test information function that is to be maximized, the sum of u_k is the total information that is already in the test, x_k is an index of whether or not we have already chosen item k (and will be coded as either 0 or 1), n is the length of the test, and j denotes the actual number of the current item under inspection compared to the total needed. For example, if we need 10 items and have already chosen 5, $j = 6$. As the d_i values decrease, the e_i values increase and this outcome represents a closer fit to the interim target for selecting item $j = 1, \dots, n$. From the previous formula, the expression

$$\frac{T - \sum_{k=1}^I u_k x_k}{n - j + 1},$$

represents the target value for the selection of the next item that will minimize the difference between the current function and the target test information function (Luecht, 1998b).

Content Constraints

The content constraints are addressed in a different way compared to the quantitative constraints. A weighting scheme is created that prioritizes the item selection within content area. This weighting scheme was devised by Luecht (1998b) to ensure optimization by selecting particular items first if they did not meet the minimum constraints assigned and thereby speeding

up the NWADH process. The weights are adjusted after each iteration of the NWADH algorithm with the items not meeting the minimum constraints receiving more weight than the items that meet the minimum or exceed the maximum. Luecht calls this prioritized outcome the “need-to-availability ratio”. This ratio is computed so that every test form will have the same number of items that meet both quantitative and content constraints with priority going to the categories that have the greatest need-to-availability ratio. Content areas with few items and with items that do not meet the minimum constraints are chosen first and therefore the content areas with a low need-to-availability ratio will be forced to make up the difference later in the iterative process.

Item and Test Information Functions

The use of item response theory (Hambleton, Swaminathan & Rogers, 1991; Lord, 1980) in automated test assembly allows for a method of item and test selection and comparison based on item and test information functions. One of the constraints in this study is that our target test information function (TTIF) is fixed across all forms. A TTIF represents the amount of measurement precision the test developer wants to achieve across the entire theta score scale. Statistically defined, the item information function is inversely proportional to the square of the width of the asymptotic confidence interval for \boldsymbol{q} . This relationship implies that the larger the information function, the smaller the confidence interval and the more accurate the measurement. For the three-parameter logistic IRT model, the item information function is calculated as (Lord, 1980, p. 73):

$$I_i(\boldsymbol{q}) = \frac{D^2 a_i^2 (1 - c_i)}{(c_i + e^{Da_i(\boldsymbol{q}-b_i)})(1 + e^{-Da_i(\boldsymbol{q}-b_i)})^2},$$

where $D=1.7$, a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the pseudo-chance level. For any given ability level, the amount of information increases with larger values of a_i and decreases with larger values of c_i . That is, item discrimination reflects the amount of information an item provides assuming the pseudo-chance level is relatively small. The test information function (i.e., $I[\boldsymbol{q}]$) is an extension of the item information function. The test information function is the sum of the item information functions at a given ability level:

$$I(\mathbf{q}) = \sum_{i=1}^n I_i(\mathbf{q}),$$

where $I_i(\mathbf{q})$ is the item information and n is the number of test items. It defines the relationship between ability and the information provided by a test. The more information each item contributes, the higher the test information function.

Method

The computerized-adaptive sequential test (CAST) procedure, which used the normalized weighted absolute deviation heuristic, was used to build parallel forms in two criterion-referenced testing situations. All analyses were conducted with the computer program CASTISEL (Luecht, 1998a). The first situation was an achievement test where we had access to a bank with 159 items from previously administered Grade 9 mathematics achievement tests from a large-scale testing program in Canada. The second situation was a medical licensure exam where we had access to a bank with 1973 items from the Medical Council of Canada.

Case #1: Criterion-Referenced Achievement Test

Data for the Grade 9 Mathematics Achievement Test came from an achievement-testing program in the Canadian province of Alberta. The blueprint for this test consists of four content areas (number, patterns and relations, shape and space, and statistics and probability) and two cognitive levels (knowledge and skills). The item bank for this test was created using five previously administered Grade 9 Mathematics Achievement Tests from 1995 to 1999. Each test contained 50 multiple-choice items with an item overlap of approximately 50% from one year to the next. To create an item bank for this test, the IRT parameters were estimated using BILOG 3.11 (Mislevy & Bock, 1997) with a random sample of 6000 students. IRT characteristic curve equating was conducted (Stocking & Lord, 1983) using a common-items nonequivalent groups design. All items were equated to the 1997 score scale. When equating was complete, the item bank had 159 items. The mean difficulty was -0.097 ($SD = 0.872$); the mean item discrimination was 0.843 ($SD = 0.293$); and the mean lower asymptote was 0.179 (SD of 0.103).

The criterion-referenced achievement test in this study is like many of the achievement tests used throughout North America. It is characterized by an average and a high cut score—that is,

the need to differentiate examinees at two points on the theta score scale. Typically, an acceptable standard of performance is identified as well as a standard of excellence. Theta scores of 0.0 and 1.0 were used for the acceptable standard and standard of excellence, respectively. Two, 50-item parallel achievement test forms were assembled from the item bank because each single form of the actual Canadian achievement test contains 50 items (thus, approximately 63% of the items from the bank were used to create two parallel forms). The targets were chosen to maximize information at the acceptable standard and the standard of excellence.

In total, the achievement tests had 24 constraints. Two parallel forms were created (2 constraints). Each form had to satisfy a test blueprint with equal numbers of items in four content areas across two cognitive levels (8 blueprint cells X 2 forms = 16 constraints). Test length was fixed to 50 items per form (2 constraints). Each form was created with unique items (2 constraints). Each form had its own statistical target which is the same target across tests when creating parallel forms (2 constraints).

Case #2: Licensure Exam

Data for the licensure exam was obtained from the Medical Council of Canada (MCC). The goal of the qualifying examinations administered by MCC is to evaluate the competencies required for licensure in Canada. These competencies are deemed to be essential for the practice of medicine by all physicians in Canada. The blueprint for this test consists of six disciplines (preventative medicine and community health, internal medicine, surgery, obstetrics and gynecology, pediatrics, and psychiatry). The item bank for this exam was obtained from the MCC. It contained 1973 items across the six disciplines. The item bank consists of dichotomously-scored items, calibrated with the 2-parameter logistic IRT model, and scaled onto the 1998 examination score scale. IRT characteristic curve equating was used with a common-items nonequivalent groups design to create the item bank. The mean difficulty was -1.881 ($SD = 2.984$) and the mean item discrimination was 0.255 ($SD = 0.129$).

The criterion-referenced licensure exam in this study is characterized by one cut-point at the lower end of the theta score scale. The cut-point was set at a theta value of -1.3 which is the

approximate passing level on this exam for recently graduated medical doctors who seek entry into a supervised practice. Seven, 168-item parallel licensure exams were assembled from the bank because the current version of the MCC Qualifying Examination, which is a fixed length computer adaptive test, contains 168 items (thus, approximately 60% of the items from the bank are used to create the parallel forms—a similar percentage to the achievement study described in the previous section). The targets were chosen to maximize information at the cut-point of -1.3 .

In total, the licensure exams had 70 constraints. Seven parallel forms were created (7 constraints). Each form had to satisfy a test blueprint with equal number of items in six discipline or content areas (6 disciplines X 7 forms = 42 constraints). Test length was fixed to 168 items per form (7 constraints). Each form was created with unique items (7 constraints). Each form had its own statistical target which is the same target across exams when creating parallel forms (7 constraints).

Results

In the following section, the means and standard deviations for the achievement tests and licensure exams are presented. The mean indicates the overall score for each form while the standard deviation is a measure of item dispersion within each form. As well, mean TIF difference is presented indicating the relative fit between the observed and target TIF. The mean fit is the average of the deviations and it can be calculated with the expression, $T(\Theta_k) - \sum_l |(\Theta_k; \xi_l)|$, where $l = 1, \dots, 50$ items computed across $k = 1, \dots, 13$ quadrature points for the achievement tests (168 items and 17 quadrature points for the licensure exam) across a fixed range of Θ . The mean square error of the TIF difference is an average of the squared deviations between the observed and target TIF (Luecht & Nungester, 1998). For this study, a MSE of .05 or less between the observed and target TIF was considered a good fit.

Criterion-Referenced Achievement Test

Table 1 contains the means, standard deviations, mean TIF differences, and MSE of the TIF differences for the two parallel form in the achievement study. Figure 2 shows the test information functions for the two CAST forms along with the five traditional forms created for the 1995 to 1999 administrations (by traditional we mean that the tests were created one at a time without the use

of an automated test assembly algorithm). Note that each of the five traditional tests have, approximately, 50% overlap from year to year and 25% overlap with every second year. The two CAST forms had no overlapping items. The means and standard deviations between the two CAST forms were comparable, although the standard deviation for the second form was larger indicating that the items on this form had a wider range of difficulties. The mean TIF differences were small indicating that the observed and the target TIFs were comparable across forms. Also, the MSE of TIF difference was small ($MSE < 0.05$) indicating good fit to the target.

Table 3 contains the means, standard deviations, mean TIF differences, and MSE of the TIF differences for the five traditional tests. These outcomes provide a basis of comparison for the CAST forms because the traditional tests were designed to be parallel across years. The means and standard deviations across the five administrations varied. The means range from -0.246 to 0.065 while the standard deviations range from 0.773 to 0.927 . The mean TIF differences also had a range, from -0.108 to 0.085 , indicating that the observed and the target TIFs were comparable for some forms but not others. The MSE of TIF difference was small ($MSE < 0.05$), indicating good fit to the target, for the 1998 administration but large from the remaining four administrations indicating poor fit.

When the CAST and traditional forms are compared, the two CAST forms had more information than three of the traditional tests (1995, 1998, and 1999) at the acceptable standard. The CAST forms also had less information than two of the traditional tests (1996 and 1997) at the acceptable standard. At the standard of excellence, the two CAST forms had more information than two of the traditional tests (1998 and 1999), equal information with one of the traditional tests (1995), and less information than two of the traditional tests (1996 and 1997).

Insert Tables 1 and 2 and Figure 2 about here

Licensure Exam

Table 3 contains the means, standard deviations, mean TIF differences, and MSE of the TIF differences for the seven parallel form in the licensure study. Figure 3 shows the test information functions for the seven parallel forms. The means and standard deviations across the seven CAST forms varied. The means range from -2.087 to -2.506 while the standard deviations range from 0.832 to 2.528 . The mean TIF differences also had a range, from 0.017 to 0.348 , indicating that the observed and the target TIFs were comparable for some forms but not others. The MSE of TIF difference was small ($MSE < 0.05$), indicating good fit to the target, for the first three forms but large from the remaining four forms indicating poor fit.

 Insert Table 3 and Figure 3 about here

Parallelism of Achievement Tests and Licensure Exams

The empirical and visual results strongly indicate that the CAST forms were parallel for the achievement tests when each form had a set content area and cognitive-level coverage, test length, item exposure limit, statistical target, and number of parallel forms. The test information functions were maximized at the acceptable standard (i.e., theta of 0) while the standard of excellence (i.e., theta of 1) had less information than the acceptable standard. The bank for the achievement study contained previously administered items. It had many items at the acceptable standard but fewer items at the standard of excellence thereby making it difficult to maximize information at both cut scores. To overcome this limitation, more discriminating items at the standard of excellence are needed in the bank. The combination of low discriminating and average difficulty items resulted in a parallel forms solution that missed the second target.

The empirical and visual results indicate the CAST forms were comparable for some of the licensure exams but not for others. Forms 1 to 3 in Table 3 indicate good fit between the observed and target test information functions although all three tests were off the target of -1.3 . In order to create parallel forms, it was necessary to shift the maximum of the target information function from -1.3 to -2.4 . This change ensured that the exams were parallel (i.e., had

overlapping test information functions) but off the cut score target of -1.3 . The remaining four forms provided inadequate fit between the observed and target test information functions with each subsequent form becoming worse relative to the target (see MSE of TIF difference in Table 3). Clearly, the licensure bank is deficit for parallel forms construction in two ways. First, item difficulty was below the cut score of -1.3 for many items making it hard to select items that functioned well at this point on the score scale. Second, discrimination power was low for many items in the bank. As a result, items with high discriminating power around the cut score of -1.3 were quickly depleted during first stages of the assembly process, and therefore, the balancing of information across all forms became problematic when more than three forms were created. While the items in the achievement bank were limited at the second cut score (i.e., the standard of excellence), the items in the licensure bank had a similar problem except at a theta cut-point of -1.3 . As a result, parallel forms from the licensure item bank diverged after three tests were created¹.

Conclusions and Discussion

Research in the area of optimal test design over the last few years has tended to focus on sorting algorithms associated with computerized test assembly problems. In fact, the simultaneous assembly of parallel forms with set content area and cognitive-level coverage, test length, item exposure limit, statistical target, and number of parallel forms has only recently been advanced. We used a constrained version of the computerized-adapted sequential testing (CAST) procedure, implemented with the computer program CASTISEL, to create criterion-referenced parallel test forms. To date, no one has used the CAST procedure for parallel forms construction. As testing organizations develop their item banks, an ever-increasing demand is applied to the already resource intensive manual assembly process. It is here where computer-assisted test assembly optimization heuristics become most applicable. If an exam similar in structure to the achievement test is required, for example, then items with maximum information can be gathered around the decision or criterion-referenced cut scores of interest using the CAST procedure to create parallel forms with many constraints. The automation of the test assembly

process using optimization heuristics is not meant to replace the test developer but only to assist the developer by solving very complex test assembly problems.

Infeasibility Problems

The design and choice of the targets is especially important in order to obtain maximum information at various cut points along the ability continuum thereby decreasing decision error at those points. In other words, the design aspect of the CAST procedure is important because it will allow test developers to tailor each test according to specific criteria. The first constraint imposed in this study was the actual size of the item banks. The exact specifications on how large or how good (i.e., item difficulty and discrimination) an item bank must be is not provided because it depends on the number of constraints that must be met in a particular testing situation. As the constraints accumulate within a particular testing program, the chance of having “infeasibility problems” becomes greater (Timminga & Adema, 1996). Timminga (1998) recently noted: “For test assembly problems, the feasible region or solution space consists of all tests that meet the model constraints in the particular model. The objective function determines the ‘best’ test in this region. If the region is empty, then the model is infeasible” (p. 280). When creating a bank, test developers must also consider the shape of the target information function for their testing program. Creating banks with items close to the target information function should reduce the chances of infeasibility. Also, it is important to note, that after the tests have been assembled using a CAST procedure, test developers can manually add or remove items to the tests. These suggestions should limit infeasibility problems that may occur when too many constraints are applied using automated test assembly procedures. It should also help developers create parallel criterion-referenced forms under highly constrained conditions with the CAST procedure.

Sequential Nature of the NWADH

As the NWADH was used to meet the increasingly stringent objectives for the achievement tests and licensure exams in this study, there was no sacrificing the quality of one form in order to build another form. Therefore, the sequential nature of the NWADH was quite beneficial because it allows the test developer to quickly determine how many parallel forms can be assembled according to the target test information function specified. If the solution is inadequate, the test

developer could pretest more items to build up the bank or to reduce the target test information function. Of course, a reduction in the target information function may result in more parallel forms but it will also reduce measurement precision. In order to create high quality parallel criterion-referenced forms with the CAST procedure, a strong item bank is required. If the item bank is depleted or devoid of high quality items then no matter which optimal test design is employed the tests will be of limited quality. Therefore, careful planning must take place when developing an item bank.

Identifying and Ensuring Parallelism Across Test Forms

Traditionally, psychometricians have used statistical definitions and criteria to operationalize parallelism. Weakly parallel forms exist, for example, when the test information functions are comparable across forms (Samejima, 1977). However to create truly parallel forms, both content and statistical targets must be met (van der Linden & Adema, 1998). We believe there are two components that must be satisfied when creating parallel test forms. The first component includes statistical evidence. For example, empirical indices such as the mean TIF difference and the mean square error of the TIF difference can be computed. The mean TIF difference is a measure of fit between the observed and target test information functions. When we take the absolute value of this difference, the MSE of the TIF difference is produced—a MSE less than or equal to 0.05 indicates good fit. To supplement empirical indices, graphical methods can be used. For example, one can examine the observed and target test information functions to see if and where the functions overlap. The second component includes substantive or judgmental evidence. For example, a substantive review by content specialists can be conducted. Content specialists could review the items to ensure test cohesion within forms, high quality across forms, and adequate content coverage. Currently, parallelism is assessed using statistical evidence but, seldom, using substantive evidence.

Two real item banks were used in this study to evaluate the CAST procedure for parallel forms construction. Parallelism was assessed using statistical evidence (i.e., empirical indices and graphical methods) but not substantive or judgmental evidence. Therefore, the authors are currently working on phase two, which involves the review process by content specialists. The

parallel forms generated for both the achievement tests and licensure exams will be reviewed by content specialist to see if the forms have integrity. The authors are also developing other statistical indices to quantify of the area between the observed and target test information functions. If the results from all three types of review (i.e., outcomes from empirical indices, graphical methods, and substantive reviews) converge, then researchers and practitioners will be assured that the CAST procedure can, in fact, produce truly parallel forms using a multi-faceted conception of parallelism. If the results do not converge, then research must be undertaken to better understand the complexities of test development and to integrate or merge these qualities into the CAST procedure outlined in this paper.

References

- Adema, J. J. (1992). Methods and models for the construction of weakly parallel tests. Applied Psychological Measurement, 16, 53-63.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.
- Luecht, R. M. (1998a). CASTISEL 3PL [Computer Program and manual].
- Luecht, R. M. (1998b). Computer-assisted test assembly using optimization heuristics. Applied Psychological Measurement, 22, 224-236
- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. Journal of Educational Measurement, 35, 229-249.
- Mislevy, R. J., & Bock, R. D. (1997). BILOG 3.11: Item analysis and test scoring with binary logistic test models [Computer Program]. Morrseville, IN: Scientific Software
- Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. Psychometrika, 42, 193-198.
- Sands, W. A., & Waters, B. (1997). Introduction to ASVAB and CAT. In W. A. Sands, B. K. Waters, & J. R. McBride. (Eds.), Computerized adaptive testing: From inquiry to operation. Washington DC: American Psychological Association.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.
- Timminga, E. (1998). Solving infeasibility problems in computerized test assembly. Applied Psychological Measurement, 22, 280-291.
- Timminga, E., & Adema, J. J. (1996). An interactive approach to modifying infeasibility 0-1 linear programming models for test construction. In G. Engelhard Jr. & M. Wilson (Eds.), Objective measurement: Theory into practice (Vol. 3). Norwood NJ: Abex.
- van der Linden, W. J. (Ed.) (1998). Optimal test assembly [Special Issue]. Applied Psychological Measurement, 22 (3).

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. Journal of Educational Measurement, 35, 185-198.

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

Wainer, H. (1990). Introduction and History. In H. Wainer (Ed.), Computer Adaptive Testing: A Primer. (pp. 1-21). New Jersey: Lawrence Erlbaum.

Wightman, L. F. (1998). Practical issues in computerized test assembly. Applied Psychological Measurement, 22, 292-302.

Notes

We thank Professor Tom Maguire for his helpful comments on an earlier version of this paper.

Support for this research was provided to the first author from the J. Gordon Kaplan Award, Faculty of Graduate Studies, University of Alberta.

¹ These characteristics of the licensure item bank—negative b-parameters and relatively low a-parameters—are not unusual for this type of testing. The majority of items are designed to tap general (rather than specific) knowledge and skills in the six discipline areas measured by the test. Moreover, the examinees consist of medical doctors who have completed and passed their studies. They now seek entry into supervised medical practice. Consequently, the examinees, as a whole, are homogenous which decreases the value of the a-parameters. Over 90% of the examinees pass the Qualifying Examination, as intended by the Medical Council of Canada.

Table 1

Parallel Form Results for the CAST Achievement Tests Using a Bank With 159 Items

| Parallel Form | No. of Items | Mean Difficulty | SD Difficulty | Mean TIF Difference | MSE of TIF Difference |
|---------------|--------------|-----------------|---------------|---------------------|-----------------------|
| 1 | 50 | -0.108 | 0.479 | 0.033 | 0.010 |
| 2 | 50 | -0.165 | 0.690 | -0.040 | 0.004 |

Table 2

Parallel Form Results for the Traditional Achievement Tests

| Form | No. of Items | Mean Difficulty | SD Difficulty | Mean TIF Difference | MSE of TIF Difference |
|------|--------------|-----------------|---------------|---------------------|-----------------------|
| 1995 | 50 | 0.065 | 0.844 | 0.018 | 0.062 |
| 1996 | 50 | -0.083 | 0.872 | -0.108 | 0.158 |
| 1997 | 50 | 0.006 | 0.773 | -0.156 | 0.407 |
| 1998 | 50 | -0.249 | 0.927 | 0.029 | 0.042 |
| 1999 | 50 | -0.246 | 0.922 | 0.085 | 0.262 |

Table 3

Parallel Form Results for the CAST Licensure Exams Using a Bank With 1973 Items

| Parallel Form | No. of Items | Mean Difficulty | SD Difficulty | Mean TIF Difference | MSE of TIF Difference |
|---------------|--------------|-----------------|---------------|---------------------|-----------------------|
| 1 | 168 | -2.475 | 0.832 | 0.017 | 0.008 |
| 2 | 168 | -2.506 | 1.100 | 0.010 | 0.004 |
| 3 | 168 | -2.460 | 1.325 | 0.007 | 0.007 |
| 4 | 168 | -2.396 | 1.648 | 0.034 | 0.066 |
| 5 | 168 | -2.424 | 1.903 | 0.203 | 1.327 |
| 6 | 168 | -2.087 | 2.028 | 0.297 | 2.685 |
| 7 | 168 | -2.367 | 2.528 | 0.348 | 3.425 |

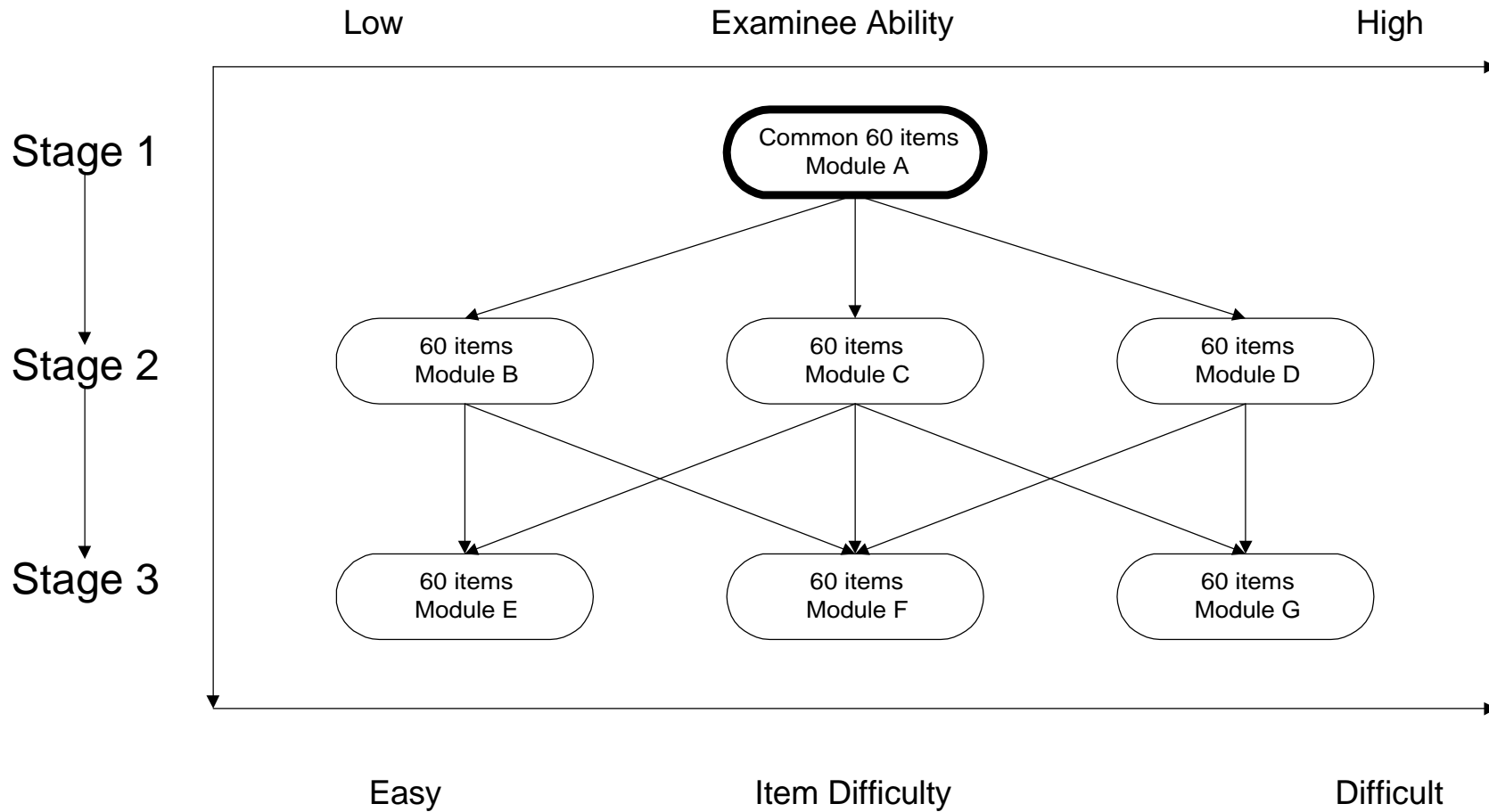
Figure Caption

Figure 1. A three-stage CAST model with seven modules and seven panels.

Figure 2. The test information functions for two parallel forms created with the CAST procedure along with the five traditional achievement tests.

Figure 3. The test information functions for seven parallel licensure exams.

Computer-Adaptive Sequential Testing 3 Stages



Test information Functions

