



2007


What's a Testlet and Why Do We Need Them?

Howard Wainer
University of Pennsylvania

Eric T. Bradlow
University of Pennsylvania

Xiaohui Wang

Follow this and additional works at: https://repository.upenn.edu/marketing_papers

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), [Human Resources Management Commons](#), [Management Sciences and Quantitative Methods Commons](#), and the [Marketing Commons](#)

Recommended Citation (OVERRIDE)

Wainer, H., Bradlow, E.T., & Wang, X. (2007). What's a Testlet and Why Do We Need Them? In *Testlet Response Theory and Its Applications*, 44-59. Cambridge University Press.

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/marketing_papers/394
For more information, please contact repository@pobox.upenn.edu.

What's a Testlet and Why Do We Need Them?

Abstract

In 1987, Wainer and Kiely proposed a name for a packet of test items that are administered together; they called such an aggregation a "testlet." Testlets had been in existence for a long time prior to 1987, albeit without this euphonious appellation. They had typically been used to boost testing efficiency in situations that examined an individual's ability to understand some sort of stimulus, for example, a reading passage, an information graph, a musical passage, or a table of numbers. In such situations, a substantial amount of examinee time is spent in processing the stimulus, and it was found to be wasteful of that effort to ask just one question about it. Consequently, large stimuli were typically paired with a set of questions. Experience helped to guide the number of questions that were used to form the testlet. It is easy to understand that if, for example, we were to ask some questions about a 250-word reading passage, we would find that as we wrote questions, it would get increasingly difficult to ask about something new. Thus, we would find that eventually the law of diminishing returns would set in and a new question would not be generating enough independent information about the examinee's ability to justify asking it. In more technical language, we might say that the within-testlet dependence among items limits the information that is available from that 250-word passage.

Disciplines

Business | Educational Assessment, Evaluation, and Research | Human Resources Management | Management Sciences and Quantitative Methods | Marketing

4

What's a testlet and why do we need them?

4.1. Introduction

In 1987, Wainer and Kiely proposed a name for a packet of test items that are administered together; they called such an aggregation a “testlet.” Testlets had been in existence for a long time prior to 1987, albeit without this euphonious appellation. They had typically been used to boost testing efficiency in situations that examined an individual’s ability to understand some sort of stimulus, for example, a reading passage, an information graph, a musical passage, or a table of numbers. In such situations, a substantial amount of examinee time is spent in processing the stimulus, and it was found to be wasteful of that effort to ask just one question about it. Consequently, large stimuli were typically paired with a set of questions. Experience helped to guide the number of questions that were used to form the testlet. It is easy to understand that if, for example, we were to ask some questions about a 250-word-reading passage, we would find that as we wrote questions, it would get increasingly difficult to ask about something new. Thus, we would find that eventually the law of diminishing returns would set in and a new question would not be generating enough independent information about the examinee’s ability to justify asking it. In more technical language, we might say that the within-testlet dependence among items limits the information that is available from that 250-word passage.

Thus, for the first century or so of its existence, testlets were used simply as a method of constructing tests that contained large stimuli in an efficient way. This all changed in the 1980s when computerized adaptive testing (CAT) became technically and economically possible. A CAT is a test that is tailored to the demonstrated ability of the examinee. It is well known (Lord & Novick, 1968) that a test item provides the most information about an examinee when the difficulty of the item is the same as the examinee’s ability. This is a self-evident

truth. Suppose we are trying to determine an athlete's high-jumping ability by asking her to jump over a sequence of hurdles. If the first hurdle is 20-feet tall, we learn very little when the athlete fails to clear it. Similarly, if the first hurdle is 6-inches high, we learn nothing when it is successfully cleared. If we wanted to ascertain the athlete's ability to an accuracy of, say, 1 inch, we might begin the test with a 6-inch hurdle and continue with 140 hurdles, one every inch, until, say, 12 feet. From the pattern of standing and fallen hurdles, we could estimate the athlete's ability. But it would take 138 hurdles to do it. Suppose the athlete could jump over a 6-foot hurdle, but would miss one at 6 feet 1 inch. We could tell because the first 66 hurdles would be cleared (assuming perfect performance) and the remaining 72 would be knocked over. Note that this "test" would be equally accurate for any athlete whose jumping ability lies between 6 inches and 12 feet. But it takes 138 hurdles to obtain this accuracy, and most of the hurdles are redundant and hence yield little or no information. Now suppose we have an adaptive test. We might use a 6-foot hurdle to begin with. If it is successfully cleared, we might use a 9-foot one (halfway between 6 and 12 feet). If that is missed, we then use a 7.5-foot one (halfway between 9 and 6 feet). If that were missed we would use a 6.75-foot jump etc. In this way, we could measure the performance of anyone whose ability lies in this wide range to within about 1 inch on a test of less than 10 items. That is the power of an adaptive test.

Of course on most tests people don't perform without error. To continue with the jumping example, we might find someone who would clear 6 feet 1 inch after missing at 6 feet, and so a sensible item selection algorithm (the way we choose what is to be the next hurdle (the next item)) should not be quite as drastic as the "split the difference" method described previously. Practical, item selection algorithms tend to move in smaller increments, depending on the character of the item pool and the pattern of the examinee's responses. If the pool is rich in items, having a broad choice at every level of difficulty, the algorithm can pick an item at or near the level desired. If it is sparser, the gaps between items may be suboptimal. So, for example, if a hurdler's true ability is 2.5 feet and after she clears a 2-foot hurdle, the only taller hurdles we have are 3 feet and above, we can never estimate her ability with an error of less than 6 inches. This lack of precision is due to limitations of the item pool.

If the examinee responds in a regular way, getting difficult items wrong and easy items correct, convergence to an accurate estimate of proficiency can be done quickly. If the response pattern is more ragged, convergence is slower and the resulting estimate is less precise. Returning to our overworked hurdler, if she knocks over a 2-foot hurdle, but clears a 30-inch one, then misses at

34 inches but clears a 36-inch one before missing all the rest, our estimate is inaccurate due to the hurdler's inconsistency, not a shortcoming of the item pool.

This example illustrates how the precision of a test can be affected by both the character of the test and the character of the examinee. The first can be controlled by the test developer, the second cannot.

The promise of adaptive testing has enthralled psychometricians for more than thirty years. It began with Lord's initial discussion of his "flexilevel test" in 1971 and, in the same year, his use of the Robbins–Monro procedures to automate it. Lord's later work (1971a, 1971b, 1974, 1980) extended and deepened the initial ideas. David Weiss's preliminary experience in military testing (1974, 1982; Weiss & Kingsbury, 1984) showed how CAT could be practical. Finally, a broad consilience was provided in the enthusiastic "how-to" manual by Wainer et al. in 1990. But in the decade following the publication of this *CAT Primer*, adaptive tests were implemented in some very large operational programs: the U.S. military's ASVAB, the Graduate Management's GMAT, the Graduate Record Examination (GRE), and many others, yielding millions of adaptive test administrations a year. The data thus generated have provided illumination on issues that previously were cloaked in darkness. We discovered that now that we had a pretty clear understanding of how to do adaptive testing, we needed to focus much more seriously on when – and more specifically, on when not to. Chapter 10 of the 2nd edition of the *CAT Primer* (Wainer et al., 2000) provides a more clear-eyed look at these issues than did its predecessor.

One of the issues that emerged from all of this experience relates to a fundamental assumption underlying CAT – that an item's characteristics remain constant regardless of its context – this is usually called the assumption of item fungibility (these are also known as context effects in survey design). This assumption within our jumping test would mean that a 3-foot hurdle is as difficult to jump over whether it follows a 1-foot hurdle or a 6-foot one; obviously a crucial assumption if an item-selection algorithm is to function properly. It is also an assumption whose credibility is rarely tested in operational CATs.

This is but one example of the kind of issues that arise when tests are custom-built for each examinee, one item at a time. In the balance of this chapter, we will discuss a number of such problems in greater detail and indicate what have been the traditional approaches toward solving them. Then we shall conclude with a discussion of how increasing the size of the fungible unit of test construction to the testlet ameliorates these problems.

4.2. Problems

4.2.1. Context effects

What is a context effect?

The term *context effect* refers to any influence that an item may acquire purely as a result of its relationship to the other items making up a specific test. In a traditional testing situation in which everyone takes the same test, context effects, such as they exist, will be the same for all examinees. And so, although they still may induce interpretative errors, there is less concern about their differential impact.

In the case of CAT, there may be a much greater problem of differential context effects, because every examinee potentially takes a different test. Thus, each item's context is potentially different for each examinee. Furthermore, in the CAT real-time approach to test construction, the phase of test development where context effects might ordinarily be identified and eliminated from nonadaptive tests has no counterpart.

Item location

One undesirable item context effect is the influence that item location within a test, or in different tests, has on the item parameter estimates. Several studies (Eignor & Cook, 1983; Kingston & Dorans, 1984; Whitely & Dawis, 1976; Yen, 1980) have found differences in difficulty parameter estimates for items as a function of their location. Whitely and Dawis found that when 15 analogy items were embedded in seven different unspeeeded 60-item tests, and parameters were then obtained separately for each test, the estimates differed significantly for 9 of the 15 items. Yen found difficulty differences due to context effects for reading comprehension items administered in a unspeeeded test, with the items becoming more difficult the later in the test they appeared. A preequating study conducted by Eignor and Cook replicated Yen's findings.

In a study on the implications of item location for both equating and adaptive testing, Kingston and Dorans (1984) examined whether the difficulty of certain types of items is affected by exposure to the same item types earlier in the test. They found that for those item types in particular (analysis of explanations and logical diagrams) difficulty decreased substantially as familiarity with the item format increased. These effects, they concluded, were due, at least partially, to the complexity and novelty of the item types. For another type of item (reading comprehension), difficulty increased moderately as the item appeared later in the test. They attributed this outcome to fatigue.

Large location effects for even moderately speeded tests are ubiquitous. In a study of speededness, Wainer et al. (2004) found the unsurprising result that the more speeded the test, the greater the increase in difficulty for items situated toward the end of an SAT math test.

Cross-information

Another undesirable effect of context that may appear on a test is cross-information. *Cross-information* refers to information that one item may inadvertently contribute about the answer to another. For example, suppose the following item appears on a test for some individuals but not for others:

1. Carbon dioxide (CO_2) is a component of all of the following except
a. seltzer b. ammonia c. "dry ice" d. photosynthesis
but then some of those who answered this incorrectly, as well as some who never saw it, were presented with the easier question
2. The symbol for carbon dioxide is
a. CO_2 b. H_2O c. NH_4 d. C_2O

Surely, those who had seen Question (1) would have an easier time with (2). Or, put in more general terms, the difficulty of (2) is dependent upon what preceded it. This is always true in test construction, but its effects are controlled for in two ways. First, test developers carefully construct tests to avoid such dependencies. And second, since everyone who takes a fixed test receives exactly the same questions, any such hints are fairly distributed – everyone gets the same one – and so no one is unfairly advantaged. With a test that is constructed by any algorithm that does not take into account the precise content of the items, dependencies among items can occur. This is a problem, but a relatively venial one. A test that is algorithmically constructed to be tailored to the individual examinee can yield dependencies that are unfairly distributed among examinees. This is a more serious problem.

Unbalanced content

A third type of context effect devolves from the content balancing of tests. All well-developed tests are built around content specifications. These are the content areas the test developer feels that the test ought to span. For example, in an arithmetic test, we might want to have 25% of the items deal with addition, 25% with subtraction, 25% with division, and 25% with multiplication. We will refer to specifications such as these as formal content specifications, since they deal with the formal contents of the subject area. There are also informal content specifications, which are usually not explicitly stated, but are both real and important. As an example of these, consider the structure of a problem in

which the actual task is embedded. Suppose our arithmetic test consists of many “word problems” such as

3. If John caught 6 fish and threw 3 back and Mary caught 7 fish but threw 4 back, who brought home more fish?
 - a. John
 - b. Mary
 - c. Both the same
 - d. Can't tell from information given

Test developers have found that it is not wise to have too many problems dealing with the same topic (fishing); nor even the same general area. In the review of one test, one criticism was that there were too many “water items” (fishing, water skiing, boating, swimming, canoeing, etc.). While it may not be obvious why too many “water items” would be unfortunate, it is easily seen how some subgroups of the general population would be disadvantaged if there were many items on golf. Aside from the obvious social class effect, it also increases the potential for contributory information across items. Thus, in addition to filling specifications regarding the formal content, test developers must be careful to balance the test with respect to the informal contents. In a fixed-form test, it is straightforward (although not always easy) to be sure that the specifications are filled satisfactorily. Moreover, test developers can also read over the form carefully to assure themselves that there is no imbalance vis-à-vis the informal content.

It is not too hard to imagine how one could structure a computer algorithm to construct a test that would balance its content (e.g., van der Linden & Boekkooi-Timminga, 1989). The candidate items would be classified by their formal content and a categorical choice algorithm (“choose one from group A and two from group B”) could be instituted. Of course, there may be need to cross-classify items by their difficulty as well, but this is just a bookkeeping task, and does not pose a complex technical problem, except for the need to write items in each content area that cover the entire range of difficulty. This does pose a problem on broad range tests, since, for example, it may not be easy to write sufficiently difficult arithmetic items or sufficiently easy calculus items. But this too can be surmounted if the unidimensionality assumption underlying IRT holds reasonably well: for in this case, the item's parameters are all that are required to characterize the item on the latent variable of interest.

It is more difficult to try to conceive of any categorization scheme that would allow a computer algorithm to determine if there was an overabundance of items on an inappropriate subject matter – where the subject matter was, in some sense, incidental to the item's content. To accomplish this requires either a finer level of item characterization (and hence a huge increase in the size of

the item pool) than is now available or a level of intelligence on the part of the algorithm that is far beyond anything currently available.

Another type of content imbalance is in reference to gender, race, or ethnic groups across items. Such an imbalance violates current test development standards and, some claim, may serve to distract examinees and create feelings of hostility or resentment that could hinder their test-taking performance. An example of this might be a selection of a succession of items that use only female, or only male, references in professional roles.

4.2.2. Item difficulty ordering

A different kind of context effect, one that is not undesirable but rather is often planned and built into linear, nonadaptive tests, is the sequential ordering of items by their increasing difficulty. In the conventional testing situation, a test or section begins with items whose difficulties are less than all but the least able of the examinees; the difficulties increase as one progresses through the test, concluding with items whose difficulty exceeds the ability of all but the most able.

It is generally believed that beginning a test with the easiest items and ordering them by increasing difficulty allows examinees, especially those who are least able, to develop confidence at the outset of the test with the easiest items. They can then work through the test to their capacity in the time allowed, without wasting time or becoming unnecessarily discouraged by items that are far too difficult for them. This approach also tends to minimize guessing, because items on which those who are least able (and therefore most prone to guessing) spend the greatest amount of time are those on which they have the greatest likelihood of knowing the answer.

In an adaptive test, however, the initial item administered is usually one of moderate difficulty for the entire group being tested. Thus, for an examinee more able than average, the difficulty of successive items increases from this rather easy start, just as with a conventional linear test. For those at the low end of the ability continuum, however, the test begins with items that are excessively difficult. These become less difficult as the test progresses and incorrect responses accumulate, until the items emerge into a range of difficulty suitable for this group. But the items come at the examinees from above rather than below, contrary to traditional testing lore, suggesting that it may disadvantage lower ability groups. Indeed, there is considerable evidence, spanning half a century, that when items are administered from most to least difficult, rather than in the more usual order, the difficulty of the test as a whole increases (Hambleton, 1986; MacNicol, 1956; Mollenkopf, 1950; Monk & Stallings, 1970; Sax & Carr, 1962; Towle & Merrill, 1975).

4.2.3. Summary of context effect problems

In the nonadaptive testing situation, a test developer can read and study each test directly during its development phase to identify and mitigate context effects. Unfortunately, CAT makes such a strategy difficult or impossible. At present, the only way to ensure that cross-information is never inadvertently provided on a CAT is to compare all conceivable subsets of items derivable from the item pool. This is at best an unreasonable undertaking, especially because a viable item pool must contain at least hundreds of items and is continuously being updated, with new items being added and old ones retired. The identification and elimination of the other context effects just discussed is even more difficult under the traditional adaptive framework. It would require an analysis of the content of each and every test produced by the CAT item-selection algorithm. This is impractical with current technology.

4.2.4. An issue of robustness

The increased efficiency that is the hallmark of adaptive testing – the same accuracy of measurement as a linear test at a greatly reduced length – is not an unalloyed good. The shorter test lacks the redundancy inherent in the almost twice as long conventional linear test and so becomes more vulnerable to idiosyncrasies in item performance. As a result, when an item is failing to perform in accordance with its parameter estimates on a CAT, the detrimental impact on validity is about twice what would result from a traditional nonadaptive test of about the same theoretical accuracy – but of greater length (Green, 1988). Furthermore, if the flawed item is one of moderate difficulty and relatively high discrimination, its detrimental influence is likely to be felt many times because most CAT item-selection algorithms favor such items (Wainer, 2000).

4.3. Traditional solutions

Unfortunately, there is no obvious solution for all of the variety of problems related to implementing CAT that we have just discussed. However, certain avenues of investigation invite exploration in the search for solutions. Most of these avenues involve a more variegated stratification of the item pool. Thus, the statistical item-selection algorithm would be limited to choosing the most informative item from within rigidly controlled strata. Inexorably, this means generating an item pool of substantially greater size than was initially envisioned by CAT enthusiasts.

4.3.1. Context effects

It may be possible to stratify finely enough to avoid cross-information, to allow the balancing of gender, race, and ethnic references, and to prevent redundancy of item topics. Such stratification does nothing to improve the fungibility of the items. Items still remain context dependent. As long as different examinees are being compared to one another on the same items in different contexts, this problem will remain. It does not seem likely that the solution to this is statistical; trying to adjust an item's parameter values on the basis of its (possibly) unique circumstance is beyond current technology.

4.3.2. Robustness

The increased influence of a single item within a CAT framework is a double-edged sword. While a flawed or inappropriate item can have an undue influence on the ability estimate, an additional corrective item can get us back on the right path doubly quickly. Thus, we feel that it is important to emphasize that CAT users ought to leave in some redundancy so as to allow for an unforeseen flaw. It is probably wise to choose a conservative estimator for the standard error of estimates of ability as well, since many CATs have stopping rules based on this criterion. This avoids stopping an examination "too soon" due to possibly incorrect items that could have the unhappy consequence of an ability estimate that is not as precise as required.

4.3.3. Item difficulty ordering

The effect of the order in which items are presented can be ameliorated by choosing a starting point for the CAT that is toward the low end of the ability continuum. The lower this is, the less the order effect problem, for more and more examinees will have the items ordered from easy to difficult. However, the lower the starting value, the less efficient the CAT algorithm. This may be a relatively venial problem if big jumps are made initially in item choice, which is consistent with most CAT item-selection algorithms.

4.4. Testlets – an alternative solution

The problems that arise from adaptive testing can be eased through an alternative approach. The key idea is to use a multi-item testlet as the fundamental unit of test construction and test administration. We shall define a *testlet* as a group of

items that may be developed as a single unit that is meant to be administered together. Although the path through a testlet could be branched, our focus, at least for now, is on linear testlets that contain n items, where n could be as few as one, but more typically would have four or more items. All examinees who are presented with a particular testlet would be confronted with the same items in the same order.

One example of a testlet is the traditional reading comprehension item type in which the examinee is presented with a passage and a bundle of items related to that passage. In similar vein, science testing commonly uses a graph or table as the focus of a set of related items, and many history and geography tests use a map as the central stimulus for a set of items.

How does this concept help us to resolve the problems stated previously? First, although a testlet-based test can be made adaptive by hierarchically structuring the testlets, the path through the testlet itself can be carefully examined to ensure that it does not suffer from the same problems as items within a standard CAT. For example, the items within a testlet can be ordered from easy to difficult. Thus because order effects are localized, their effects are greatly diminished. This is roughly akin to the way that on linearly administered tests, test sections are difficulty ordered, but when the examinee moves to a new section, the items within it are ordered from easy to difficult. Obviously, most testlets are shorter than a typical test section, but the general idea is the same.

In addition, when blunders do occur with the testlet model, they can be more easily overcome, since all examinees who receive a particular testlet receive exactly the same testlet. Thus any blunder that occurs, occurs repeatedly and can be more easily detected, and test scores of those affected adjusted accordingly.

Of course, testlets have other advantages, quite apart from those associated with providing greater control within an adaptive framework. By far, their greatest value is that they provide another formal mechanism for test developers to test the construct in a way that makes the most sense.

In the late 1980s, when adaptive testing was still in its infancy, the first large-scale adaptive test was the adaptive version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). The ASVAB, fully described by Sands, Waters, and McBride (1997), is used by the military to classify and select about two million young men and women every year, making it the largest single testing program in the world. By making the ASVAB adaptive, substantial savings in time and money were realized. But the test had to have a few modifications in its transition to computer-based administration. One key change, important for this discussion, was in the section on paragraph comprehension. The CAT-ASVAB was scored using traditional IRT. As we mentioned in

Chapter 1 and discussed in greater detail in Chapter 3, a key requirement for IRT to hold is that the items of the test should be conditionally independent (otherwise Equation (3.4) would not be valid). Yet, it was clear that traditional paragraph comprehension items, because they are connected together with a common passage, are not likely to be conditionally independent. This put the developers of the CAT-ASVAB into a quandary. If they left the paragraph comprehension section alone, the scoring model was very likely to be incorrect. But how to change it? The typical passage was about 300 words long and was accompanied by five items.¹ If they just used a single item, the local dependency problem would go away. But getting only one item's worth of information after asking an examinee to spend so much time reading and figuring out a long passage seemed wasteful of effort. It was also antithetical to the efficiency goals that are fundamental to adaptive testing. What to do? It was decided to ask but a single question while shortening the typical paragraph to about 120 words.² The new items showed an increase in their correlation with word knowledge and correlated only 0.38 with the prior form of paragraph comprehension items administered in the paper-and-pencil format (ASVAB P&P form 10B – Moreno & Segall, 1997, p. 172). The median correlation of all of the other ASVAB subtests with their CAT counterparts was 0.72. Thus, the shortening of the passage solved the immediate problem, but, as it turned out, changed the construct being measured.

If, instead, the passage and its associated items were thought of not as a set of independent items but rather as a testlet, to be administered as a unified whole, the problem would have solved itself. But how?

In 1988, Paul Rosenbaum made an important distinction. He pointed out that local independence could be violated because of idiosyncratic features of the test format (e.g., the same passage associated with several items) or because of actual departures from the unidimensionality of the test construct. Rosenbaum (1988) proved (Theorem 1) that given the loss of conditional independence within testlets, unidimensionality can still prevail between testlets. Hence, if we use the testlet as the unit of measurement, a unidimensional item-response model may still be appropriate.

Thus, Rosenbaum's theorem permits the use of items with desirable characteristics, with less concern for violation of the assumption of conditional independence. Hence, traditional paragraph comprehension items, when scored as testlets, can be used without additional concern by the test developer.

¹ P&P ASVAB forms 23, 24, 25, and 26 had 5 items per passage, and the typical passage ranged up to 326 words (John Welsh, personal communication, July 9, 2004).

² CAT ASVAB forms 1, 2, 3, and 4 had 1 item per passage, and the typical passage had no more than 120 words (John Welsh, personal communication, July 9, 2004).

One of the strengths of Rosenbaum's result is that if a traditional IRT model shows a lack of fit when applied to testlet data, a statistical test using contingency table methods can be applied to determine whether the lack of fit is due to a loss of conditional independence within testlets. If this is found to be the case, examinee performance can still be summarized with a single parameter, as the between-testlet scoring may still satisfy the requirements of a unidimensional model.

4.5. An initial summing up

The problems discussed in this chapter are not certain to occur within the context of adaptive testing, nor are they guaranteed to be cured through the use of testlets. Rather, we believe that the bundling of items into testlets provides a mechanism through which the likelihood of these unfortunate circumstances will be reduced without seriously diminishing the efficiency of an adaptive test and will allow the test's structure to more closely match the constructs that the users of the test want to draw inferences about. More specifically:

1. Context effects

- a. *Cross-information*: By enlarging the fungible unit of the test from the item to the testlet, we are reducing the "boundary effects" of presentation. Only the first item in the testlet has an unknown predecessor; all the others follow an item that experts have judged to satisfy the canons of good practice. These boundary effects include such things as cross-information, which, though not entirely eliminated, has had its likelihood reduced substantially. Moreover, if it does occur, the contaminating item is likely to be more distant and so reduces its effect.
- b. *Unbalanced content*: By controlling the presentation more tightly, we can better ensure that the test specifications are satisfied for each individual test. For example, we know of an adaptive science test that consisted of biology, chemistry, and physics questions. The physics questions turned out to be somewhat more difficult than either of the other areas. One student, who studied biology and physics, but not chemistry, found that the questions presented to him vacillated between biology (which he tended to answer correctly) and chemistry (which he tended to miss). He never got to demonstrate his expertise in physics. Such a situation reflects the lack of unidimensionality within the test and could be solved within a traditional CAT framework, but its solution is accomplished so much more gracefully with three testlets in series.

2. *Robustness*: Once again, this is a double-edged sword. If a flawed item appears in a testlet, measurement accuracy will be affected, but fairness can still be served. With a testlet, examinees whose abilities are near one another will receive similar tests. The closer together they are, the closer together their tests will be.³ Thus, the appearance of a flawed item will tend to have a diminished impact on individual comparisons. CATs provide protection by allowing recovery from what turns out to be an anomalous response. Testlets also provide protection, because they are constructed as a whole and the plausible interplay among its members could be subjected to expert judgment in preparation, and statistical quality control in pilot testing. Also, because a testlet is self-contained, any test construction errors that do show are localized and the offending part can be replaced without overly disturbing the rest of the test.
3. *Order effects*: Item difficulty ordering can be controlled by controlling starting points. Even if this is not done, we once again obtain fairness because of the testlet characteristic of having those whose ability estimates are close having closely matching tests.

The strength of the testlet approach derives from two of its characteristics. The first is that it allows greater control of item presentation. The benefits of this were just elaborated upon. The second is that it enforces the notion that the more difficult it is to discriminate between two examinees, the more identical the test they took is likely to have been. Thus, while we may still have problems with multidimensionality – flawed items and other deviations from the test model – those problems will be the same for the particular individuals being compared. We do not mean to imply that test equating does not yield valid scores that allow comparisons across different test forms. This is certainly true to some extent. But the legitimacy of such comparisons depends on the extent to which all of the assumptions of the test scoring model and the equating method are upheld in the data; the smaller the difference between the two individuals being compared, the greater the likelihood that the standard error of the scores will affect the validity of the comparison. When two individuals take precisely the same test, the error of equating is removed from the mix (and substituted for it are the ills associated with problems of security – if one person took the test

³ With the security precautions now common, it is possible that two examinees with identical abilities can, in theory, get two very different but parallel test forms. But this outcome relies on an item pool of substantial size – indeed, it needs to be very large at all levels of ability. Even for tests with hundreds of thousands of examinees, and with the resources that such volume provides, test items are sparse at some levels of ability, and very few items account for most of the administered tests (Wainer, 2000). Thus, for tests constructed within more modest circumstances, it is likely that individuals with similar abilities will receive a substantial number of identical items.

well before the other, there is always the chance of some sort of security leak that would allow the later examinee to have access to the items).

Testlets will not cure this, but they can help to ameliorate it by ensuring that if two examinees' test forms share a common testlet, they will share all items in the testlet. This means that, for those items, no equating is required, and hence the error of equating is reduced.

The use of testlets also requires a shift in our thinking about psychometric modeling. True score theory (Lord & Novick, 1968) might also be called "test response theory" since it uses as its fungible unit something that is too large for adaptive testing – the whole test. Traditional summary statistics, such as true score, test reliability, and test validity focus on the entire test as the testing unit. Item response theory focuses on the item and it produces statistics, such as item difficulty, which are points on an underlying continuum. The examinee's score on the test is also estimated as a point on the same continuum. These estimates and their associated measures of precision stand alone and are not referred to the specific items taken; the underlying trait is all that matters. We are recommending a middle path that we term "testlet response theory" since it uses, as its unit of measurement, pieces of the test that are simultaneously small enough to be usefully adaptive and large enough to maintain some stability. Tests have long been divided into different sections; our contribution is to suggest a more flexible unit, the testlet. The testlet can be as small as a single item (although in this extreme case, none of the advantages discussed here would hold), as large as the entire test, or anything in-between. Our point is that the size of the testlet should be determined by the character of the constructs being measured, not the available psychometric theory. In the balance of this monograph, we describe formal psychometric theories that can accommodate testlets of any character.

Questions

1. What testing problems was the invention of the testlet meant to address?
2. Why was the solution to these problems made more urgent with the popularization of CAT?
3. How does the use of testlets ameliorate these problems?
4. What crucial role did Rosenbaum's theorem on the independence of item bundles play in TRT?

References

- Eignor, D. R., & Cook, L. L. (1983, April). *An investigation of the feasibility of using item response theory in the preequating of aptitude tests*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

- Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 77–86). Hillsdale, NJ: Erlbaum.
- Hambleton, R. K. (1986, February). *Effects of item order and anxiety on test performance and stress*. Paper presented at the annual meeting of Division D, the American Educational Research Association, Chicago.
- Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT and adaptive testing. *Applied Psychological Measurement*, 8, 146–154.
- van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika*, 54, 237–248.
- Lord, F. M. (1971a). The self-scoring flexilevel test. *Journal of Educational Measurement*, 8, 147–151.
- Lord, F. M. (1971b). Robbins-Munro procedures for tailored testing. *Educational and Psychological Measurement*, 31, 3–31.
- Lord, F. M. (1974). Estimation of latent ability and item parameter when there are Omitted responses. *Psychometrika*, 39, 247–264.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- MacNicol, K. (1956). *Effects of varying order of item difficulty in an unsped verbal test*. Unpublished manuscript. Educational Testing Service, Princeton, NJ.
- Mollenkopf, W. G. (1950). An experimental study of the effects on item analysis data of changing item placement and test-time limit. *Psychometrika*, 15, 291–315.
- Monk, J. J., & Stallings, W. M. (1970). Effect of item order on test scores. *Journal of Educational Research*, 63, 463–465.
- Moreno, K. E., & Segall, D. O. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (chap. 17, pp. 169–174). Washington, DC: American Psychological Association.
- Rosenbaum, P. R. (1988). A note on item bundles. *Psychometrika*, 53, 349–360.
- Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.
- Sax, G., & Carr, A. (1962). An investigation of response sets on altered parallel forms. *Educational and Psychological Measurement*, 22, 371–376.
- Towle, N. J., & Merrill, P. F. (1975). Effects of anxiety type and item difficulty sequencing on mathematics test performance. *Journal of Educational Measurement*, 12, 241–249.
- Wainer, H. (2000). Rescuing computerized testing by breaking Zipf's Law. *Journal of Educational and Behavioral Statistics*, 25, 203–224.
- Wainer, H., Bridgeman, B., Najarian, M., & Trapani, C. (2004). How much does extra time on the SAT help? *Chance*, 17(2), 17–23.
- Wainer, H., Dorans, D. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer (2nd ed.)*. Hillsdale, NJ: Erlbaum.

- Wainer, H., Dorans, D. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–201.
- Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Psychometric Methods Program.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473–492.
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.
- Whitely, S. E., & Dawis, R. V. (1976). The influence of test context effects on item difficulty. *Educational and Psychological Measurement, 36*, 329–337.
- Yen, W. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*, 297–311.