DOCUMENT RESUME

ED 392 813 TM 024 451

AUTHOR	Wainer, Howard; And Others
TITLE	Some Empirical Guidelines for Building Testlets. Program Statistics Research Technical Report No. 91-14.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-91-56
PUB DATE	Oct 91
NOTE	14p.
PUB TYPE	Guides - Non-Classroom Use (055) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; Computer Simulation; *Item Banks; Test Construction; Test Format; Test Items; Test Length; *Test Validity
IDENTIFIERS	*Testlets

ABSTRACT

A series of computer simulations was run to measure the relationship between testlet validity and the factors of item pool size and testlet length for both adaptive and linearly constructed testlets. Results confirmed the generality of earlier empirical findings of H. Wainer and others (1991) that making a testlet adaptive yields only marginal increases in aggregate validity because of the peakedness of the typical proficiency distribution. Findings suggest that if a linear test is constructed from a much larger calibrated item pool, it can compare quite favorably to an adaptive test. The larger the selection of items from which the test is built, the better the final result. (Contains one figure, four tables, and seven references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

□ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

RR-91-56

ED 392 813

Some Empirical Guidelines for Building Testlets

Howard Wainer
Bruce Kaplan
Charles Lewis
Educational Testing Service

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 91-14

Educational Testing Service
Princeton, New Jersey 08541

TM024457

BEST COPY AVAILABLE

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

Some Empirical Guidelines for Building Testlets

Howard Wainer
Bruce Kaplan
Charles Lewis
Educational Testing Service

Program Statistics Research
Technical Report No. 91-14

Research Report No. 91-56

Educational Testing Service
Princeton, New Jersey 08541

October 1991

Copyright © 1991 by Educational Testing Service. All rights reserved.

Some empirical guidelines for building testlets[1]

Howard Wainer Bruce Kaplan Charles Lewis
Educational Testing Service

Abstract

A series of computer simulations were run to measure the relationship between testlet validity and the factors of item pool size and testlet length for both adaptive and linearly constructed testlets. We confirmed the generality of earlier empirical findings (Wainer, Lewis, Kaplan & Braswell , 1991) that making a testlet adaptive yields only marginal increases in aggregate validity because of the peakedness of the typical proficiency distribution.

[1] This research was supported by the Educational Testing Service's *Program Research Planning Council*; we are grateful for the help that this provided. We are also grateful for the comments and criticism we received on an earlier draft from Jim Braswell and Bill Ward.

Some empirical guidelines for building testlets

1. Introduction

There is an increasing awareness that tests should be built out of units larger than an item. The term *testlet* was explicitly introduced by Wainer & Kiely (1987, p. 190) to characterize "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." The use of the testlet as the unit of construction and analysis for computerized adaptive tests was proposed 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. In addition, a testlet can provide explicitly a coherent measure of a larger set of skills than would ordinarily be possible with a single item. It can also allow the test builder to provide some guides through a complex problem by suggesting, through the judicious use of subproblems, a path toward the solution of a larger question. This can provide both instructional help and an explicit framework for awarding partial credit.

In earlier work we described a testlet-based algebra exam (Wainer & Lewis, 1990; Wainer, Lewis, Kaplan & Braswell, 1991) and compared the efficacy of a linear 4-item testlet with a hierarchically constructed testlet. The former is analogous to a fixed format test, the latter to an adaptive test. Our results suggested only a marginal gain in validity of the hierarchically constructed testlet, but were of uncertain generality.

What would have happened if the item pool was larger? Would the broader choice of items allow the item selection algorithm to do much better? Would the adaptive testlet profit more from the increased choice? If so, how much more?

How much precision is gained if the testlets were longer? Is the gain greater for hierarchical testlets than for linear ones? If so, how much more? Does the Spearman-Brown prophesy apply? To what extent are any gains due to capitalization on chance?

To clarify our intuition on these issues we embarked on a series of simulations. This paper is an account of those simulations.

The basic idea of the simulations was to generate 'item pools' of four different sizes (15, 30, 50 and 100 items) and then construct testlets of five different lengths (4, 5, 6, 7, and 8) from those item pools. We then generated item response vectors for a substantial sample (1,000) of simulees and, from these 'data' choose items from the pools to build a testlet such that the performance of our simulees on it would yield estimates of proficiency that correlated maximally with the proficiency parameter (θ) that generated each response pattern. We built two different kinds of testlets:

(i) linear testlets, analogous to a traditional fixed length test, in which all simulees were confronted with precisely the same items, and

(ii) hierarchical testlets, analogous to an adaptive test, in which items are chosen for presentation to each simulee on the basis of performance on prior items.

The measure we used as an objective function is the squared correlation ratio η^2 (Pearson, 1905; Hays, 1973, p. 683), relating the groups into which the testlet divided simulees to the proficiencies for those simulees. After 'items' were chosen, trees built, and correlations determined, we took those testlets and calculated the value of η^2 on a neutral sample of simulees. We report both the initial (exploratory) results and the validation (confirmatory) results.

This study parallels quite precisely what we did earlier, except we are substituting known values of proficiency (θ) for estimated ones, and carefully determined item parameters for estimated parameters from real items. Since our findings, under the conditions of the earlier study, match our earlier empirical results, we are confident that the results in other situations are credible.

2. Procedure

2.1. Generating the Simulated Data

Two samples of 1,000 observations were generated for each of four item pools which consisted of 15, 30, 50 and 100 items respectively.

To generate an item pool of size M, we used a three parameter logistic item response model (3-PL) with $a = 1$, $c = .2$, and the b's evenly spaced from -3 to 3. The value for any given item difficulty is therefore given by

$$b_i = -3 + (i - 1) \times [\ 6 / (M - 1)], \text{ for } i = 1, \ldots, M.$$

Next we generated 2,000 proficiency values from a random normal distribution with mean 0 and variance 1. For each proficiency we calculated p_{ij}, the probability of person j with proficiency θ_j getting item i correct using the 3-PL model (shown in Equation 1).

$$P(\theta) = c + \frac{1 - c}{1 + e^{-a(\theta - b)}} \qquad (1)$$

And then we drew a random number from a Uniform [0,1] distribution. If it was greater than or equal to p_{ij} then item i was scored "correct" for person j, otherwise it was scored as wrong. This created the item response pattern for the 2,000 simulees. Half of them were set aside to form the confirmatory sample and the other half formed the exploratory sample.

2.2. Best Test of Fixed Length *m*

Using the exploratory sample for each of the four item pools, the best tests of lengths 4, 5, 6, 7 and 8 were determined. The criterion for the best test was the highest η^2 calculated as follows:

1. For a given set of *m* binary items the data were divided into the 2^m possible partitions. For example for a test of fixed length 4, the $2^m = 16$ combinations range from 0 (0000 in binary) to 15 (1111 in binary). 0000 corresponds to getting none of the four items right, 1111 corresponds to getting all four items correct.

2. For each of the 2^m partitions the variance of the θs is computed and then multiplied by (n_k - 1), the number of people in that partition minus one. This is the within cell error sum of squares and is added across all 2^m cells to form the pooled within cell sum of squares (SSE). η^2 is then defined as :

$$\eta^2 = 1\text{-(SSE/SST)},$$

where SST is (N-1) × Variance of θ for the entire sample. N is the total sample size, 1,000 in our case.

The number of sets of *m* items is the number of ways that *m* items can be drawn from an item pool of *M* items. This simple combinatorial problem is usually stated as $\binom{M}{m}$, which is defined as $M!/[(M\text{-}m)!\ m!]$ For our smallest case there are $\binom{15}{4}$ = 1,365 possible combinations of items, all of which need to be computed to find that combination with the largest η^2. For small numbers of combinations we checked all combinations, but for many situations $\left(\binom{100}{8}\right.$ equals over 186 billion combinations$\left.\right)$ there are obviously too many to enumerate all of them. But using some knowledge of how the item pool was constructed, as well as a little experimentation, taught us that it is sufficient to compute η^2 for $\binom{m+10}{m}$ combinations, taking the *m* middle items and 5 items before and 5 items after those *m* middle items. This shrunk the $\binom{100}{8}$ to $\binom{18}{8}$ = 43,758 number of combinations to evaluate.

2.3. "Best" Tree of Length *m*

The competition for the best linear testlet is a hierarchically constructed testlet. This can be profitably thought of as a tree of length *m*. We constructed such trees on the exploratory sample, for each size item pool for lengths of 4, 5, 6, 7, and 8. The η^2 as described above, was used on the 2^m cells of this tree.

The best tree was formed by choosing, as the start, the item that yielded the minimum SSE for two groups (one group got the item correct and the other group got the item incorrect). Since SST is constant for a given sample of simulees, SSE is linearly related to η^2 and thus minimizing one is identical to maximizing the other. The second item on each branch was chosen as the one that, when combined with the first, minimized the SSE in the four groups thus formed. This was continued until a test of length *m* was formed. Items were allowed to be reused, so that an item may appear on more than one node of a given tree.

2.4. Confirmatory Analysis

Four cells in our design were used for a confirmatory analysis. We chose the largest (100) and smallest (15) item pools and the largest (8) and smallest (4) tree and test length to do our analysis. Using the items for the best test of lengths 4 and 8 and the best tree of lengths 4 and 8 established by in the exploratory sample, the η^2 values were calculated for the confirmatory sample to see the effect of the sample on the tree selection.

3. Results

3.1. Hierarchical testlets

In Table 1 is shown the mean value of η^2 for trees of five sizes and item pools of four sizes. We see that although in the exploratory sample there appears to be an increase in precision of measurement for the same length test as the size of the test pool increases, most of this increase is probably due to capitalization on chance, at least for shorter testlets. Note that a 15 item pool is plenty for building a four item testlet. Any increase in precision we see with pool size disappears in the confirmatory sample. This is not quite as true for longer testlets. We see that the extra freedom we have when we use a 100 item pool to build an 8 item testlet yields some increase in the size of η^2.

Intuition in interpreting our results is helped by remembering the structure of a hierarchical testlet. If the testlet length is four the total tree can consist of as many as 15 items; the apex of the tree is a single item, which branches to two items, thence to four and finally to eight. The sum of these is 15. Of course since an item can be reused on any branch, 15 is the upper bound. In general for a testlet length of *m* items the maximum number of items required is $2^m - 1$. Thus for an eight item testlet as many as 255 items might be required.

Table 1

Best Possible Tree

		Exploratory Sample				*Confirmatory Sample*	
Test		*Item Pool Size*				*Item Pool Size*	
Length	**15**	**30**	**50**	**100**	*Mean*	**15**	**100**
4	0.59	0.63	0.65	0.68	**0.63**	0.55	0.55
5	0.66	0.71	0.73	0.76	**0.72**		
6	0.72	0.78	0.80	0.84	**0.78**		
7	0.77	0.83	0.86	0.90	**0.84**		
8	0.80	0.88	0.91	0.94	**0.88**	0.70	0.75
Mean	**0.71**	**0.76**	**0.79**	**0.82**			

3.2. Linear testlets

In Table 2 are shown the analogous results for a fixed (linearly presented) testlet. There is a small gain in precision with an increasing size item pool . The confirmatory results indicate that there is smaller shrinkage on cross-validation when the testlet is drawn from a larger pool than from a smaller one. Gains in precision with test length do not match

ERIC
Full Text Provided by ERIC

what would be expected from Spearman-Brown, reinforcing the fact that this prophecy formula is inappropriate within this context.

Tahle 2

Best Possible Linear Testlet

	Exploratory Sample					*Confirmatory Sample*	
Test	*Item Pool Size*					*Item Pool Size*	
Length	**15**	**30**	**50**	**100**	*Mean*	**15**	**100**
4	0.55	0.58	0.57	0.57	**0.57**	0.49	0.55
5	0.61	0.63	0.62	0.62	**0.62**		
6	0.66	0.67	0.68	0.67	**0.67**		
7	0.70	0.72	0.73	0.72	**0.72**		
8	0.74	0.77	0.77	0.77	**0.76**	0.69	0.75
Mean	**0.65**	**0.67**	**0.67**	**0.67**			

Table 3 contains the differences between respective entries in Tables 1 and 2. What we see is that even though there is a substantial apparent advantage to making a testlet adaptive, this advantage practically disappears in the confirmatory sample. Table 4 has the same results displayed as percentage gains.

Table 3

Differences between hierarchical & linear testlets

	Exploratory Sample					*Confirmatory Sample*	
Test	*Item Pool Size*					*Item Pool Size*	
Length	**15**	**30**	**50**	**100**	*Mean*	**15**	**100**
4	0.04	0.05	0.08	0.12	**0.07**	0.05	0.00
5	0.05	0.09	0.10	0.14	**0.10**		
6	0.06	0.10	0.13	0.17	**0.12**		
7	0.07	0.11	0.14	0.18	**0.13**		
8	0.07	0.11	0.14	0.17	**0.12**	0.01	-0.01
Mean	**0.06**	**0.09**	**0.12**	**0.16**			

Table 4

Percentage advantage of hierarchical over linear

	Exploratory Sample					*Confirmatory Sample*	
Test	*Item Pool Size*					*Item Pool Size*	
Length	**15**	**30**	**50**	**100**	*Mean*	**15**	**100**
4	7%	8%	13%	20%	**12%**	11%	0%
5	8%	14%	16%	22%	**15%**		
6	9%	15%	19%	25%	**17%**		
7	9%	15%	19%	24%	**17%**		
8	9%	14%	17%	22%	**16%**	1%	-1%
Mean	**8%**	**13%**	**17%**	**23%**	**15%**		

4. Discussion and Conclusions

This set of simulations provides us with a confirmation that the empirical results we reported earlier (Wainer et al, 1991) on 4-item testlets drawn from 15 item pools generalize to broader circumstances. The performance of a carefully chosen linear test relative to an adaptive test was encouraging. One criticism of adaptive testing (Wainer et al, 1990) is that it requires the development and calibration of large item pools, whereas a linear test only requires the items that actually appear on the test. Our findings suggest that if a linear test is constructed from a much larger calibrated pool it can compare quite favorably to an adaptive test. But all bets are off if a 4 item test is built from a 4 item pool. Good items make for good tests, and the larger the selection of items from which we construct our tests the better will be the final result. Our simulations showed that with a high quality pool like this (remember all a's were one), a 15 item pool does rather well even for a testlet as long as 8 items. Certainly a 30 item pool would be ample if the distribution of item difficulty was peaked in the middle rather than being uniform as in this case. The other advantage of a linear format is that, because there is less chance to capitalize on chance, there is less shrinkage on cross-validation. The large size of the shrinkage for an adaptive format and consequently the small size of the cross-validated difference was a bit of a surprise for us.

The disappointing performance of the hierarchical testlets is principally due to the unconditional nature of the measure we chose to characterize performance. The linear test is centered in the middle of the proficiency distribution and so does very well where there are a lot of simulees. Its performance deteriorates on simulees in the tails of the distribution.

Shown in Figure 1 are typical 4-item testlets chosen from a 100 item pool: o..e linear and one hierarchical. Note how the items for the linear testlet are clustered near the center of the proficiency scale, whereas the items chosen for the tree are spread out further. This provides more accurate estimation in the tails. However since there are so few simulees in the tails relative to the middle, an aggregate statistic like η^2 is insensitive in the area of the hierarchical testlet's superiority. Had we used a measure of quality of performance that was conditional on θ, this superiority in the tails would have been visible. For a variety of reasons, discussed in our earlier paper (Wainer et al, 1991), we opted for this unconditioned measure. However we feel it is important to emphasize this important limit to the consequences of our findings. We are not suggesting that linear testlets are just as good as hierarchical ones; only that the advantage of adaptive testing is in the tails and is only required when accurate measurement is required throughout the proficiency range. For tests with cut-scores within the middle of the proficiency range there is little advantage in making the test adaptive.

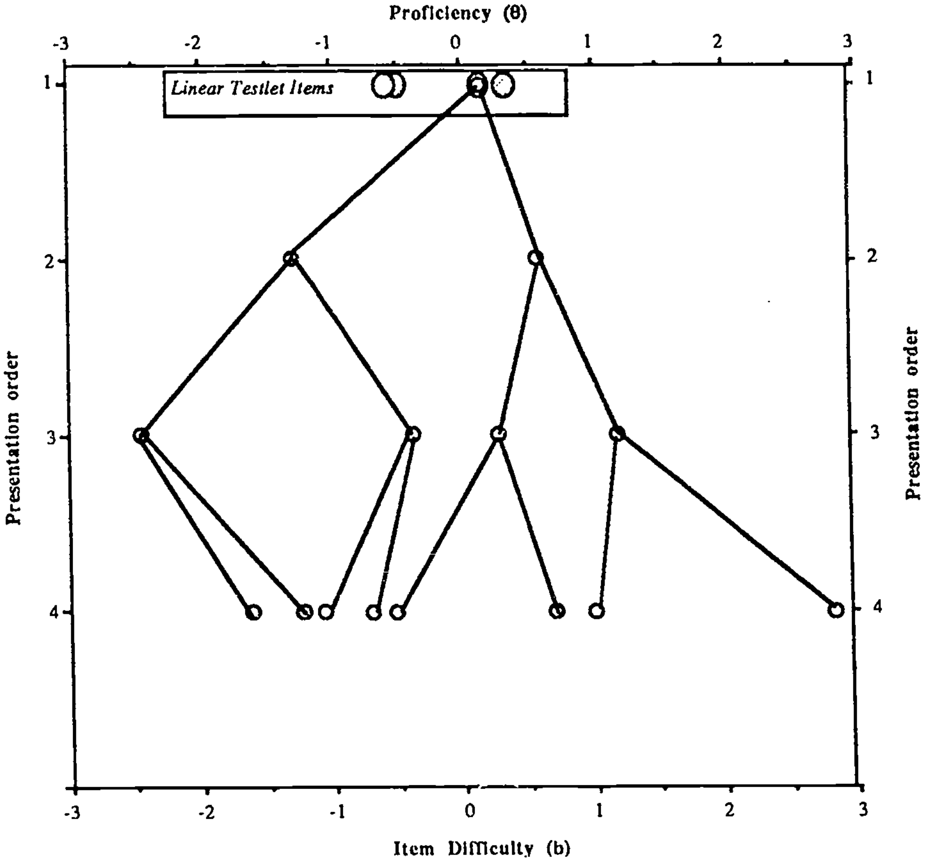

Figure 1: A 4-item hierarchical and linear testlet.

5. References

Hays, W. L. (1973). *Statistics for the social sciences* (2nd edition). New York: Holt, Rinehart and Winston.

Pearson, K. (1905). Mathematical contributions to the theory of evolution, XIV: on the general theory of skew correlation and non-linear regression.*Draper's Company Research Memoirs,* biometric series 2. Reprinted in Pearson, 1956, pp. 477-528.

Pearson, K. (1956). *Karl Pearson's Early Statistical Papers.* Cambridge: Cambridge University Press (first issued, 1948).

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

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1-14.

Wainer, H., Dorans, N., Green, B., Mislevy, R., Steinberg, L., & Thissen, D. (1990). "Future challenges." In H. Wainer, et al, *Computerized Adaptive Testing: A Primer.* Hillsdale, NJ: Lawrence Erlbaum Associates, Chapter 9, pps. 233-272.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1991). Building algebra testlets: A comparison of hierarchical and linear structures. *Journal of Educational Measurement, 28,* xxx-xxx.

ERIC Full Text Provided by ERIC