Psychometric Properties of Finite-State Scores Versus Number-Correct and Formula Scores: A Simulation Study

Miguel A. García-Pérez
Universidad Complutense de Madrid

Robert B. Frary
Virginia Polytechnic Institute and State University

As developed by García-Pérez (1987), finite-state scores are nonlinear transformations of the proportions of conventional multiple-choice responses that are correct, incorrect, and omitted. They estimate the proportions of item alternatives which the examinees had the knowledge needed to classify (as correct or incorrect) before seeing them together in the items. The present study used simulation techniques to generate conventional test responses and to track the proportions of alternatives the examinees could classify independently before taking the test and the proportions they could classify after taking the test. Then the finite-state scores were computed and compared with these actual values and with number-correct and formula scores based on the conventional responses. Highly favorable results were obtained leading to recommendations for the use of finite-state scores. These results were almost the same when the simulation proceeded according to the model and when it was based on a naturalistic process completely independent of the model. Hence the scoring procedures on which finite-state scores are based are both accurate and robust. *Index terms: applied measurement models, examinee behavior, finite-state scores, guessing, multiple-choice tests, test scoring.*

Many psychological constructs, such as knowledge or educational achievement, can be measured on interval or even ratio scales. Yet, for a typical multiple-choice test, number-correct scores cannot reasonably be claimed to provide direct estimates of the amounts of knowledge that examinees have. With rare exceptions, this scoring method provides only ranking information about the examinees, and important information is lost when it is used. This deficiency arises in part because it is not possible to establish a homomorphism that maps levels of knowledge into the set of integers constituting the number-correct scores (see Krantz, Luce, Suppes, & Tversky, 1971, pp. 8–9). Conventional formula scoring (correction for guessing) was initially invoked with the goal of achieving such a homomorphism, namely, mapping the numbers of items known (without guessing) into the formula scores. Of course, this outcome is possible only if, for every item, an examinee either knows the answer or is reduced to omitting it or guessing at random among all of the choices. As a result, in practice, formula scores provide only the same (ranking) information about the examinees as number-correct scores (Lord, 1975).

This unsatisfactory state of affairs has led to a great deal of research to develop response/scoring systems for multiple-choice tests that eliminate or control guessing, that produce scores on interval or ratio scales, or that provide better reliability and validity than number-correct or formula scores. An

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 4, December 1989, pp. 403–417
© Copyright 1989 Applied Psychological Measurement Inc.
0146-6216/89/040403-15$2.00

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

extensive review by Frary (1989), covering methods that attempt to evaluate partial information at the item level, concluded that none of the methods was uniformly desirable or efficacious with respect to examinee acceptance, resource usage, and psychometric properties of the scores.

Item response theory (IRT), on the other hand, is said to produce ability estimates on an interval scale. However, for multiple-choice testing, this benefit of IRT is likely to be attainable only for testing programs involving rather large numbers of examinees. Furthermore, when IRT methods providing maximum likelihood ability estimates are used, certain incorrect answers can yield higher ability estimates than correct responses to the same items, which, to say the least, should be very difficult to explain to examinees (Thissen & Steinberg, 1984). Such an outcome is, at least indirectly, the result of failure to adhere to the advice of Molenaar (1981), who, commenting on mathematical models for achievement testing, recommended that ''not only mathematics and statistics but also psychology (in the form of cognitive models and empirical data) should preferably play a dominant role'' (p. 228).

The following section reviews an approach to producing scoring models (García-Pérez, 1987) that fills some of the gaps just outlined. The resulting scoring models may be adapted to any assumptions about the items and the guessing behavior of the examinees that may hold for a given situation, and they yield *finite-state scores*. These scores have a psychologically sound interpretation: Under the conditions stated for each scoring model, a finite-state score estimates the proportion of all the options on the test that an examinee could correctly classify as correct versus incorrect responses to the item stems, without seeing them together in the items. If the test consists of a random sample of all available items, then a finite-state score will also estimate the proportion of options known in the population of all available options. The process of determining a finite-state score takes item characteristics and examinee behavior into consideration and results in an estimate of the examinee's knowledge as it was *before* taking the test. During this time, learning is possible (e.g., learning the answer to an item by recognizing all of the distractors).

The goal of the study was to evaluate the psychometric properties of finite-state scores in comparison with number-correct and formula scores. Examinee responses were simulated using two different approaches. Simulation was a virtual necessity, because attaining the goals of the study required knowing each examinee's overall ability and state of knowledge with respect to each option before seeing them together in the items.

Finite-State Models and Finite-State Scores

Hutchinson (1982) described a class of theories of performance on multiple-choice tests, which he called finite-state theories. These theories assume that there are only two possible states of an examinee with regard to an item: total ignorance or total knowledge. As developed by Hutchinson, finite-state theories cannot account for partial knowledge; for this reason he favored his continuous-distribution theories. However, the continuous-distribution theories do not account for whether an examinee guesses when the answer is not known. This discrepancy led García-Pérez (1987) to propose finite-state theories that account for guessing and incorporate partial knowledge. Common to García-Pérez's finite-state theories is the assumption that, for each option of an item, the examinee has or does not have the knowledge needed to classify the option as correct or incorrect *without having read the other options*.

A detailed presentation of the theory can be found in García-Pérez (1987), but a different mode of presentation is used below. First, it is shown how quantitative operationalization of knowledge and an assumption about guessing behavior can be combined to produce a particular realization of finite-state theory. Then the scoring model that is appropriate for use in such circumstances is derived. Finally, variations of the theory and the scoring model that result from considering different guessing behaviors are discussed.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Examinee Characteristics and Behavior

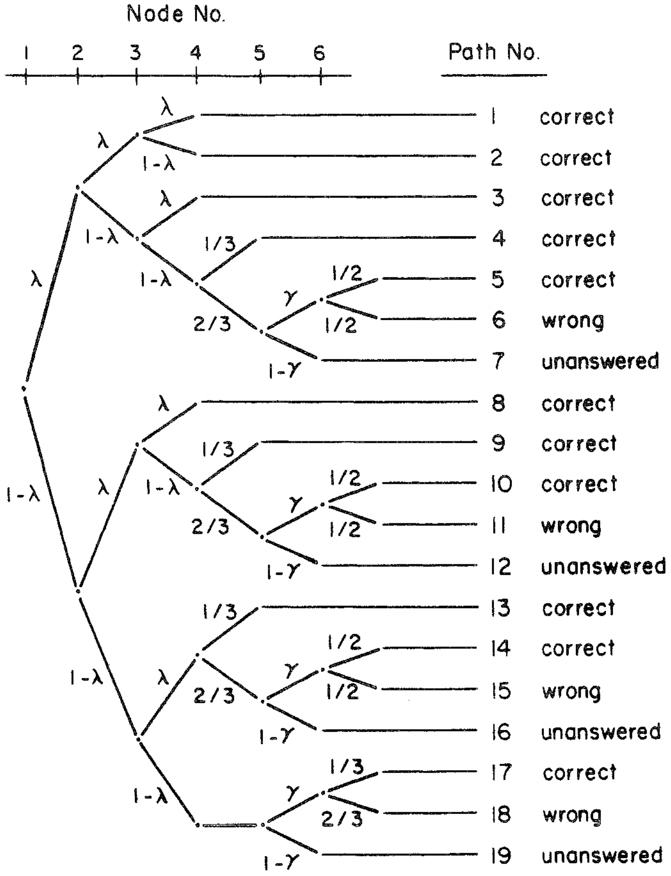
For an N-option item, there are $N + 1$ possible states of knowledge, ranging from total ignorance (when no option can be classified) to total knowledge (when all options can be classified). Also, there are $N - 1$ states of partial knowledge corresponding to having or not having the knowledge needed to classify 1, 2, ..., $N - 1$ options, regardless of whether the (single) correct choice (the answer) is among them.

An examinee is assumed to know how to classify a proportion λ ($0 \leq \lambda \leq 1$) of the options (whether answers or distractors) of all the items appropriate for testing the subject matter of interest. Therefore, λ represents the examinee's ability, and (assuming that the test is a random sample of all the appropriate items) it also denotes the probability that he/she will have the knowledge to classify any option (answer or distractor) included in the test. If the examinee's knowledge over all the options of any item does not result in assured identification of the correct answer (disregarding the possibility of misinformation), it is assumed that there is a probability γ ($0 \leq \gamma \leq 1$) that he/she will guess at random among the unclassified choices.

Figure 1 shows how λ and γ interact to determine whether an examinee answers correctly or incorrectly or omits an item. The tree diagram represents all possible sequences of events when an examinee confronts

Figure 1

Tree Diagram for a Three-Choice Item Assuming Random Omission Behavior



Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

a three-choice item under instructions to mark the choice believed to be correct. The first, second, and third nodes (left to right) represent independent attempts at classifying each option as correct or incorrect. The fourth nodes, in the paths where they appear, represent whether the option that is the answer to the item is among those correctly classified. The fifth nodes represent decisions to guess or not guess, and the sixth nodes represent guesses when they are made.

The first path (from top to bottom) represents total knowledge (knowledge to classify all three choices) and results in a correct answer, as indicated to the right of that path. The highest degree of partial knowledge (when two of the three choices are classified) is represented by paths 2, 3, and 8 and always results in a correct answer. The next and lowest degree of partial knowledge (one choice classified) is represented by paths 4–7, 9–12, and 13–16.

In this case, a correct answer may be given either when the only classified choice is the correct one (paths 4, 9, and 13) or when guessing occurs after elimination of one distractor (paths 5, 10, and 14). Guessing may also result in an incorrect answer (paths 6, 11, and 15). Finally, the examinee with the lowest level of partial knowledge may omit (paths 7, 12, and 16). Total ignorance is represented by paths 17–19, with the possible outcomes of guessing and succeeding (path 17), guessing and failing (path 18), and omitting (path 19).

Finite-State Scoring Formulas

From Figure 1, therefore, the probability c of correctly responding to an item is

$$c = \lambda^3 + 3\lambda^2(1-\lambda) + \lambda(1-\lambda)^2 + \lambda(1-\lambda)^2\gamma + \frac{(1-\lambda)^3\gamma}{3} \quad . \qquad (1)$$

The probability w of an incorrect response is

$$w = \lambda(1-\lambda)^2\gamma + \frac{2(1-\lambda)^3\gamma}{3} \quad . \qquad (2)$$

The probability u of omitting the item is

$$u = 2\lambda(1-\lambda)^2(1-\gamma) + (1-\lambda)^3(1-\gamma) \quad . \qquad (3)$$

Note that the five addends in the right side of Equation 1 respectively indicate that the correct answer may be given on the basis of total knowledge, a high level of partial information, a low level of partial information that includes knowledge of the answer, a low level of partial information with a successful guess in the absence of knowledge of the answer, and total ignorance with a successful guess. Because estimating λ is of interest, Equations 1 and 2 can be combined to yield

$$\lambda^4 + \lambda^3 - 3\lambda^2 + (c - 2w - 2)\lambda + (2c - w) = 0 \quad , \qquad (4)$$

which represents the nonlinear scoring formula derived from the model for the situation described. In practice, the proportion of correct and incorrect responses for a given examinee would be taken as estimates of c and w. Then the value of λ satisfying Equation 4 for these values would be the examinee's finite-state score. It can easily be proven that the scoring polynomials have exactly one root in the interval [0,1]. Numerical methods are effective for finding that root.

Additional Behaviors and Scoring Formulas

Although Equation 4 is appropriate for the situation just described (namely, a test with three-choice items and examinee behavior determined according to the tree diagram in Figure 1), there are obviously any number of behaviors examinees might adopt regardless of the number of choices. To some extent

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

these behaviors may be the result of personal idiosyncrasies, but for the most part they are probably influenced by the instructions for the test and the examinees' knowledge of the optimal guessing strategy. The behavior depicted in Figure 1 will be referred to as *random omission* (RO), and might be adopted by examinees who were not well motivated or did not understand the probable gain from guessing after elimination of incorrect choices even under formula scoring. RO behavior might also occur if the examinees were not given any particular instructions about guessing or failed to attend to them. In addition to resulting in inappropriate omissions, RO behavior could also result in guessing when formula-scoring instructions would advise against it (i.e., when an examinee is completely ignorant).

A second response behavior is one that might be induced by instructions to the effect that the test will be scored by number-correct. In the *number-correct* (NC) behavior, it is assumed that all examinees answer all items, guessing among the options they cannot classify when the answer is unknown. The tree diagram for this behavior is a simplified version of that in Figure 1, with $\gamma = 1$ everywhere. The probabilities of the response outcomes are then

$$c = \lambda^3 + 3\lambda^2(1-\lambda) + 2\lambda(1-\lambda)^2 + \frac{(1-\lambda)^3}{3} \tag{5}$$

and

$$w = \lambda(1-\lambda)^2 + \frac{2(1-\lambda)^3}{3} \quad , \tag{6}$$

and the scoring formula is

$$\lambda^3 - 3\lambda + (3c-1) = 0 \quad . \tag{7}$$

A third possible basis for responding will be referred to as the *formula scoring* (FS) behavior. This behavior is characterized by guessing at random among the unclassified options when at least one distractor has been identified, and leaving items unanswered in the case of total ignorance. As compared with Figure 1, the tree diagram for this behavior has $\gamma = 1$ everywhere except in the branching that leads to paths 17–19, where $\gamma = 0$. In this case,

$$c = \lambda^3 + 3\lambda^2(1-\lambda) + 2\lambda(1-\lambda)^2 \quad , \tag{8}$$

$$w = \lambda(1-\lambda)^2 \quad , \tag{9}$$

$$u = (1-\lambda)^3 \quad , \tag{10}$$

and the appropriate scoring formula is

$$\lambda^3 - 3\lambda + (2c-w) = 0 \quad . \tag{11}$$

Finite-state theory is not limited to producing the relatively simple models used to introduce it. As shown by García-Pérez (1987), it can be adapted to almost any testing format and response behavior, including subset selection methods (e.g., Coombs' method) and answer-until-correct. The present authors are aware of no other approach to developing scoring models for multiple-choice tests that can be tailored both to the response format and to variations in examinee behavior/strategy. Furthermore, resulting models have been tested, with some success, using data from a dual administration of a test using conventional and Coombs-type directions (García-Pérez, 1987) and from an answer-until-correct administration of another test (García-Pérez, in press).

Assumptions About Items

The derivations of the finite-state scoring polynomials just presented were based on two implicit assumptions that are made explicit here. The first is that each option is classified independently by an

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

examinee. As a result, a test to be scored using these finite-state polynomials should not have two options when one is simply the negation or opposite of the other, as are sometimes found together in the same poorly written item. Similarly, knowledge needed to classify a distractor should not be equivalent to the knowledge needed to recognize the answer. Well-written items should be free of these problems.

On the other hand, items asking questions such as ''Which of the following is the best example of . . . ?'' or items with choices such as ''Both b. and c. are correct'' would have to be avoided, because at least one of their options cannot be classified without having read the others. This point aside, a model could be tailored to accommodate these latter types of questions if an entire test consisted of a single type (e.g., a test in which all items offered the choice ''None of the above'').

The second assumption is that examinees have equal probability of having the knowledge to classify answers and distractors. As a result, when an examinee can classify only one option of a three-choice item, the model applies a probability of 1/3 that the classified option is the answer to the item. This characteristic weighs against using Equations 4, 7, and 11 for tests containing substantial numbers of easily eliminated distractors when the remaining options are relatively difficult to classify. Indeed, some multiple-choice tests contain absurd choices, which almost all examinees avoid.

While the deficient item characteristics just mentioned are inconsistent with use of Equations 4, 7, and 11, it should be noted that tree diagrams could be developed to accommodate them. For example, it could be assumed that if an examinee can classify only one option of an item, then it is a distractor. However, the present study will be limited to tests with items meeting the above assumptions.

Organization of the Study and Procedures

Scoring Formulas for Simulation Use

Scoring formulas for three-choice items were developed above for the sake of simplicity. A more realistic case for evaluation of finite-state scores is a test with four-choice items. Accordingly, the scoring polynomials for four-choice items corresponding to the three behaviors defined earlier are listed here. For RO behavior,

$$\lambda^6+\lambda^5+\lambda^4-4\lambda^3+(c-3w-2)\lambda^2+(2c-2w-3)\lambda+(3c-w) = 0 \quad . \tag{12}$$

For NC behavior,

$$\lambda^4-4\lambda+(4c-1) = 0 \quad . \tag{13}$$

For FS behavior,

$$\lambda^4-4\lambda+(3c-w) = 0 \quad . \tag{14}$$

The finite-state scores arising from use of these equations are respectively designated $\hat{\lambda}_{\mathrm{RO}}$, $\hat{\lambda}_{\mathrm{NC}}$, and $\hat{\lambda}_{\mathrm{FS}}$ to distinguish them from the true λs that they estimate.

Areas of Investigation

The evaluation of the psychometric properties of the finite-state scores arising from Equations 12 through 14 involved the following areas of investigation:

Ranking characteristics. Number-correct and formula scores provide essentially the same rankings as more complex response/scoring procedures such as subset selection, answer-until-correct, etc. Although the more complex procedures are impractical for many uses, they show promise for evaluating knowledge at the option level (Frary, 1982). Therefore, it was of interest to determine how finite-state rankings

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

compare with those from number-correct and formula scores. Of particular interest were the rank-order correlations among the finite-state scores and the number-correct and formula scores.

Psychological interpretability. Because, theoretically, finite-state scores are quantitative estimates of knowledge, it is important to determine the extent to which this characteristic holds in practice. Specifically, scores from Equations 12 through 14 were compared to what they are said to estimate: actual proportions of options which the simulated examinees had the knowledge to classify before seeing them together in the items.

In this regard, it was also of interest to compare the intercorrelations among the finite-state scores, the number-correct and formula scores, the true abilities of the examinees, and the observed numbers of options correctly classified before and after testing. Of particular interest was a comparison of finite-state score correlations with those from number-correct and formula scores with respect to true ability and the proportions of options correctly classified.

Effect of inappropriate behavior. A major concern regarding all response/scoring procedures is the effect of examinee failure to respond as the scoring model assumes. Therefore, the responses arising from each of the three examinee behaviors defined above were evaluated by each of the related finite-state scoring formulas to detect the effect of failure to follow directions.

Initial versus final knowledge level. A comparison was made between the number of options the examinees could classify correctly before taking the test and the number correctly classified upon completion of the test. Number-correct and formula scores were evaluated to determine their adequacy as estimators of either of these quantities.

Simulation Procedures

Approaches to simulation. The scoring models of this study are based on the probability λ that an examinee will be able to classify a randomly drawn option, with no specification regarding the difficulty of the option. However, a real test has a collection of options that vary in difficulty. An average examinee can almost certainly be expected to have the knowledge to classify some of the easy options, whereas it may be unlikely that the examinee will have the knowledge needed for the most difficult options.

This raises the question of the extent to which the practical implications of the model will be applicable to real testing situations. Two approaches to simulation were adopted to cope with this problem. In the "model-based" simulation, responses were generated using the functional relationships specified by the model. In the "naturalistic" simulation, responses were generated using an arbitrarily determined but plausible procedure that operated independently of the model and took option difficulties into consideration.

This distinction allowed investigation of the properties of finite-state scoring formulas in two different ways. The model-based simulation allowed conclusions about the accuracy with which scoring polynomials attain their goal when the assumptions from which they are derived hold. The naturalistic simulation permitted study of the robustness of these scoring models with respect to a violation of one of these assumptions—namely, variation in difficulty across items and options—which is likely to occur in a real testing situation.

Programs were written to simulate the responses of 300 examinees to a four-choice, 60-item test. IMSL (1987) subroutines GGNML and GGUBS were used to generate unit-normal deviates and pseudorandom numbers between 0 and 1 at various points in the programs. Responses were simulated under the assumption that the test was unspeeded, that is, all examinees had ample time to read and consider all items. Both the model-based simulation and the naturalistic simulation began by drawing samples of 300 unit-normal deviates to represent examinee ability levels. These were used differently in each simulation approach, as described below. Also, in both simulation approaches, a second sample of 300 unit-normal deviates,

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

uncorrelated with the first, was drawn to represent each examinee's guessing proclivity under the RO behavior. These were changed into probabilities of guessing in the absence of knowing the answer by using their mean and standard deviation to transform them to have a mean of .7 and a standard deviation of .1. Results above 1 or below 0 were truncated to 1 or 0. The standardization values were chosen because they produced reasonable numbers of omissions, approximately the same as occur under the FS behavior.

Model-based simulation. The mean and standard deviation of the 300 unit-normal ability values were used to transform them to have a mean of .5 and a standard deviation of .15, truncating values above 1 or below 0 if necessary. These were taken to represent the true λs of the sample. For each examinee, the program then proceeded through the nodes of the applicable model, once for each item. This was accomplished by drawing pseudorandom numbers between 0 and 1 to make decisions at every node.

This can be elucidated by considering the case of an examinee being processed under the behavior of the model in Figure 1 (RO behavior). Suppose that for this examinee $\lambda = .6$ and $\gamma = .8$. For the first item, the first node decision is made by drawing a pseudorandom number between 0 and 1; if this number is smaller than .6, the next node decision is at the highest second node, and so on through the model, repeating for each item. As this process continues, counts are kept of the number of options the examinee could classify correctly before and after seeing the test. (For example, if an examinee only classifies the correct choice on a four-choice item, then he/she will be able to classify all four choices after seeing them together in the item.)

The first of these two variables was designated ℓ_{PRE}, using ℓ to designate that it is the observed proportion of options classified rather than the true λ. The proportion of options correctly classifiable after taking the test was designated ℓ_{POS}, though, strictly speaking, its only relation to λ as defined earlier is that it is greater than or equal to this value. In addition, counts were kept of the number of correct and incorrect responses to the items. When all items were completed, the proportion correct and the proportion incorrect were computed and entered into Equations 12 through 14, which were then solved for the finite-state scores (λ estimates), namely $\hat{\lambda}_{\text{RO}}$, $\hat{\lambda}_{\text{NC}}$, and $\hat{\lambda}_{\text{FS}}$.

This process was repeated for the NC and FS behaviors, yielding six additional finite-state scores. Thus an examinee completed the simulation process with the true λ intact, nine finite-state scores (three for each behavior), ℓ_{PRE} (the proportion of options correctly classifiable before seeing the test, which is the same under each behavior), ℓ_{POS} (the proportion of options correctly classifiable after seeing the test, also the same under each behavior), and the score appropriate for the test (number-correct for responses arising from NC behavior, and formula scores for responses arising from FS or RO behavior). The decision to apply only formula scoring to the responses arising from FS or RO behavior was made because it seemed implausible to encounter fairly widespread omissions on a test scored number-correct and because inappropriate omissions may well occur when formula-scoring instructions are used.

Naturalistic simulation. This procedure was based on the intuitive notion that whether an examinee has the knowledge to classify an option depends on the examinee's ability and on the difficulty of the option. The initially drawn normal deviates representing ability were not modified but were taken directly as measures of ability, with 0 representing average ability. However, values outside the range of -3 to $+3$ were truncated to these values for a reason that will become apparent. To represent the difficulty of the options of the test, a vector of 240 unit-normal deviates was drawn (60 items × 4 choices). Based on the mean and standard deviation of this sample, it was standardized to have a mean of .5 and a standard deviation of .15, with truncation of values above .99 or below .01 if these occurred. These values were then taken to be the probabilities that an examinee of average ability would independently have the knowledge to classify each option as correct or incorrect.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Figure 2 defines the basis for determining the knowledge of a single examinee with respect to a single option. The two linear functions in Figure 2 meet at the point $(0,P)$, in which P represents the appropriate probability (from the 240-element vector) that an examinee of average ability has the knowledge to classify the option. To determine whether a specific examinee could classify an option, the height of the intersection of the unit ordinate at the examinee's ability level was determined, as shown in Figure 2 for a sample case with ability of -1. A pseudorandom number between 0 and 1 was then drawn. If this number was smaller than the height of the intersection, the examinee "knew" the option. This procedure progressed through all 240 options, and counts of items answered correctly, omissions, options classified, etc., were kept as in the case of the model-based simulation. Similarly, the various finite-state scores and conventional test scores were computed. Ability values were rescaled linearly, with -3 (see Figure 2) corresponding to 0 and with 3 corresponding to 1, to make them comparable to the true λs from the model-based simulation.

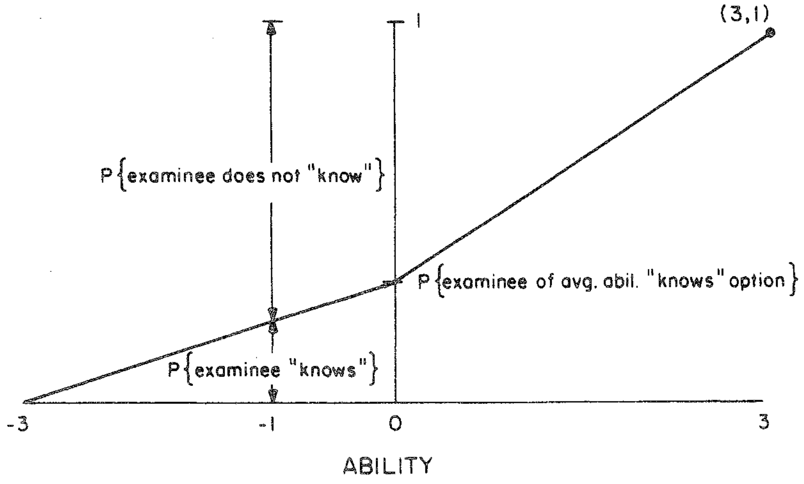
Analysis of Data

The simulation procedures described above were invoked several times using different seeds for the pseudorandom-number generator. All runs yielded highly similar results, and only the first is reported here.

Table 1 lists the variables of interest and reports their means and standard deviations. Note that in this and the following tables, the results for the three finite-state scores (λ estimates) resulting from NC behavior are identical. This is because the roots of Equations 12, 13, and 14 between 0 and 1 are equal when an examinee has no omissions (i.e., when $w = 1 - c$). This may be confirmed by substituting $1 - c$ for w in Equation 14, which makes it identical with Equation 13. Similarly, substituting $1 - c$ for w in Equation 12 and dividing the result by $\lambda^2 + \lambda + 1$ yields Equation 13.

To investigate questions concerning the interpretability of finite-state scores, new variables were constituted consisting of the differences between the various finite-state scores and the observed proportions

Figure 2
Linear Functions for Determining Whether Any Examinee "Knows" an Option
Illustrated for an Examinee With Ability of -1


1
(3,1)
P{examinee does not "know"}
P{examinee of avg. abil. "knows" option}
P{examinee "knows"}
-3
-1
0
3
ABILITY

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1
Descriptive Statistics, $N = 300$

		Method of Simulation			
		Model-Based		Naturalistic	
Behavior	Score	Mean	SD	Mean	SD
-	ℓ_{PRE}	.50	.15	.50	.18
-	ℓ_{POS}	.68	.16	.68	.18
NC	Number Correct/60	.73	.13	.72	.15
NC	Number Correct rescaled[1]	.64	.18	.63	.21
FS	Formula Score/60	.63	.18	.63	.20
RO	Formula Score/60	.61	.18	.61	.21
NC	$\hat{\lambda}_{\text{NC}}$	.51	.16	.50	.19
FS	$\hat{\lambda}_{\text{NC}}$	.47	.19	.48	.21
RO	$\hat{\lambda}_{\text{NC}}$	.45	.18	.45	.20
NC	$\hat{\lambda}_{\text{FS}}$	.51	.16	.50	.19
FS	$\hat{\lambda}_{\text{FS}}$	.50	.17	.51	.18
RO	$\hat{\lambda}_{\text{FS}}$	.48	.17	.49	.19
NC	$\hat{\lambda}_{\text{RO}}$	.51	.16	.50	.19
FS	$\hat{\lambda}_{\text{RO}}$	.51	.17	.52	.18
RO	$\hat{\lambda}_{\text{RO}}$	.50	.17	.51	.19

[1] Rescaled linearly with 15 or less (chance level) corresponding to 0 and 60 (perfect score) corresponding to 1. These scores are equivalent to formula scores in the absence of omissions, i.e., formula scores under NC behavior.

of options correctly classified by examinees both before and after seeing them together in the test items (ℓ_{PRE} and ℓ_{POS}). In addition, differences between number-correct and formula scores (expressed as proportions) and these observed proportions of correctly classified options were produced. For this use the number-correct scores were linearly rescaled, with 0 corresponding to 15 or less (the chance level) and 60 (a perfect score) corresponding to 1.

Table 2 lists the more relevant of these difference variables and provides their means and standard deviations. The difference variables were created by subtracting the criterion variables listed in the table from the appropriate scores; hence positive means indicate overall overestimation and negative means indicate underestimation. In addition, Table 2 shows the results of t tests of the null hypotheses that the mean population differences are 0. Contrary to common usage, only instances of *failure* to reject this null hypothesis at the .01 level of significance are indicated by asterisks in Table 2. This level was chosen arbitrarily. The magnitudes of the significant differences were the main focus of this analysis.

Table 3 gives product-moment correlations between three relevant criteria and various finite-state scores, number-correct scores, and formula scores. The first criterion is ability, the true λ for the case of the model-based simulation and the unit-normal ability measure for the naturalistic simulation. The second and third criteria are ℓ_{PRE} and ℓ_{POS}.

Results

Ranking Characteristics

The functional relationships between the finite-state scores and the number-correct or formula scores (as appropriate) are shown in Figure 3. For the RO behavior and Equation 12, there are actually an infinite

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 2
Accuracy of Estimation of Ability and Proportions of Alternatives Correctly Classified, $N = 300$

			Score-Criterion Differences			
			Model-Based Simulation		Naturalistic Simulation	
Behavior	Score	Criterion	Mean	SD	Mean	SD
NC	$\hat{\lambda}_{NC}$	ℓ_{PRE}	.01*	.06	.00*	.06
FS	$\hat{\lambda}_{NC}$	ℓ_{PRE}	−.03	.06	−.02	.06
RO	$\hat{\lambda}_{NC}$	ℓ_{PRE}	−.05	.06	−.05	.06
NC	$\hat{\lambda}_{FS}$	ℓ_{PRE}	.01*	.06	.00*	.06
FS	$\hat{\lambda}_{FS}$	ℓ_{PRE}	.00*	.05	.01*	.05
RO	$\hat{\lambda}_{FS}$	ℓ_{PRE}	−.02	.05	−.01	.05
NC	$\hat{\lambda}_{RO}$	ℓ_{PRE}	.01*	.06	.00*	.06
FS	$\hat{\lambda}_{RO}$	ℓ_{PRE}	.01	.06	.02	.05
RO	$\hat{\lambda}_{RO}$	ℓ_{PRE}	.00*	.05	.01*	.05
NC	Number Correct[1]	ℓ_{PRE}	.14	.07	.13	.07
FS	Formula Score/60	ℓ_{PRE}	.13	.07	.13	.06
RO	Formula Score/60	ℓ_{PRE}	.11	.07	.11	.07
NC	$\hat{\lambda}_{NC}$	λ or Abil.	.01*	.06	.01	.06
FS	$\hat{\lambda}_{NC}$	λ or Abil.	−.03	.07	−.01	.06
RO	$\hat{\lambda}_{NC}$	λ or Abil.	−.05	.07	−.04	.07
NC	$\hat{\lambda}_{FS}$	λ or Abil.	.01*	.06	.01	.06
FS	$\hat{\lambda}_{FS}$	λ or Abil.	.00*	.06	.02	.06
RO	$\hat{\lambda}_{FS}$	λ or Abil.	−.02	.06	.00*	.06
NC	$\hat{\lambda}_{RO}$	λ or Abil.	.01*	.06	.01	.06
FS	$\hat{\lambda}_{RO}$	λ or Abil.	.01	.06	.03	.06
RO	$\hat{\lambda}_{RO}$	λ or Abil.	.00*	.06	.02	.06
NC	Number Correct[1]	λ or Abil.	.14	.07	.14	.08
FS	Formula Score/60	λ or Abil.	.13	.07	.14	.07
RO	Formula Score/60	λ or Abil.	.11	.07	.12	.08
NC	Number Correct[1]	ℓ_{POS}	−.05	.05	−.05	.06
FS	Formula Score/60	ℓ_{POS}	−.06	.07	−.05	.05
RO	Formula Score/60	ℓ_{POS}	−.07	.07	−.07	.05

*Probability that population mean score-criterion > 0 exceeds .01.

[1] Rescaled linearly with a score of 15 or less (chance level) corresponding to 0 and 60 (perfect score) corresponding to 1. This change puts the number-correct scores on the same scale as the true λs and the already rescaled ability measures for the naturalistic simulation. These scores are equivalent to formula scores under NC behavior.

number of functions corresponding to the possible values of γ, all of which are within the darkened region for this relationship. As may be noted from Figure 3, the relationships are very nearly linear over most of their ranges and are all monotone increasing. Accordingly, the Pearson correlations between number-correct and formula scores and the corresponding finite-state scores were all .99 or higher, and the Spearman correlations were perfect or nearly perfect. Relationships between finite-state scores and

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 3
Product-Moment Correlations of Finite-State, Number-Correct and Formula Scores with Criteria, $N = 300$

		Simulation Method and Criteria					
		Model-Based			Naturalistic		
Behavior	Score	ℓ_{PRE}	ℓ_{POS}	λ	ℓ_{PRE}	ℓ_{POS}	Abil.
NC	Number Correct	.93	.96	.92	.94	.96	.93
FS	Formula Score	.94	.93	.93	.95	.97	.94
RO	Formula Score	.94	.92	.92	.95	.97	.93
NC	$\hat{\lambda}_{NC}$	.94	.94	.93	.96	.94	.95
FS	$\hat{\lambda}_{NC}$	.96	.93	.95	.97	.97	.96
RO	$\hat{\lambda}_{NC}$	.94	.90	.93	.96	.95	.95
NC	$\hat{\lambda}_{FS}$	.94	.94	.93	.96	.94	.95
FS	$\hat{\lambda}_{FS}$	.95	.91	.94	.96	.95	.95
RO	$\hat{\lambda}_{FS}$	.95	.91	.93	.96	.95	.95
NC	$\hat{\lambda}_{RO}$	.94	.94	.93	.96	.94	.95
FS	$\hat{\lambda}_{RO}$	.94	.90	.93	.95	.95	.95
RO	$\hat{\lambda}_{RO}$	.95	.91	.93	.96	.96	.95

number-correct or formula scores when the response strategy is not consistent with the scoring equation were only slightly lower, with Pearson correlations between .98 and .99, and with Spearman correlations also nearly perfect. Both the model-based and naturalistic simulations produced these results. Finite-state scores, therefore, provide essentially the same rankings of the examinees according to knowledge or ability as do number-correct or formula scores.

An interesting point arises from inspection of the relationship for the RO behavior depicted in Figure 3. Note that for any given finite-state score, there is a more or less broad range of possible formula scores that are due to variation in guessing propensity. This illustrates the well-known fact that formula scores are affected by examinees' willingness to guess after identification of distractors. Interestingly, use of $\hat{\lambda}_{RO}$ prevents guessing propensity from affecting finite-state scores, because these two parameters are separately considered by the model in their relation to partial information. Appropriate formulas can be produced for estimating γ, analogous to those given above for estimating λ.

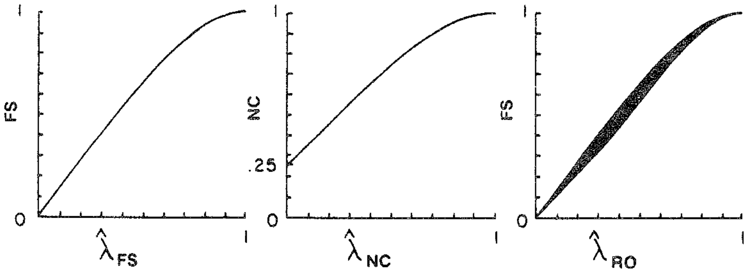
Psychological Interpretability

Table 2 shows that the various finite-state scores provide direct and much closer estimates of ℓ_{PRE}, true λ, and ability than do the transformed number-correct or formula scores. Generally, these estimates are slightly closer when the response behaviors and scoring equations are consistent. Indeed, in most such cases, the finite-state score means are not significantly different from the true λ or ℓ_{PRE} means ($p > .01$), as may be noted from the location of the asterisks in Table 2. Therefore, with reasonable examinee adherence to test directions, finite-state scores have very good potential for providing direct and accurate estimates of the proportions of options classifiable by the examinees before taking the test. This result confirms that finite-state scores are informative about knowledge at the option level even when the conventional response procedure is used.

Table 3 reveals that generally the finite-state scores are minimally more strongly related to the λs and ability than number-correct or formula scores. When the criterion is the proportions of options correctly

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Figure 3
Functional Relationships Between λ Estimates and Appropriate Conventional Scores



classified, the differences between coefficients are even smaller and less systematic. Moreover, throughout Table 3 there is little difference between correlations when the scoring model is appropriate for the response strategy or when it is not. However, there is a slight tendency for the finite-state scores to yield higher coefficients when the response strategy is consistent. In any case, a result that appears systematically in Table 3 is that, for any given response behavior, finite-state scores have slightly higher correlations with ability, λ, and ℓ_{PRE} than conventional scores, whereas conventional scores are slightly more highly correlated with ℓ_{POS} than finite-state scores.

The fact that these linear correlations are so high implies that the criteria shown in Table 3 could be estimated almost equally well from any of the scoring models. Table 2 makes the point that finite-state scores are direct estimates of those criteria, whereas conventional scores, being high overestimators of them, would require a transformation (regression) based on statistics that are not available in a real testing situation.

Effect of Inappropriate Behavior

Although, in general, examinee adherence to behavior consistent with the scoring model yielded higher correlations with and better estimates of criteria, this tendency was not uniform in the results reported above. In any case, differences were quite small, which suggests a somewhat optimistic outlook with respect to the unavoidable deviations of examinee behavior from that which is optimal for the scoring model to be applied. Perhaps surprisingly, the λ equation for the RO behavior (Equation 12) was quite robust when used to evaluate responses generated under the other two behaviors, especially when used to estimate ℓ_{PRE}. This robustness may be due to the fact that the equation for the RO behavior was derived without making any assumption about γ, unlike the equations for NC or FS behaviors. Hence it would seem to accommodate what amounts to a high γ in the case of NC behavior and a low γ in the case of FS behavior (reflecting the tendency to omit only when completely ignorant).

Initial Versus Final Knowledge Level

The results just reviewed show substantial differences in initial versus final proportions of options the examinees can classify as correct or incorrect. From Table 1, the mean proportions increased from about .50 to .68 in either simulation method. Clearly the finite-state scores provide close estimates of the proportions known before the options were seen together in the items. As shown in Table 2, the transformed formula or number-correct scores provide (less accurate) underestimates of the proportions classifiable at the termination of the test. However, they do underestimate this quantity somewhat less than they overestimate the proportion known in advance.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Approaches to Simulation

For all practical purposes, there appears to be no difference in the accuracy of the finite-state scores arising from the naturalistic versus the model-based simulation. Tables 2 and 3 confirm this claim, though there is some indication of minimal degradation (basically with respect to statistical significance only, not magnitude) of the estimation properties of the λ estimates arising from the naturalistic versus the model-based simulation. This robustness provides some support for the expectation that finite-state scores should perform well with real data. It may also be viewed as an indication that variations in the difficulty of items and options do not need to be taken into consideration when applying finite-state scoring to ability estimation.

Effect of Test Length

Because test length was held constant at 60 items in these simulations to avoid undue complexity, consideration should be given to the outcomes that might be expected for shorter tests. The concern here is not with the reliability of finite-state scores because, as (nearly linear) transformations of the number of correct responses (and possibly omissions), they would be about as reliable as the conventional scores. Instead, concern would center on the accuracy of estimation for the finite-state scores. In this regard, it is noteworthy that the 60-item simulations yielded average $(\hat{\lambda} - \ell_{\mathrm{PRE}})$ errors of less than 1% of ℓ_{PRE} when behavior was consistent with the scoring model. Because this accuracy is somewhat better than that provided by conventional scores (interpreted as percentages) for the same number of items ($1/60 = 1.7\%$), finite-state scores would be expected to perform relatively as well for tests with fewer items. This is confirmed by the tables in García-Pérez (1989).

Discussion and Conclusions

The results just reviewed confirm very strongly that finite-state scores may be interpreted as estimates of the proportion of options examinees can classify before seeing them together in the items of a test. This finding has substantial educational implications. However, before discussing these implications, some practical matters should be considered.

The circumstances under which finite-state scores would be preferable to number-correct or formula scores (or their transformations, such as *T* scores) must be determined. The rank correlation results reported earlier clearly indicate that there is no advantage in providing finite-state scores when the scores will be used only for ranking purposes. Thus, when scores are reported in terms of norms or percentile ranks, the use of finite-state scores would probably only serve to confuse the already difficult process of communicating testing results.

In contrast, nearly all teachers tend to interpret scores from classroom tests in a criterion-referenced sense (see Bowman & Frary, 1983). The deficiency of number-correct and formula scores for this purpose, as outlined above, can make such interpretations highly misleading. A score of, say, 70% on a test scored number-correct (or, for that matter, formula scored) might raise the question "70% of what?" At the same time many instructors and students believe that such a score implies having learned 70% of some body of subject matter. The use of finite-state scores could bring new insights to instructors and students concerning what levels and types of knowledge they possess.

Obviously, the use of finite-state scores would not be feasible for all classroom testing. The solution of polynomials of degree four or higher cannot reasonably be done by hand. However, when multiple-choice test responses are collected for processing by an optical mark reader, there would be little additional burden on the scoring/analysis system to provide finite-state scores along with the usual number-correct

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

or formula scores. Alternatively, it would be possible to compile tables giving $\hat{\lambda}$ as a function of c and w for different numbers of choices and different assumptions about guessing behavior.

However, finite-state scores should not be provided without extensive efforts to explain their characteristics to users. The present results indicate that the proportion of options classifiable before seeing the test is likely to be somewhat less than afterward, and also less than the proportion correct or even the formula score expressed as a proportion (see Table 1). Accordingly, instructors and students would need to consider much more analytically what their states of knowledge are with respect not only to the correct answers but also to the distractors. For instructors, this consideration should lead to more thoughtful development of test items, especially their distractors. Students should gain from consideration not only of their knowledge of the answers but of why each distractor fails to serve as an answer.

Another potential use for finite-state scores is in the area of criterion-referenced testing. The use of finite-state scores in this environment is described in García-Pérez (1989), where the number of items needed to estimate with prescribed precision the true λ of an examinee is considered as a function of the format of administration of the test and the response behavior of the examinees. Currently, many instructional systems use scores on multiple-choice tests as criteria for placement and advancement. In contrast to the finite-state approach just described, the bases for establishing these criteria have often been no more than trial-and-error judgments or have been based on rather simplistic models. Thus finite-state scores should provide a much better basis for setting criteria.

References

Bowman, R. W., Jr., & Frary, R. B. (1983). *Difficulty level of classroom tests: Communicating with teachers.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Frary, R. B. (1982). A simulation study of the reliability and validity of multiple-choice test scores under six response-scoring modes. *Journal of Educational Statistics, 7,* 333–351.

Frary, R. B. (1989). Partial credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2,* 79–96.

García-Pérez, M. A. (1987). A finite state theory of performance in multiple-choice tests. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology-I* (pp. 455–464). Amsterdam: Elsevier.

García-Pérez, M. A. (1989). Item sampling, guessing, partial information, and decision-making in achievement testing. In E. E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 249–265). New York: Springer-Verlag.

García-Pérez, M. A. (in press). A comparison of two models of performance in multiple-choice tests: Finite states versus continuous distributions. *British Journal of Mathematical and Statistical Psychology.*

Hutchinson, T. P. (1982). Some theories of performance in multiple choice tests, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology, 35,* 71–89.

IMSL, Inc. (1987). *FORTRAN subroutines for statistical analysis.* Houston TX: Author.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: Vol. 1. Additive and polynomial representations.* New York: Academic Press.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12,* 7–11.

Molenaar, I. W. (1981). On Wilcox's latent structure model for guessing. *British Journal of Mathematical and Statistical Psychology, 34,* 224–228.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49,* 501–509.

Acknowledgments

A travel grant provided by the Virginia Tech Foundation permitted the authors to meet and greatly facilitated their collaboration in this research.

Author's Address

Send requests for reprints or further information to Robert B. Frary, Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg VA 24061, U.S.A.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/