Some New Developments in Adaptive Testing Technology

Wim J. van der Linden

University of Twente, The Netherlands

Abstract. In an ironic twist of history, modern psychological testing has returned to an adaptive format quite common when testing was not yet standardized. Important stimuli to the renewed interest in adaptive testing have been the development of item-response theory in psychometrics, which models the responses on test items using separate parameters for the items and test takers, and the use of computers in test administration, which enables us to estimate the parameter for a test taker and select the items in real time. This article reviews a selection from the latest developments in the technology of adaptive testing, such as constrained adaptive item selection, adaptive testing using rule-based item generation, multidimensional adaptive testing, adaptive use of test batteries, and the use of response times in adaptive testing.

Keywords: adaptive testing, item-response theory (IRT), response-time modeling, rule-based item generation, test specifications, test speededness

The idea of adapting a test to the ability level of the individual test taker is as old as testing itself. Prototypes of this adaptation are the oral examination in education and the individual diagnostic interview in psychology. Ever since the first oral examinations have been conducted, examiners are aware of the fact that it would be a waste of time to ask examinees questions that are too difficult or too easy. When an examinee produces plainly wrong answers to a series of difficult questions, each examiner will resort to easier questions until their level of difficulty leads to uncertainty as to whether the next answer will be right or wrong. The reverse strategy will be followed if the examinee produces a series of perfect answers. Any sensitive psychologist involved in individual diagnosis or informal testing will follow comparable principles.

These examiners and psychologists must have had ideas about the difficulties of their test items and how to account for them when scoring or classifying test takers. Otherwise, they would simply have been unable to do their jobs. In hindsight, we could even argue that these ideas must have functioned as intuitive versions of the later item-response models, which organized their quantitative impressions of the items and test takers and guided them in the selection of the former as well as the scoring of the latter.

The idea of adapting a test to the level of the test taker was ingrained so deeply in the practice of examination and psychological diagnosis that it was automatically adaptive when more formal psychological testing was introduced. The prime example is the first intelligence test constructed by Binet in the beginning of the 20th century. In spite of its thorough standardization, this test was fully adaptive; its protocol contained precise descriptions as to how to select the next item for a test taker as a function of his or her previous responses (Binet & Simon, 1905). This pioneer must thus have been convinced that for a test to be standardized it was unnecessary to give each test taker the same selection of items – only that they be subjected to identical *rules* of item selection.

Although the Binet intelligence test has been generally hailed as the first standardized test in the history of psychological testing, this author believes that the major innovation by Binet was not the standardization – the methodological necessity of it was already fully accepted in German experimental psychology in the 19th century – but the replacement of the *intuitive response models* of his predecessors by explicit scaling of the test items and test takers. As is well known, Binet used chronological age as a scale for intelligence, extensively pretested his items to estimate their position on this scale, and scored his test takers by estimating their position on the same scale, which he referred to as their mental age. It was no coincidence that Thurstone, in the very first article on the statistical aspects of scaling in 1925, used a data set for the Binet intelligence test to demonstrate his new scaling model.

But by the time Thurstone began his work on scaling, educational and psychological testing had already been greatly influenced by the invention of paper-and-pencil, group-based testing as the result of the necessity to test large numbers of conscripts in the United States during mobilization in World War I (DuBois, 1970). The same format of group-based testing, with an identical linear test form for each candidate and the use of observed-score equating to maintain comparability of scores over time, soon became popular in testing for college admission as well. Although efficient to administer for the testing agency, this format does not permit any adaptation. As a matter of fact, it even led to loss of efficiency for the test takers because of the waste of their time by including items in the test that were too easy or too difficult for them.

DOI 10.1027/0044-3409.216.1.3

We had to wait for two new developments before adaptive testing returned. The first was the introduction of what is now known as item-response theory (IRT). There is a direct line of descent between IRT and Thurstone's work but he eventually became more interested in the scaling of objects than the measurement of persons. A key feature of IRT modeling is its explanation of the probability distribution of the responses on a test item by separate parameters for the ability of the test taker and the relevant features of the item. One of the early examples of an IRT model is that by Rasch (1960), which has exactly one ability parameter and one parameter for the difficulty of the item. The model assumes that the probability of a correct response U_i by a test taker on an item can be written as

$$\Pr\{U_i = 1\} = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}, \qquad (1)$$

where θ is the ability parameter for the test taker and b_i the difficulty parameter of item i. When the items in a test have been calibrated using pretest data (i.e., their difficulty parameters have been estimated with enough precision to treat them as known), the estimates of the ability parameter from the test takers' responses are automatically adjusted for the difficulties of the items used in the test. This feature permits the use of any selection of items from a calibrated pool as a test without loosing the comparability of their scores and, therefore, makes adaptive selection of test items possible.

In adaptive item selection, the test begins with an initial estimate of the ability parameter, $\hat{\theta}_0$. The next item is then selected to be optimal at $\hat{\theta}_0$ and the response to it is used to re-estimate the ability, that is, calculate an estimate $\hat{\theta}_1$. The procedure is then repeated again, and the result is a new estimate $\hat{\theta}_2$, and so forth. For the model in (1), an obvious procedure is to select each next item so that it has a difficulty parameter b_i as closely as possible to the current estimate of θ. For the more flexible models typically used in educational testing, this criterion is not appropriate. One popular criterion is the use of Fisher's information measure in statistics as an item information function, $I(\theta)$. We demonstrate its use only graphically. Figure 1 shows the selection of the first three items in an adaptive test (top to bottom). Each next item is selected to have its peak as closely as possible to the last estimate of θ. Item information functions have the advantage of being additive; that is, the test information function is just the sum of the information functions of its items. As shown in Figure 1, even after as few as three items, the test information function already reveals a tendency to become peaked over a small area of the ability scale. For the mainstream IRT models used in testing, it can be proven that the location of the peak converges to the test taker's true ability.

The second development that led to the reintroduction of adaptive testing was the availability of computers with ample computational power. When these became affordable in the second half of the 1980s, they were immediately used for test administration. Their power has enabled the testing industry to estimate the θ parameters of test takers

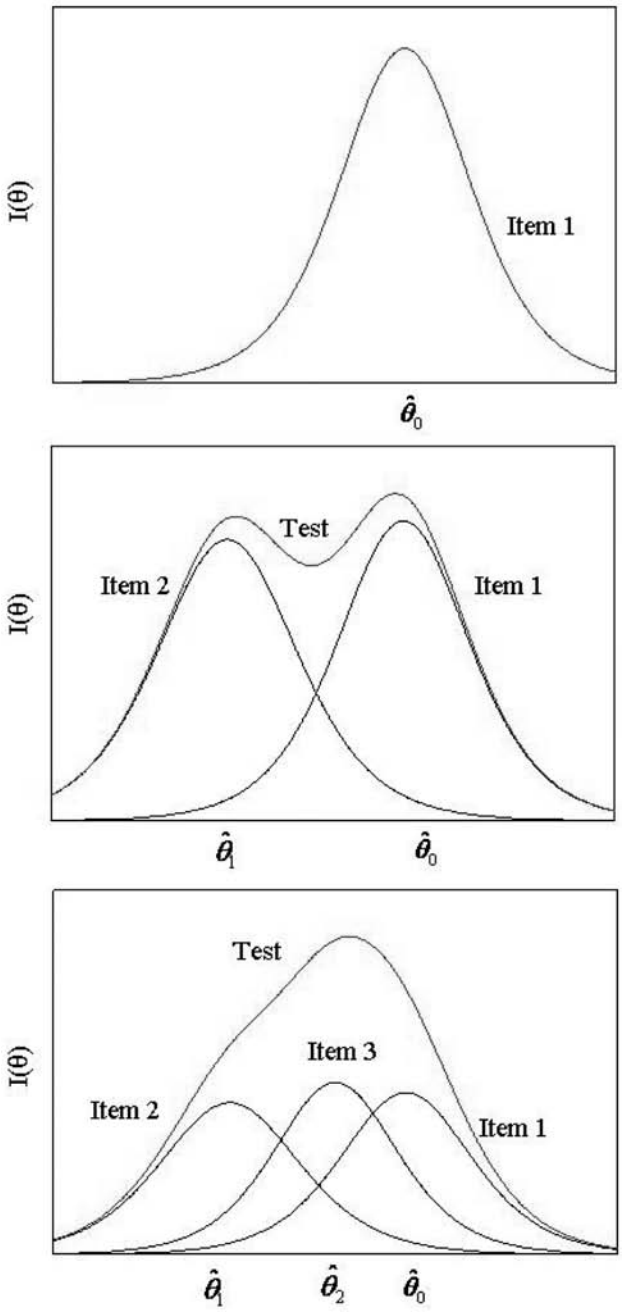

Figure 1. Graphical example of the selection of the first three items in an adaptive test using the information functions of the items in the pool.

in statistically sophisticated ways and to select items optimally at the estimates from large pools of items in real time.

This does not mean that there were no earlier attempts at adaptive testing. David Weiss did his pioneering work on adaptive testing at the University of Minnesota in the 1970s on a mainframe computer. Others tried to simplify the necessary estimation and item-selection procedures. For example, Lord (1970) studied a simple up-and-down method of item selection that had his roots in the Robbins-

Monro technique for stochastic approximation developed for other applications in statistics. He also developed paper versions of adaptive tests in the form of multi-stage tests and an ingenious flexilevel test for which the test takers had to scratch their answer sheets to find the next item (Lord, 1980, chaps. 8–9).

The first large-scale applications of adaptive testing were launched in the mid-1990s. Surprisingly, none of the applications were in psychological testing but in such areas as university admission, job certification, and placement in the military. However, because of the large scale of these applications and the high security of their nature, test specialists soon became aware of the fact that their original ideas about adaptive testing had been rather naive and its implementation involved much more than just picking the best items from a pool one at a time. In fact, these early experiences opened up the eyes to a whole array of challenging new problems.

For example, just like standardized linear tests, adaptive tests have to meet extensive sets of specifications to maintain their validity. These are not only content specifications but also rules with respect to the degree of speededness of the test, the exclusion of items that are too close to each other, the frequency of item reusage, and so forth. Such specifications have to be imposed while selecting items sequentially in real time – a situation differing fundamentally from the mixing and matching that is possible in manual test assembly. As we will see below, the consequence is a complicated constrained sequential optimization problem with large numbers of constraints.

Also, adaptive testing is usually adopted in combination with testing schedules that allow candidates to show up when they are ready to take the test (*walk-in testing*). Such continuous testing service is valuable but places a heavy demand on the security of the testing material, especially because adaptive testing tends to capitalize on a small set of best items in the pool and to ignore the others. This feature has a serious price in the form of additional security measures that are necessary to protect this subset and to write and pretest massive numbers of items to replenish the item pool frequently.

The last example has to do with the timing of the test. Continuous testing is mostly offered in the form of time slots of fixed length for which candidates sign up. But test items usually vary considerably in the time they take; in educational testing, differences by a factor of five or more are quite common. Therefore, it did not take too long to discover that on adaptive tests the more able test takers sometimes run out of time. The reason for this differential speededness is the positive correlation between the difficulty of the test items and their time intensity. As a result, the algorithm tends to give the more time-intensive items to the more able test takers. On the other hand, in adaptive testing, the response times are automatically recorded. They can be used not only to fix this problem but even to increase the efficiency of the test beyond what is possible on the basis of the responses only.

The review of adaptive testing research in the next sections reports on these and other topics. The selection of these topics is strongly biased by the author's own research agenda and certainly not exhaustive. For instance, we do not touch on any of the topics in the other contributions to this special issue. Also, the review is not technical and, therefore, misses critical aspects of the developments. Readers with a more technical interest in adaptive testing are referred to a recent review by van der Linden and Glas (2007), which addresses several of the same topics but deals with their statistical aspects only.

Constrained Adaptive Testing

Even though the imposition of content specifications on the selection of items from a calibrated pool seems superfluous from a psychometric perspective, important validity reasons exist to do so on adaptive tests, especially when they test knowledge and skills that are the result of learning. Otherwise, if test takers found out that the item selection tends to favor certain content areas and ignores others, they might change their learning and, hence, invalidate the item calibrations. In addition to the content specifications, adaptive tests usually have to satisfy numerous other conditions, some also related to their validity but others just practical. Rather than using ad hoc modifications of the algorithm to deal with these specifications individually, we should be interested in a general approach that can be trusted for any type of test specification.

Key to the development of such an approach is the notion of a test specification as a constraint on the selection of the items by the algorithm. Its adoption implies that we can treat adaptive testing as an instance of constrained combinatorial optimization; that is, a problem in which we pick an optimal combination of items from the pool that has to meet a well-defined set of constraints. However, unlike regular constrained combinatorial optimization, the combination has to be found sequentially, each time updating the objective function that is to be optimized – in Figure 1: The information about θ in the test – once a new item is picked.

The sequential nature of the problem prevents us from any backtracking – that is, undoing the choice of an earlier item if it appears to lead to a solution that is suboptimal or even infeasible because some of the constraints have to be violated later in the test. Backtracking is typical of algorithms for the solution of regular combinatorial optimization problems (e.g., a branch-and-bound algorithm). When backtracking is impossible, the only approach left is to look ahead and project the consequences of each step. This alternative is followed in the shadow-test approach to adaptive testing (van der Linden, 2000; 2005, chap. 9; van der Linden & Reese, 1998).

The approach is summarized in the following pseudo-algorithm:

– Select a full test that meets all constraints and is optimal at the initial ability estimate $\hat{\theta}_0$;
– Pick the best item from this test for administration;
– Record the response and calculate the new ability estimate $\hat{\theta}_1$;
– Reassemble the full test from Step 1 to be optimal at $\hat{\theta}_1$ while still meeting all constraints and fixing the item that has already been administered;
– Repeat Step 2–4 until the adaptive test is completed.

The full tests that are (re)assembled are shadow tests; they are never seen by the test taker, but only serve as an intermediate step in the selection of the items for the adaptive test (see Figure 2). Because each shadow test meets all of the constraints, the adaptive test automatically meets them. Likewise, because each shadow test is optimal and its best item is always used, the adaptive test taken is optimal given the set of constraints.

An ideal way of implementing a shadow-test approach is through 0–1 integer programming. In this technique, the objective function and the constraints are modeled using 0–1 variables for the selection of the items, whereupon the model is solved for its optimum. Software programs with powerful solvers for 0–1 problems are available in most commercial optimization packages. A catalog of examples of how to formulate test specifications as constraints using 0–1 variables is given in van der Linden (2005). For a typical adaptive testing problem with hundreds of constraints and a well-implemented solver, it takes no more than a split second to calculate the next shadow test. Also, the best item is selected much faster than in unconstrained adaptive testing because it is picked no longer from the entire pool but from the free items in the shadow test.

The shadow-test approach is somewhat counterintuitive; rather than picking the best item from the pool, its first step is the assembly of a traditional linear test. But it has been demonstrated to work excellently in a series of recent studies with such applications as highly constrained real-world tests (van der Linden & Reese, 1998), constraints that control the degree of speededness of the individual test administrations (van der Linden, 2008a; van der Linden, Scrams, & Schnipke, 1999), constraints that keep the exposure rates of the test items in the pool below given security levels (van der Linden & Veldkamp, 2004, 2007), as well as constraints that stratify the test with respect to its item-discrimination parameters (= α-stratification) to prevent suboptimal selection of items early in the test because of large errors in the estimates of θ (van der Linden & Chang, 2003). An unorthodox application of the shadow-test approach to adaptive testing is to create number-correct scores on the adaptive test that are automatically equated to those on a linear reference test used for score reporting (van der Linden, 2001).

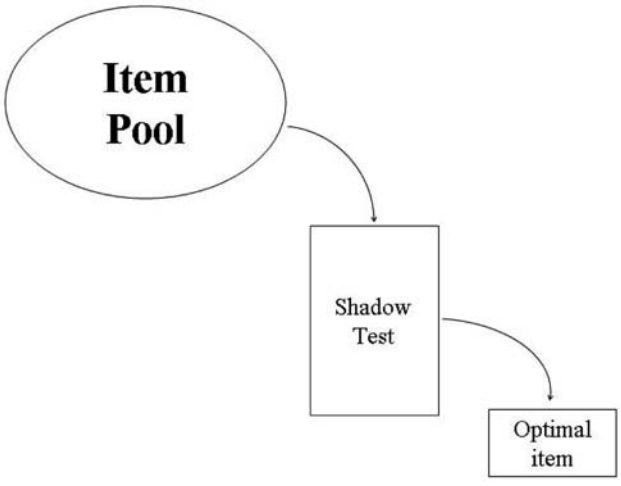

Figure 2. Graphical illustration of constrained adaptive testing using a shadow-test approach.

Adaptive Testing with Rule-Based Item Generation

In the traditional linear format of testing, new items are written and pretested for a single test form each time the previous form has become obsolete. But each new version of an adaptive test requires the replacement of a full item pool. If candidates can take the adaptive test continuously and the security risks are high, the replacement of the pool will draw heavily on the resources of the testing agency and item writers may easily run out of ideas for new but equivalent test items.

One of the earlier solutions to this problem was using item-exposure control techniques to better exploit the item pool. Although, they were originally invented to avoid the risk of item compromise by reducing the exposure rates of the popular items in the pool (Sympson & Hetter, 1985), it was soon recognized that the same techniques have a positive impact on the rates of items in the pool that normally tend to be hardly used (Chang & Ying, 1999). Of course, such applications only work satisfactorily when all items in the pool are of high quality.

A more fundamental approach to the problem is to look into the possibility of mass production of high-quality items by computer algorithms. Ideas for rule-based item generation were already explored in the 1960s, mainly for use in domain-referenced testing (e.g., Hively, Patterson, & Page, 1968; Osburn, 1968). The possibilities to computerize item generation directly for use in adaptive testing has revitalized this area of research.

Different types of item generation have been studied. One straightforward type is the use of item forms or shells in which existent items of superior quality are selected and some of their elements are replaced by large sets, from which elements are randomly substituted. Another approach is to clone items using transformation rules (e.g., linguistic rules or rules that are content based). Research in the area of rule-based item generation is rapidly growing

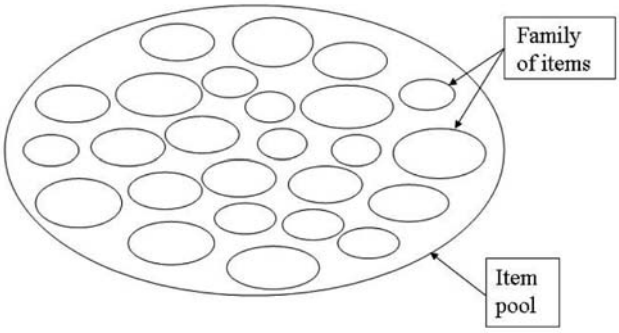

Figure 3. Graphical example of a pool of item families for adaptive testing.

because of its promising profits. For a review of the techniques that are currently under investigation, see Irvine and Kyllonen (2002). A recent example of a successful application, which allows the generation of numerous items for the testing of figural reasoning from a limited set of rules, is presented in Freund, Hofer, and Holling (in press).

However, a problem not yet addressed by rule-based item generation is the costs of item pretesting. In fact, these costs are much higher that those of just writing the items; they involve extensive item review, empirical pretesting, and statistical item calibration. It would be naive to think that these activities could be skipped because the parameters of an item can be used automatically for any other item generated from it. The general impression from empirical research into this issue is that the differences between item parameter estimates are much smaller within item families than between families, but that they are still large enough to not ignore them. For a graphical impression of the typical variation between the item parameters in a pool with families of items generated from a small set of parent items or item rules, see Figure 3.

Substantial savings on item calibration are possible when the selection of the items follows a two-stage procedure in which (i) a family is identified that is optimal at the current estimate of θ and (ii) an item is randomly generated from this family. The first step exploits the between-family differences in the pool and still allows us to be adaptive. The second step randomizes with respect to the generally much smaller within-family differences.

This type of item selection is facilitated by the replacement of the response model by a model with a two-level structure. Suppose we have items families $f = 1,\ldots,F$. The items from family f are denoted as $i_f = 1,\ldots,I_f$. For the Rasch model in ((1)), an appropriate two-level structure would be

$$\Pr\{U_{i_f} = 1\} = \frac{\exp(\theta - b_{i_f})}{1+\exp(\theta - b_{i_f})}, i_f = 1,...,I_f,\ f = 1,...,F, \quad (2)$$

with

$$b_{i_f} \sim N(\mu_f, \sigma_f^2), f = 1,...,F. \quad (3)$$

That is, each family f has a normal distribution of the difficulty parameter with its own mean μ_f and variance σ_f^2. The differences between the families are captured by the parameters μ_f, the differences within the families by the parameters σ_f^2.

Under this model, *item calibration* can be replaced by *family calibration*, that is, estimation of the family parameters from samples of items drawn from them. Once these parameters have been estimated with enough precision, an adaptive test from the pool can be run by choosing each next family to be optimal at the current θ estimate and sampling an item from the family. The savings in item calibration are, thus, directly related to the difference between the size of the sample of items in the calibration and the total number of distinct items that can be generated from the families.

The idea of modeling item pools as a collection of families of items was presented for the more general three-parameters logistic in Glas and van der Linden (2001). Bayesian estimation of the same model was also studied in Sinharay, Johnson, and Williams (2003). A study in which the family model was applied to adaptive testing is reported in Glas and van der Linden (2003).

It is important to notice that this type of two-stage item selection is different from that in the shadow-test approach. In constrained adaptive testing, the two need to be combined: the optimization model for the shadow test is then for the choice of the item families in the test. Once a shadow test is selected, the best family is picked from the free families in it. Finally, an item is randomly sampled from this best family.

Multidimensional Adaptive Testing

Most IRT models have been developed for the measurement of unidimensional abilities and knowledge domains. But recently the interest has shifted to multidimensional models as well. One reason for this shift is that, although such models were already proposed in the 1960s, they were always hard to use in operational testing because of their statistical intractability. But thanks to recent advances in statistics and the availability of plentiful cheap computer power, the situation has changed dramatically and routine use of multidimensional IRT has now become feasible.

We expect multidimensional response models to be particularly useful for adaptive testing. One of the frequent reasons of a response model showing less than satisfactory fit is lack of unidimensionality of the ability parameter. Usual cases are items whose formulation rely too heavily on language or analytic reasoning skills, whereas the focus is on the measurement of a more substantive ability. In traditional pretesting of a new linear test form, generally more items are pretested than needed. And if the assumption of unidimensionality is violated for a small minority of the items, a standard solution is to ignore them and assemble the form from the rest of the items. If the number of violations is large, it is sometimes possible to detect a simple

structure and reorganize the test as a battery of smaller unidimensional tests (multidimensional ability vs. multiple unidimensional abilities; see the next section).

We expect the problem of multidimensionality to be more dominant for adaptive testing because of the necessity of much larger numbers of items required to show a satisfactory fit to the response model. (One exception might be the rule-based generation of item pools discussed in the previous section, which can be expected to result into more homogeneous item pools than pools left entirely to the creativity of a team of different item writers. But more practical experience is needed to pass judgment on this issue.) If multidimensionality appears to be unavoidable, the only route left is to model it and adjust the adaptive testing algorithm for use with the multidimensional model.

This view of multidimensional response modeling as a last resort when violations of unidimensionality become dominant is too negative. A more positive motivation is the recent interest in performance-based testing, with its emphasis on the testing of complex skills in a real-world context. The testing of such performances should always be approached from a multidimensional point of view. And if the goal is diagnosis and each of the dimensions should be tested carefully, the choice of an adaptive format seems obvious.

The change from unidimensional to multidimensional adaptive testing involves an important modification of the item-selection criterion. For instance, for item selection based on item information functions (see Figure 2), the presence of more than one parameter to be estimated during the test complicates the item-selection process considerably because the information functions are replaced by information matrices of size $p \times p$ (with p the number of item parameters), which not only reflects the accuracy of the estimates but also their correlations. Alternatively, in a Bayesian context, the item-selection criterion generalizes to that for a multivariate posterior distribution of the ability parameters.

How to reduce these multivariate entities to a single criterion depends on the goal of the test. For a two-dimensional test, three different goals should be distinguished: (i) both ability parameters are intentional and should be estimated accurately, (ii) only one parameter is intentional and the other is a nuisance parameter, and (iii) the interest is only in a combination of the two parameters, such as their (weighted) mean. Rules for item selection and item pool assembly for these different goals can be derived from the optimal design principle used in statistics to optimize experimental designs or sampling procedures. Applications are given in Mulder and van der Linden (2007) and van der Linden (2005, chap. 9). One optimal design criterion – known as the criterion of D-optimality – has been studied for adaptive testing by Luecht (1996) and Segall (1996). Alternatively, we could follow an entirely different procedure and use the Kullback-Leibler measure to maximize the distance between the joint sampling distribution of the ability estimates before and after the selection of the item (or between the posterior distributions of the ability parameters). For this idea, see Mulder and van der Linden (2008) and Veldkamp and van der Linden (2002).

Adaptive Use of Test Batteries

When a linear test becomes adaptive, the savings in testing time can be used to shorten the length of the test or to increase the accuracy of the scores. These options are helpful to testing programs that struggle with limitations in efficiency. A prominent example are test batteries that are to be administered in a single session but for which the burden of test taking is too large. Examples are the use of testing batteries for vocational counseling or diagnosis for remedial instruction. For both, the goal is to produce profiles of scores for individual test takers. When the profiles are used for high-stakes decision making, the accuracy of their individual scores should be as high as for a regular test. But it is generally impossible to administer a battery of, say, five tests of regular lengths. One of the first applications of adaptive testing was to solve this dilemma between accuracy and total testing time for test batteries (Brown & Weiss, 1977). As a rule of thumb, an adaptive test needs some 40–50% of the items to reach the same level of accuracy as a linear test. This gain in efficiency allows us to spend, for example, 1 h of testing time on a battery of five 10-item adaptive achievement tests for use in remedial teaching and be equally efficient as a battery of five linear tests of some 20–24 items.

The efficiency can be increased further by optimizing the order in which the tests are administered to the individual test takers. Obviously, the best strategy is to choose the order adaptively: The first test is then selected to be optimal over initial estimates of each of the abilities measured by the tests. After the first test is completed, the second test is chosen to be optimal over the predicted abilities on the remaining tests given the test taker's responses on the first test, and so on.

The reason why the efficiency of an adaptively sequenced test battery can be expected to be profitable resides in the typical pattern of convergence of ability estimates on an adaptive test. Due to the randomness of the responses, the estimates are likely to wander around for a while early in the test and convergence only when the responses begin to show an obvious trend. The use of the information from the responses on the previous tests gives the next test a much better start than the typical choice of an arbitrary initial value ability estimate somewhere in the middle of the scale.

A natural statistical framework for adaptive sequencing of test batteries is multilevel IRT. The framework should then consists of a distinct response model for each of the item pools for the battery as first-level models in combination with a second-level model for the joint distribution of their ability parameters for the population of test takers. For

the Rasch model in (1) and a multivariate normal distribution for abilities $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_H)$, the framework becomes

$$\Pr\{U_{ih}=1\} = \frac{\exp(\theta_h - b_{i_h})}{1+\exp(\theta_h - b_{i_h})}, i_h = 1,\ldots,I_h, h = 1,\ldots,H, \quad (4)$$

with

$$\boldsymbol{\theta} \sim MVN(\mu,\Sigma). \quad (5)$$

It is important to notice the differences between this framework and the multilevel structure in (2)–(3), which had a distinct model for each item family and a second-level distribution of their difficulty parameters. Also, this case of multiple unidimensional adaptive tests should not be confused with that of multidimensional adaptive testing in the previous section.

The multilevel structure in (4)–(5) lends itself perfectly for an implementation of adaptive test sequencing using an empirical Bayes approach: The items in the first test are then selected using updates of the posterior distribution of its ability parameter. At the end of the test, the responses are used to calculate the posterior predictive distribution of the abilities on each of the remaining tests in the battery and the test with the most informative predictive distribution is selected. This distribution is subsequently used as a prior distribution for the selection of the first item in the new test. The selection of later tests and items is analogous.

The approach is empirical because the second-level model for the distribution of the abilities is estimated from actual test data during item calibration. For more details of this adaptive sequencing of a test battery and empirical results both for batteries with constrained and unconstrained adaptive tests, see van der Linden (2007). Generally, the gain in efficiency due to the adaptive sequencing is a function of the pattern of correlations between the abilities measured by the tests. Real-world test batteries are usually constructed to measure a set of related but distinct abilities. Hence, we expect these correlations to be substantial, and it would be a waste to ignore the information in them.

Use of Response Times in Adaptive Testing

In adaptive testing, the response times (RTs) by the test takers are automatically recorded. Their information can be explored to improve adaptive testing.

One of the possibilities alluded to earlier is the control of the degree of the speededness of the test for each individual test taker. This control is necessary because test items vary considerably in the amount of time they take. Usually, the more difficult items take more time. As a result, the adaptive algorithm penalizes the more able test takers by giving them the more time-intensive items. Control of differential speededness is only possible if we have an appropriate statistical model for the distributions of the RTs on the items in the pool.

A moment's reflection shows that, analogously to a unidimensional IRT model, the RT model should have separate parameters for the test taker and the items. An RT model with this structure is the lognormal model, which postulates the following normal distribution for the logarithm of the time for a fixed person on an items

$$\ln T_i \sim N(\beta_i - \tau, \alpha_i^{-2}), \quad (6)$$

where τ is a parameter for the speed at which the test taker operates, β_i a parameter for the time intensity of item i, and α_i is its discrimination parameter (van der Linden, 2006). Parameters α_i and β_i can be estimated from the response times collected during item pretesting along with the regular calibration of the item pool. Once these parameters are known, the RTs for any selection of the items from the pool can be used to estimate the speed parameter τ for the test taker.

To control the degree of speededness of an adaptive test, the test taker's τ should be estimated from his/her response times during the test, just as θ is estimated from the responses. As the item parameters are already known, we can estimate the RT distributions of the test taker for each of the remaining items in the pool as soon as we have an estimate of the speed parameter. Initially, the estimates of these distributions are poor but they become quite accurate toward the end of the test, which is exactly where the degree of speededness of the test becomes critical.

The basic idea is to impose a constraint on the selection of the items that requires an estimate of the total time on the remaining items to be no larger than the remaining time available for the test. It is simple to impose such a constraint using the shadow-test approach discussed earlier. Technical details of the approach and empirical examples are given in van der Linden (2008a) and van der Linden et al. (1999).

Another use of the RTs is for the improvement of the item selection. When the ability of the test takers correlates with their speed, which typically happens, the RTs contain valuable information about the test takers' ability. The information can be used to accelerate the convergence of the ability estimates and, hence, of the adaptation of the test.

In a recent study (van der Linden, 2008b), the idea was executed using a hierarchical framework with the 3PL model and the RT model in (6) as first-level models and a second-level model for the joint distribution of ability parameter θ and speed parameter τ. The framework allows us to import the information in the RTs in the estimation of θ in the form of an empirical prior distribution of it. In an study with simulated data, we found considerable gains of efficiency even for moderate (positive or negative) correlation between θ and τ. For example, a 10-item adaptive test with the use of RTs tended to be equally efficient as a 20-item test without the use of them.

The final example of the use of RTs reviewed here is to detect possible aberrances in adaptive testing due to, for instance, differentially functioning test items, ambiguous

instructions, test takers who need other special test accommodations, and cheating in the form of answer copying, item memorization, or item preknowledge. Traditionally procedures for detecting such behavior focus on unexpected responses indicative of item- or person-misfit under a response model that has been shown to fit regular test behavior.

There are at least three reasons why a focus on unexpected RTs might be more efficient. First, during an adaptive test the probabilities for a correct and incorrect response converge to a value close to .50. As a result, response patterns that would normally be indicative of aberrances become also indicative of regular test behavior, and any statistical test based on them loses its power. RTs are insensitive to such effects, and statistical tests based on them keep their power in adaptive testing. Second, RTs are continuous instead of binary and therefore contain much more information on the size of aberrances. Third, it is nearly impossible to fake realistic RTs for test takers who are trying to cheat. RT models with parameter structures as in (6) allow us to adjust the test takers' RTs for their actual speed and to check if the results follow the pattern of time intensities for the items in their test. Even for sophisticated cheaters, it will be impossible to find out during the adaptive test what a regular pattern would be, because they would have to do so while their time on the items elapses. For technical details on RT-based detection of aberrances in adaptive testing, see van der Linden and Guo (2008).

Concluding Remarks

This review shows that the original idea of adaptive testing as simply picking the best item when a new ability estimate becomes available has become somewhat naive. To maintain the validity of the test, items selection has to be constrained considerably, and we may have to deal with complications due to the multidimensionality of the abilities that are tested. On the other hand, substantial further improvements of adaptive testing are possible in the form of rule-base item generation, adaptive sequencing of test use, and exploiting the RTs that are recorded during the test.

The new adaptive testing technology that is emerging has profited much from a decade of pioneering applications in educational testing. We expect it to develop further when we learn from applications in psychological testing as well.

Acknowledgments

A portion of this research was funded by Deutsche Forschungsgemeinschaft (DFG), Schwerpunktprogramm "Kompetenzmodelle der Erfassung individuelle Lernergebnisse and zur Bilanzierung von Bildungsprozessen" (Competence Models for the Measurement of Individual Learning Results and the Monitoring of Educational Processes; SPP 1293), Project "Rule-Based Item Generation of Algebra Word Problems Based upon Linear Logistic Test Models for Item Cloning and Optimal Design."

References

Binet, A., & Simon, Th. A. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectual des anormoux [New methods for the diagnosis of abnormal levels of intellect]. *L'Anneé Psychologie, 11,* 191–336.

Brown, J.M., & Weiss, D.J. (1977). *An adaptive testing strategy for achievement test batteries* (Research report 77–6). Minneapolis, MN: University of Minnesota, Psychometric Methods Program.

Chang, H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211–222.

Dubois, Ph. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon.

Freund, Ph. A., Hofer, S., & Holling, H. (in press). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*.

Glas, C.A.W., & van der Linden, W.J. (2001). *Modeling variability in item parameters in item response models* (Research report 01–11). Enschede, The Netherlands: University of Twente.

Glas, C.A.W., & van der Linden, W.J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*, 247–261.

Hively, W., Patterson, H.L., & Page, S.H. (1968). A "universe-defined" system of arithmetic achievement items. *Journal of Educational Measurement, 5,* 275–290.

Irvine, S.H., & Kyllonen, P.C. (Eds.). (2002). *Item generation for test development.* Mahwah, NJ: Erlbaum.

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

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

Mulder, J., & van der Linden, W.J. (2007). Multidimensional adaptive testing with optimal design criteria for item selection. Submitted for publication.

Mulder, J., & van der Linden, W.J. (2008). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W.J. van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer. In press.

Osburn, H.G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurements, 28*, 95–104.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Segall, D.O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331–351.

Sinharay, S., Johnson, M.S., & Williamson, D.M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics, 28*, 295–313

Sympson, J.B., & Hetter, R.D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings*

of the 27th annual meeting of the Military Testing Association (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

Thurstone, L.L. (1925). A method of scaling educational and psychological tests. *Journal of Educational Psychology, 16*, 433–451.

van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Boston: Kluwer.

van der Linden, W.J. (2001). Computerized adaptive testing with equated number-correct scoring. *Applied Psychological Measurement, 25*, 343–355.

van der Linden, W.J. (2005). *Linear models for optimal test design*. New York: Springer-Verlag.

van der Linden, W.J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181–204.

van der Linden, W.J. (2007). Sequencing an adaptive testing battery. Submitted for publication.

van der Linden, W.J. (2008a). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 32.* In press.

van der Linden, W.J. (2008b). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33.* In press.

van der Linden, W.J., Breithaupt, K., Chuah, S.C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement, 44*, 117–130.

van der Linden, W.J., & Chang, H.-H. (2003). Implementing content constraints in α-stratified adaptive using a shadow test approach. *Applied Psychological Measurement, 27*, 107–120.

van der Linden, W.J., & Glas, C.A.W. (2007). Statistical aspects of adaptive testing. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 27: Psychometrics* (pp. 801–838). Amsterdam: Elsevier.

van der Linden, W.J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73.* In press.

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

van der Linden, W.J., Scrams, D.J., & Schnipke, D.L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement, 23*, 195–210.

van der Linden, W.J., & Veldkamp, B.P. (2004). Constraining item-exposure rates in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273–291.

van der Linden, W.J., & Veldkamp, B.P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32*, 398–418.

Veldkamp, B.P., & van der Linden, W.J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*, 575 588.

W.J. van der Linden

Department of Research Methodology,
Measurement, and Data Analysis
University of Twente
P.O. Box 217
NL-7500 AE Enschede
The Netherlands
Tel. +31 53 489-3581
Fax +31 53 489-4239
E-mail w.j.vanderlinden@utwente.nl

