RR-89-29

RESEARCH REPORT

TOWARD A PSYCHOMETRICS FOR TESTLETS

Howard Wainer
Charles Lewis

Educational Testing Service
Princeton, New Jersey
June 1989

Toward a Psychometrics for Testlets [§]

Howard Wainer
and
Charles Lewis

Educational Testing Service

Abstract

In 1987 the *testlet* was introduced as one way of dealing with a variety of problems that might occur with the algorithmic construction of tests. In the short time since this introduction its range of plausible usages has been broadened considerably through the work of other researchers. In this paper we examine three different applications of the testlet concept and describe the psychometric models which seem most suitable for each application. In each case, the testlet concept gracefully solves a problem that would have been awkward with other, more traditional approaches.

I. Introduction

The concept of the *testlet* was explicitly introduced by Wainer & Kiely (1987, p. 190) as: " a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow." We proposed the use of the testlet as the unit of construction and analysis for computerized adap-

[§] This research was supported by the Educational Testing Service's *Program Research Planning Council*; we are grateful for the help that this provided. We would like to express our thanks to Jim Braswell for his development of the hierarchical algebra testlets described in section IV; to David Thissen for his thoughts on the psychometrics of testlets; to Kathy Sheehan for her help on section II; to Michael Zieky for his advice on the plausibility of testlets in test construction; to Vic Bunderson for his enthusiasm and support of the testlet notion; to Jamie Algina and two wise, careful, but anonymous referees.

This paper will be appearing in Volume 27 of the *Journal of Educational Measurement*, all references should be made to that source.

tive tests (CATs) with the expectation that they could ease some of the observed and prospective difficulties associated with most current algorithmic methods of test construction. Principal among these difficulties are problems with context effects, item-ordering and content balancing. *Context effects* arise when the appearance of a particular item has an effect on the difficulty of a subsequent item. 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	*c. "dry ice"*
b. ammonia	*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	*c.* NH_4
b. H_2O	*d.* π.

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 which 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. Testlets reduce the likelihood of occurrence of such events as well as the severity of their impact. They do this by allowing test developers to construct testing units that are larger than a single item. This unit, the testlet, is developed and then screened specifically so that there are no unfortunate dependencies within it. Moreover, any examinee who gets one item from the testlet gets all of the other ones as well. Thus any test development algorithm used can safely choose any testlet without concern that such an error will occur within that testlet. Of course, this does not prevent these problems from showing up between testlets, but their likelihood is reduced (because different testlets usually reflect somewhat different subject areas, and so items which inform on one another are less likely to occur in separate testlets). Moreover their impact is reduced as well, since they are likely to be spaced further apart in the presentation sequence. This is an especially effective preventative in a computerized test, since most presentation algorithms do not allow the examinee to return to earlier items.

A second area where the testlet concept may be of some help is in reducing the untoward effects of *item ordering*. Test developers have evolved useful rules about the ordering of items within tests. Specifically, tests are often designed to start out easy and end up hard. Such designs, so-called *power tests*, are designed this way for several reasons. One is so that examinees of lower proficiency are encouraged by initial success, and so will work harder at the solution to the more difficult items that may be at the boundaries of their performance. There is some empirical evidence (Hambleton, 1968; MacNicol, 1956; Mollenkopf, 1950; Monk & Stallings, 1970; Sax & Carr, 1962; Towle & Merrill, 1975) that this is true. Yet item selection algorithms for CATs frequently violate these ordering rules. Maximal efficiency is obtained when the initial item in a CAT is one of middling difficulty, and the subsequent ones are chosen as a function of the examinee's success. If the examinee gets these items correct, more difficult items are chosen. This yields a test whose items are ordered according to the time-honored tradition described previously. Yet for all examinees whose proficiency is below the middle have items coming at them from the top down. This can mean that those examinees with below middle proficiency will have a tougher time of it. Of course this can be modified by choosing a lower-than-optimal starting point. This will reduce the ordering effect, but also reduce the efficiency gains possible through adaptive testing. A second possibility is to ignore the effects of ordering, since they will be small locally (those examinees with similar proficiencies will have tests ordered in about the same way) and the prospective size of the order effects is probably not enough to endanger the validity of inferences about the gross ordering of examinees. As we will see in sections II and III, testlets can help with this problem as well, since items-within-testlets follow a fixed order which is predetermined by a human test developer. In section IV, where we describe a prestructured hierarchical testlet, sometimes this extra control still doesn't completely solve the problem of difficulty ordering, although the ordering is strictly controlled.

The third area of testlet usefulness is *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" like:

(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	*c. Both the same*
b. Mary	*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, swim-

ming, 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 dealing with polo chukkers. 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, they can also read over the form carefully to assure themselves that there is no imbalance vis-a-vis the informal content.

It is not too hard to imagine (in theory at least) how one could structure a computer algorithm to construct a test that would balance its content. The candidate items would be classified by their formal content and a "Chinese Restaurant" choice algorithm ("choose one from column A and two from column 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 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 item response theory (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 over abundance 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 which is far beyond anything currently available.

Once again, the concept of a testlet can be quite useful in content balancing a test. While this is somewhat different from Wainer and Kiely's original formulation (quoted at the beginning of this paper) one might, for example, balance the content of each testlet. This could be done with the aid of human test developers and would prevent the over-dependence on any one area of subject matter. Each testlet could also be balanced in terms of the average difficulty for applications where that is appropriate (see Section II). Testlets constructed in such a way could then be combined using a rather simple algorithm. One successful application of such a scheme is described in the next section.

We can summarize the purposes for which testlets were developed into two categories; *control* and *fairness*.

Control, in the sense that by redefining the fungible unit of test construction as something larger than the item, the test developer can recover some of the control over the structure of the finished test that was relinquished when it was decided to use an automatic test construction algorithm.

Fairness, in the sense that all examinees who are administered a particular testlet, in addition to getting a sequence of items whose content and order have been prescreened and approved by a test developer and an associated test development process, also get the same sequence as other examinees whose observed proficiencies are near theirs. Thus when comparisons among examinees of very similar proficiency are made, those comparisons will be made on scores derived from tests of very similar content.

Since the introduction of testlets a number of prospective uses have emerged. Some of these were foreseen in the original presentation, some were not. In this paper we describe three uses of testlets, which illustrate three different kinds of testlet construction. We derive psychometric models which can be used to score tests composed of testlets in each of these forms. We do not, by our categorization here, mean to imply that these are the only way that testlets can be constructed, nor are these the only psychometric models that will prove to be efficacious. Rather we believe only that these are reasonable ways to proceed. Further, we have purposely chosen quite different psychometric characterizations, for we do not yet have enough experience with these models to have formed strong biases vis-a-vis their relative appropriateness. Therefore rather than being prescriptive here we have decided to try to be expansive. Actually, it is our current belief that no single psychometric characterization can comfortably wear that euphonious appellation of "the best." We believe that, depending upon the prospective situation, different methods will be needed in different circumstances, to serve our needs best.

II. NCARB Example — Content balanced testlets, randomly selected

(Lewis & Sheehan, 1988)

The National Council of Architectural Registration Boards (NCARB) commissioned ETS to develop a Test of Seismic Knowledge for use in the architectural certification process. There was a very broad range of knowledge and experience among the prospective examinees. Some were so expert in these matters that even a very cursory oral examination would reveal this quite clearly. On the other extreme there were some examinees whose knowledge of this area was so sparse that this too would be revealed with only a quick look. Many prospective examinees fell comfortably between these two extremes, and so required an examination in more depth. It was hoped that a test could be constructed that would adapt itself to the level of discrimination required. If the judgement was an easy one ("pass" or "fail") it would do so with alacrity (using as few items as required for the precision of measurement needed). If the judgement was more difficult, the test would lengthen itself to yield the required accuracy. Thus it was desired to have a test that as questions were asked of the examinee, and their responses noted, the testing algorithm would continually ask itself, "Pass?" "Fail?" "Keep testing?" This determination would be made on the basis of a Bayesian decision process in which a loss function was constructed based upon the costs associated with (1) passing someone who should have failed, (2) failing someone who should have passed, or (3) asking more questions. It must be remembered that this is a certification test where only a "pass-fail" decision is to be made. Thus it is more efficiency to concentrate testlet difficulty in the region of the decision point than to attempt

to confront examinees with items which are appropriate for their proficiency level, as would be the case in a traditional adaptive test.

The powerful statistical machinery of sequential decision making can be used to build a test with an adaptive stopping rule. This is most practical when all of the test's component parts are exchangeable. That is, that any piece of the test can be substituted for any other piece with no degradation of the test. It is also important that the test that any examinee is administered be comparable, in terms of difficulty and content, to the test that any other examinee received, **regardless of their respective lengths**. Obviously trying to develop an item pool in which all of the items are exchangeable is at least a very difficult task, and probably an impossible one. Testlets proved to be a useful tool in the successful completion of this ambitious application since, as we shall soon show, they can be constructed to be exchangeable.

There was an existing item pool of 110 items which had previously been administered to another sample of the examinee population in a paper and pencil format. Their responses to these items had previously been fit with a three parameter logistic model (3PL) and these estimated parameters were available (Kingston, 1987). The characteristic functions of these 110 items varied in difficulty and slope. They also fell into one of two content categories. Sixty percent of the items were Type I items, which dealt with physical and technical aspects of seismic knowledge. Forty percent were Type II items, which dealt with economic, legal and perceptual aspects of seismic knowledge.

It was decided that if the item pool was divided into ten-item testlets, in which each testlet was balanced for content and equal in average difficulty and discrimination, a simple item presentation algorithm would suffice to achieve the desired ends. Specifically, if such interchangeable testlets could be constructed **one could choose any testlet at random for presentation** to an examinee. After the completion of that testlet, a statistical determination could be made to pass the examinee, fail him, or present another testlet. Since all testlets would be content balanced, we could be sure that all examinees, regardless of the length of the test that they received, would have received a test of identically balanced content. Moreover, since all testlets were of the same average difficulty, we could be sure that no examinees were unfairly advantaged (of disadvantaged) with a too easy (or too difficult) exam. Last, since all testlets were of equal average discrimination, the precision of all tests were strictly proportional to their length (and not of their particular makeup). This latter aspect is of importance for some technical issues dealing with the calculation of the posterior expected loss. When testlets are constructed in this way, the *number right score* carries all of the information we need to be able to implement the Bayesian decision process that was employed in this application.

The testlets were constructed by cross classifying the item pool according to (a) Type designation, and (b) estimated item difficulty. Next the items were assigned sequentially to testlets. Each resulting testlet had six Type I items and four Type II items. The six testlets which appeared most "parallel" were then selected from the eleven available. The

obtained testlets were then examined by test developers who were experts in seismology. [1] The validity of the testlet interchangeability assumption was evaluated by determining the degree to which the six selected testlets varied with respect to the average likelihood of a particular number right score. Likelihoods were evaluated at five different points on the latent proficiency scale. These five points corresponded to important decision points surrounding the anticipated cutscore. This validity check showed that for examinees near the cutscore the average number right score had about **the same probability regardless of which testlet was administered.** This strongly supported the assumption of "parallel testlets" for examinees whose number right score was in the regions examined. That such parallel testlets, with balanced content, could be easily constructed provides us with a powerful and practical tool for building fair and balanced tests which have an adaptive stopping rule.

The NCARB test is being employed currently with a minimum test length of 20 items (2 testlets) and a maximum of 60 items (6 testlets). The testlet building blocks make it possible to easily utilize the powerful machinery of sequential probability tests that was originally developed by Abraham Wald (1947) for quality control problems. It allows us to maximize the probability of correctly classifying individuals while at the same time minimizing the amount of testing that must be done.

III. Reading Comprehension Example — Linear testlets, linearly administered

(Thissen, Steinberg & Mooney, 1989)

Reading comprehension items, formed by a single passage followed by a number of related questions, have long been a common and useful component of most verbal tests, yet their use has been restricted within IRT scored CATs. The reason for this is that it was (properly) recognized that the (sometimes) crucial assumption of conditional independence among items was very likely to be violated if several items share the same stem. There have been three different responses to this problem. One, decided upon by the development team for the CAT-ASVAB, (*Computerized Adaptive Test version of the Armed Services Vocational Aptitude Battery*) is to modify the items so that each passage is queried by only a single item. This solves the issue of loss of conditional independence, but at a considerable cost in efficiency — one only gets a single response's worth of information from the time taken to read an entire passage. The response to this was to make the passages shorter. This increased the efficiency, but changed somewhat the trait that was being measured by

---

[1] This individual examination of the algorithmically formed testlets revealed that the algorithm worked reasonably well, but was still not perfect. There were a few items paired (in the same testlet) that ought not to have been, and it was discovered that all items involving a graphic had been omitted. The flaws were corrected. The former problem was relieved by replacing offending items with others that had been matched for content area and difficulty. The latter problem was solved by inserting graphics items into testlets and removing a previously included item that matched the new one in terms of content and difficulty. This fiddling around with the testlet resulted in very little change in terms of the other criteria.

the test. When factor analyzed, this new kind of item loaded more heavily on word knowledge than had previous reading comprehension items. A second approach, quite commonly taken, is to ignore the interdependencies among the items associated with a single passage, fit a binary response model and hope that everything is alright. This approach tends to overestimate the information yielded by the item. The third approach, one which obviously we favor, is the testlet approach that we shall describe in this section. In it, we explicitly define the passage with its m associated questions as a single item; or, more precisely, a single testlet. In this formulation we will consider the examinees' responses to the m questions as a polychotomous response, and then score it either 0, 1, 2, ..., or m depending upon how many of the m questions the examinee gets correct.

Using this approach, we remove concerns about the lack of conditional independence among the questions associated with a single passage. The model we will use (developed by Bock, 1972) does not require conditional independence within testlets, only between them. This latter requirement involves independence between passages (after conditioning on examinee proficiency). Rosenbaum (1988) has shown that when conditional independence does not obtain, one can determine if the dependencies are just within certain item clusters, or endemic throughout. If only the former, such polychotomous response models as we shall describe, can fit the data.

Bock's Model

Suppose we have J testlets, indexed by j, where $j = 1, 2, ..., J$. On each testlet there are m_j questions, so that for the j th testlet there is the possibility for the polychotomous response $x_j = 0, 1, 2, ..., m_j$. The statistical testlet scoring model posits a single underlying (and unobserved) dimension which we call *latent proficiency*, and denote θ. The model then represents the probability of obtaining any particular score as a function of proficiency. For each testlet there is a set of functions, one for each response category. These functions are sometimes called item characteristic curves (Lord & Novick, 1968), item operating curves (Samejima, 1969) or trace lines (Thissen, Steinberg & Mooney, 1989). We shall follow Thissen *et al*'s notation and nomenclature.

The trace line for score $x = 0, 1, ..., m_j$, for testlet j is

$$T_{jx} = \frac{exp\,[a_{jk}\theta + c_{jk}]}{\sum_{k=0}^{m_j} exp\,[a_{jk}\theta + c_{jk}]}$$

where the $\{a_k, c_k\}_j$, $k = 0, 1, ..., m_j$ are the item category parameters, which characterize the shape of the individual response trace lines. The model is not fully identified, and so we need to impose some additional constraints. It is convenient to insist that the sum of each of the parameters equal zero, i.e.

$$\sum_{k=0}^{m_j} a_{jk} = \sum_{k=0}^{m_j} c_{jk} = 0$$

This model was fit to a 4-passage, 22 item test of reading comprehension by Thissen *et al*, in which there were 7 items associated with the first passage, 4 with the second, 3 with the third and 8 with the fourth [or, in the notation just introduced, $m_j = (7, 4, 3, 8)$]. They did this after they had performed an item factor analysis (Bock, Gibbons & Muraki, 1988) and found that a multifactor structure existed. The (at least) 4-factor structure found among these 22 items made the unidimensional (conditional independence) assumption of traditional IRT models untenable. After considering the test as four testlets and fitting Bock's nominal response model to the data generated by the almost 4,000 examinees they compared the results obtained with what would have been the case if they had ignored the lack of conditional independence and merely fit a standard IRT model. They found two things. First that there seemed to be a slightly greater validity of the testlet derived scores when correlated with an external criterion. Second, the test information function yielded by the traditional analysis was much too high. This was caused by this model's not being able to deal with the excess intra-passage correlations among the items (*excess* after conditioning on θ). The testlet approach thus provided a more accurate estimate of the accuracy of the assessment.

IV. Algebra Test — Hierarchical testlet, linearly administered

Mathematical knowledge has a partially hierarchical structure which lends itself to the construction of hierarchical testlets, which are, by their very nature, adaptive. Two testlets related to elementary algebra were prepared by ETS Test Development staff members James Braswell, Jeanne Elbich and Jane Kupin. In Figure 1 is a sample testlet which might form part of a larger test consisting of a number of such testlets, each of which covers a different topic in a broader mathematical unit. These testlets would be administered linearly, that is, each student would respond to items in *all* testlets, so that balancing of content occurs *between* testlets. A given testlet could focus on content related to what are felt to be important topics in the subject.

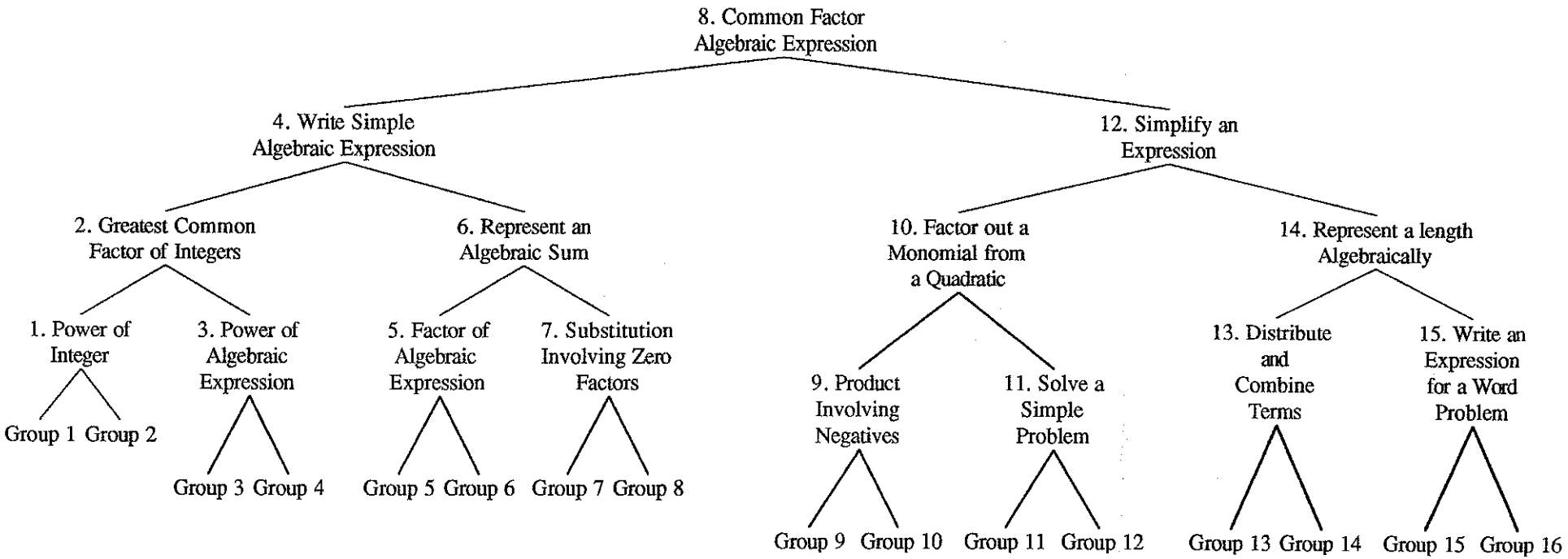
The idea of the hierarchical testlet is that the student is routed through the items according to his or her performance. After a correct answer, an item addressing a more difficult concept is presented. After an incorrect answer, an item testing a less difficult concept is given.

Insert Figure 1 About Here

As an illustration, consider a student who answers the first item in Figure 1 (*identifying the greatest common factor in two algebraic expressions)* incorrectly. This student would then be asked to perform the theoretically simpler task of writing a simple algebraic expression. Suppose the student is able to answer this item correctly. In this case the student would next be asked to represent an algebraic sum. A correct answer here, and the fourth item presented to the student would require a substitution involving zero factors. The actual items that this hypothetical student would receive are presented in Figure 2.

Figure 1. Elementary Algebraic Expressions Testlet - 15 Questions

*The **left** path from a node always indicates an **incorrect** response to the question represented by the topic shown at that node.*



Insert Figure 2 About Here

At the end of the sequence, students have been grouped into 16 theoretically ordered levels, based on the patterns of their responses. Conceptually, such a test is very appealing, since, under certain conditions, it provides the same resolution among the students taking the test as does the "number correct score" of a 15 item test, with each student taking only four items. It does, however present some difficult questions about how to model responses and estimate proficiency. Additionally, while the resolution of the test is the same as that of a 15 item test, its precision is not.

Although these questions could be addressed in the context of IRT (see Wainer & Kiely, p. 195), we will consider an alternative approach, namely *Validity-Based Scoring* (Lewis, in preparation). This approach is based on the assumption that it is possible to obtain information on some criterion measure (or measures), at least for a calibration sample of students. In the present context, this information might consist of scores on a battery of longer tests of various algebraic skills. As with any consideration of predictive validity, the choice of criterion is both crucial and difficult. *Validity-Based Scoring* forces the test developer to make this choice explicit, rather than avoiding it.

The grouping of students based on their responses to the items in the testlet may be expressed using indicator (0,1) variables, one for each group. These indicator variables are then used as predictors for the criterion measures in the calibration sample of students. *Validity-Based Scoring* assigns the predicted values on the criterion as the scores for each possible outcome for the testlet. These scores are simply the mean criterion values for the group of students with each given testlet result. The group standard deviations on the criterion variables may be interpreted as conditional standard errors of prediction for these scores.

Insert Figure 3 About Here

Figure 3 illustrates this idea for the algebra testlet of figure 1. Imagine a calibration sample of students, all of whom had responded to the algebra testlet, and for all of whom a criterion score (denoted by C) was available. The group of students (Group 1 in Figure 3) who incorrectly answered all questions presented to them (items 1, 2, 4, and 8) have a mean score on the criterion which is denoted by $\overline{C_1}$. This is the score which will subsequently be given to any student with this response pattern. Similarly, mean criterion scores for each of the remaining 15 groups in the calibration sample ($\overline{C_2}$ through $\overline{C_{16}}$) will be used as scores for students with any of the remaining response patterns. The standard deviation

Figure 2. A sample sequence of items (and responses) from the testlet represented in Figure 1

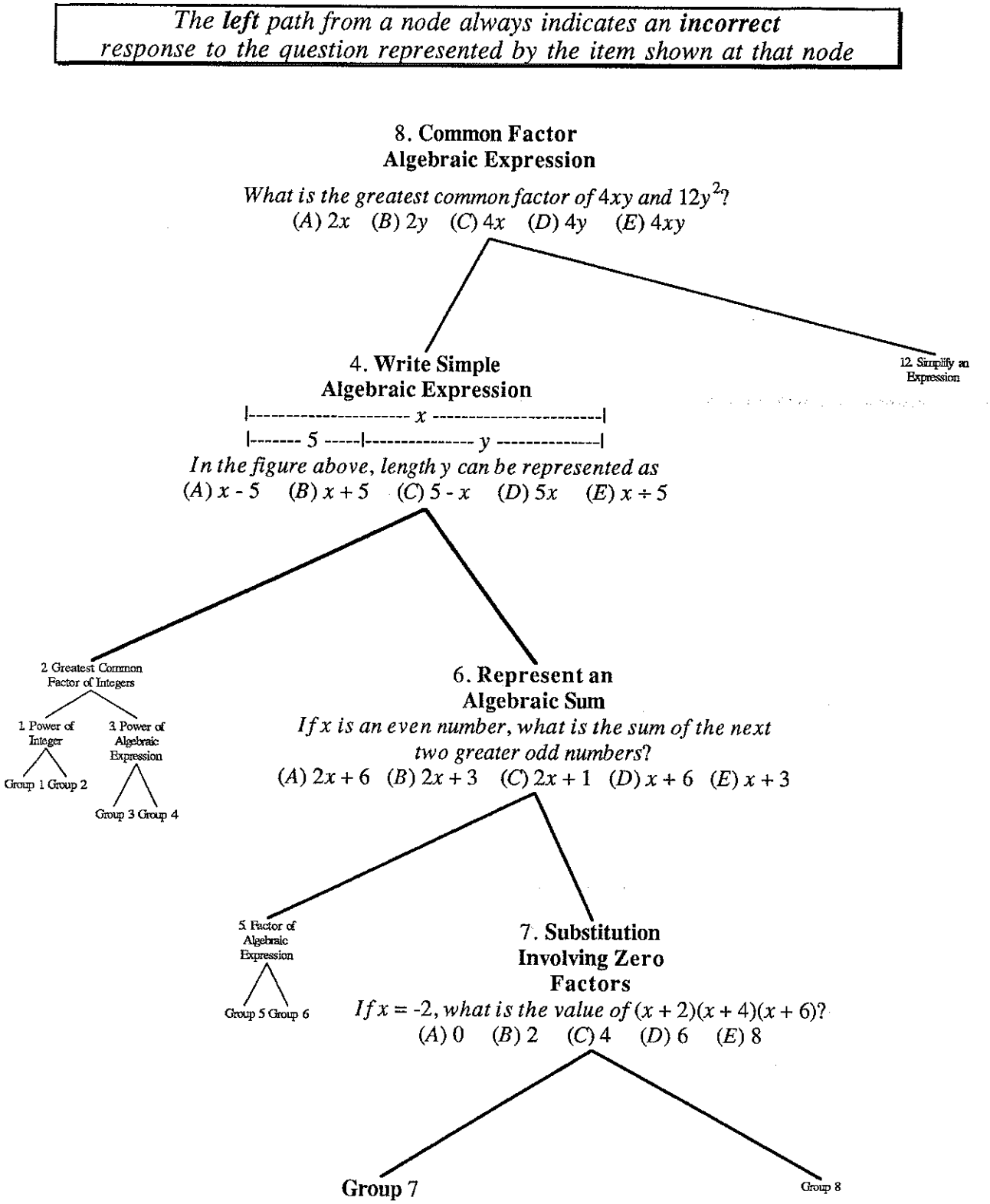


Figure 3. Response patterns, the item responses yielding them, and their associated criterion scores

Group	*Pattern*	*Item* 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15	*Criterion Score*
1	0000	0 0 0 0	C_1
2	0001	1 0 0 0	C_2
3	0010	1 0 0 0	C_3
4	0011	1 1 0 0	C_4
5	0100	1 0 0 0	C_5
6	0101	1 1 0 0	C_6
7	0110	1 1 0 0	C_7
8	0111	1 1 1 0	C_8
9	1000	1 0 0 0	C_9
10	1001	1 1 0 0	C_{10}
11	1010	1 1 0 0	C_{11}
12	1011	1 1 1 0	C_{12}
13	1100	1 1 0 0	C_{13}
14	1101	1 1 1 0	C_{14}
15	1110	1 1 1 0	C_{15}
16	1111	1 1 1 1	C_{16}

of the observed C_i , *within each group,* serves as the standard error of prediction for the score.

It should be noted that *Validity-Based Scoring* provides a direct check on the theoretical ordering of the response groups. If the scores for the groups do not reflect their ordering, or if differences between scores for adjacent groups are small relative to the standard errors, follow-up diagnostics should be explored. For example, a person who correctly guesses the answer to the first item cannot finish lower than Group 9. If this happens too often, the mean criterion score for Groups 9 and above will be depressed. This will result in a reduction of the differences among the group means relative to the within group variation. Thus one possible diagnostic procedure is to apply *Validity-Based Scoring* at each hierarchical level of the testlet, and see at what point the expected ordering begins to break down. This would then provide aid in the redefining of the hierarchical structure (and the associated items) that characterize both the theoretical structure of the testlet and its operational realization.

It may be useful at this point to briefly compare *Validity-Based Scoring* and Item Response Theory, considered as alternative approaches to scoring and assessing the precision of hierarchical testlet results. The advantage of IRT is that it provides a formal theoretical framework for modelling individual item responses and allows scoring and assessment of precision within that framework. Its disadvantages are that it makes strong assumptions about the responses (such as conditional independence) and that it provides no reference to the external validity of the results. *Validity-Based Scoring,* on the other hand, has no formal theoretical basis and mat best be characterized as being empirically oriented. Its strength is that it gives information which is directly relevant to test use, in the sense of prediction of a relevant criterion (assuming one can be identified). IRT results must be validated against such a criterion as a separate (and often neglected) step in test development.

V. Summary and Conclusions

Testlets, as the name implies, are small tests. We first proposed them as convenient units from which to construct a test. They are small enough to manipulate but big enough to carry their own context. They can be used to guarantee content balance to an algorithmically constructed test by being sure that each testlet is balanced (we refer to this as "within testlet balancing" as in the NCARB example), or by letting each testlet span one aspect of the test specifications, so that the test contents are balanced by judicious choice of testlets. We refer to this as "between testlet balancing" (as in the Algebra test example).

In this paper we have described three different ways that testlets can be used to improve the quality of tests and the flexibility of their mode of construction. While the original formulation of the testlet was phrased within the context of adaptive testing, we have shown that they can be useful in other situations as well. A testlet formulation provides us with a more accurate estimate of test quality in the paragraph comprehension example, it allows us to use the powerful statistical machinery of sequential decision making in the NCARB example, and it provides us with a much more efficient test in the Algebra exam-

ple — without the risks associated with putting the entire task of test construction into the hands of an imperfect algorithm.

The three examples we elaborated here were chosen for several reasons. One of these is that they illustrate the role that item response theory (IRT) plays in testlet construction and use. Specifically, IRT and testlets are two notions that are somewhat independent. One can use the testlet approach, even in an adaptive mode without recourse to IRT at all (the Algebra test). Or one can tie the testlet's construction and scoring intimately to IRT (the paragraph comprehension test). Or, it can be in between, where IRT is used to construct the testlets, but is then not used in the scoring (the NCARB example).

Obviously, there are many IRT models that can appropriately be utilized for the representation of testlet difficulty and for testlet scoring. The paragraph comprehension example only touches the surface of such models. It is conceivable that more information might have been gleaned from the NCARB test if the full response pattern had be used rather than merely the number right within testlet. Such information could have been got with the fuller use of IRT, but toward what end? As currently configured, the NCARB test is as efficient and accurate as required [2].

In addition there are other modes of testlet construction and combination that we have not touched on. One of the most obvious ones would be testlets (either hierarchically or linearly formed) combined hierarchically. We leave a description of the scoring of such tests to a later account.

We would like to emphasize one aspect of the Algebra test's scoring scheme; the way that the scoring of the testlet is integrated fully with a validity criterion. We feel that this is very important, for we feel that a test's validity is its most important characteristic and that too often validity studies are just tagged on at the end. If this scheme is used, one must confront validity right from the start, as an integral part of the test. If this was always done, we would face far fewer brouhahas about the legitimacy of test usage.

---

[2] Indeed, when this test was first developed it was shown that for some individuals one testlet (or even none!) was sometimes sufficient to reach a decision with the desired accuracy, however the members of NCARB felt that the minimum test length should be at least 2 testlets (20 items), so that examinees would believe that they were getting fair value for their testing fee.

VI. References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika,* **37**, 29-51.

Bock, R. D., Gibbons, R. & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement,* **12,** 261-280.

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. (1987). *Feasibility of using IRT-Based methods for Divisions D, E and I of the Architect Registration Examination.* Report prepared for the National Council of Architectural Registration Boards. Princeton, N.J.: Educational Testing Service.

Lewis, C. (in preparation) *Validity-Based Scoring.* Manuscript in preparation, Princeton, N.J.: Educational Testing Service.

Lewis, C. & Sheehan, K. (1988). *Using Bayesian decision theory to design a computerized mastery test.* Unpublished manuscript, Princeton, N.J.: Educational Testing Service.

Lord, F. M. & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison Wesley.

MacNicol, K. (1956). *Effects of varying order of item difficulty in an unspeeded verbal test.* Unpublished manuscript. Princeton, N.J.: Educational Testing Service.

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.

Rosenbaum, P. R. (1988) A note on item bundles. *Psychometrika*, **53,** 349 -360.

Sax, G. & Carr, A. (1962). An investigation of response sets on altered parallel forms. *Educational and Psychological Measurement,* **22,** 371-376.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph,* No. 17.

Thissen, D., Steinberg, L. & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement,* **26,** xxx-xxx.

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. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement,* **24,** 185-201.

Wald, A. (1947). *Sequential Analysis*, New York: John Wiley & Son .