Research Report

Reliability of Speeded Number-right Multiple-choice Tests

Yigal Attali

April 2004
RR-04-15

Reliability of Speeded Number-right Multiple-choice Tests

Yigal Attali

ETS, Princeton, NJ

April 2004

Abstract

Contrary to common belief, reliability estimates of number-right multiple-choice tests are not inflated by speededness. Because examinees guess on questions when they run out of time, the responses to these questions show less consistency with the responses of other questions, and the reliability of the test will be decreased. The surprising implication is that adding questions to a multiple-choice test may lower its reliability when the test is speeded. This paper develops the mathematical derivations and shows the effects of speededness on reliability in simulations.

Key words: Reliability, speededness, multiple-choice tests

Introduction

In their introduction to test theory, Crocker and Algina (1986) describe the effects of time limits on the reliability of power tests:

> When a test has a rigid time limit such that some examinees finish but others do not, an examinee's working rate will systematically influence his or her performance on all forms of the test. . . . On [power] types of tests, time limits should be long enough to allow all, or nearly all, examinees to finish. Otherwise, the reliability estimate may be artificially inflated because of consistencies in performance caused by the test's time limit. (p. 145)

This argument for inflated reliability estimates is best presented with the odd-even split-half procedure for estimating reliability. For example, Crocker and Algina (1986, p. 145) note that "Once an examinee runs out of time, performance on all remaining uncompleted odd- and even-numbered items will be perfectly consistent." However, this argument holds only if, indeed, items that are not reached are also omitted by examinees. With multiple-choice (MC) tests that do not have correction for guessing (i.e., are scored as number-right), examinees may earn some points by guessing the answers to the remaining items, those that were not reached. With these guessed answers there will be no consistency in performance, because these responses reflect no knowledge, but some of them will be correct nonetheless.

Intuitively, it is not plausible that reliability estimates will be inflated by these random responses. Yet, psychometric textbooks do not refer to this possibility in their analysis of the effects of a speed component on reliability estimates (e.g., Anastasi & Urbina, 1997; Crocker & Algina, 1986; Gulliksen, 1950; Lord & Novick, 1968; Stanley, 1971; Traub, 1994).

Take, for example, Gulliksen's (1950) discussion of this topic. Gulliksen makes a thorough analysis of speed versus power tests, but implicitly assumes a test with no opportunity to successfully guess the answer of items. Gulliksen's (p. 230) distinction between speed and power tests is based on the decomposition of total error score, X, to W (number of wrong answers) and U (number of not reached answers). In Gulliksen's discussion a "pure speed" test is a test where W will be zero for each examinee, hence X = U, and a "pure power" test is a test where U is zero for each examinee, and, hence, X = W. However, for a MC test in which examinees who run out of time randomly guess the answer to the questions that were not reached, U will be zero, even though the test is actually speeded. Thus, the decomposition of the error score into W and U does not distinguish between speeded and nonspeeded MC tests, when examinees guess the answers to the items they did not reach.

Despite the fact that many tests use MC items, the psychometric literature did not seem to acknowledge this effect of random guessing on the reliability of speeded tests. The purpose of this paper is to provide a theoretical account of the effects of speededness on reliability of MC power tests, and to support this account with simulation data.

The Effect of Random Guessing on the Performance of a Single Item

Following Schnipke and Scrams (1997) and Yamamoto (1995), it is assumed that examinees choose to engage in either solution behavior or rapid-guessing behavior in answering each item. Because examinees spend very little time on the item in rapid-guessing, their answers may be characterized as random guessing, and the probability of a correct answer for these responses can be assumed to be $1/k$, where k is the number of options for the MC item. This assumption is obviously a simplification of the actual process that examinees are experiencing. In real life, examinees might experience more and more time-related pressure and react by gradually shortening the time they spend on each item (and gradually lowering their probability of answering correctly). Because the research on this issue is very limited, it is difficult to determine the correctness of this two-state assumption. However, Schnipke and Scrams (1997) were successful in modeling item response times with such a two-state mixture model. The two-state assumption is plausible because examinees know in advance that the time limits of a particular test are very strict. Consequently, the time pressure they feel in speeded tests is reasonably high from the start of the test, and towards the end of the test, examinees find themselves unable to solve items even more rapidly.

A useful way to analyze the effect of random guessing on internal consistency measures of reliability is to concentrate on the item level. The following discussion will show that when some of the examinees are not answering an item and instead engage in rapid-guessing, scores on this item are less correlated with other measures of performance, either another item in the test or the test as a whole. The analysis is based on the observation that when examinees are randomly guessing the answer to an item, their responses will, by definition, be independent and uncorrelated with their responses to any other item.

It should be noted that the decrease in correlations would occur only insofar as response time is indeed irrelevant to the trait being tested, as is assumed for power tests. When differences in speed of response are positively correlated with differences in the trait being measured (e.g., perceptual speed tests), the speed component might not lower these correlations. However, the effects of moderate correlations between speed and performance will be examined.

First, it will be shown that the correlation between a partially speeded item and any other item is decreasing *more or less linearly* with the rate of random responding for the partially speeded item. The following derivations will examine the correlations between the (dichotomous) responses to an arbitrary item *i*, a completely unspeeded item *j*, a completely speeded item *g*, and a partially speeded item *k*. The correlation coefficient (the *phi* coefficient) between the arbitrary item *i* and the unspeeded item *j* is given by

$$\phi_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i q_i p_j q_j}},$$

where p_{ij} is the joint proportion of examinees answering both items correctly, p_i and p_j are the proportions of correct responses for the items, q_i and q_j are the proportions of incorrect responses for the items.

For an item *g* that is completely speeded, that is, all examinees are randomly responding to it, p_g is equal to one over the number of options. Moreover, because all examinees are randomly responding to this item, the responses to this item are independent of the responses to any other item, and the joint proportion of correct answers to this item and any other item *i*, p_{ig}, is equal to the product of the marginal proportions $p_i p_g$, and consequently ϕ_{ig} is equal to 0.

For an item *k* that is partially speeded, it will be assumed that its difficulty for examinees engaged in solution behavior is equal to the difficulty of the unspeeded item *j* (in other words, if *k* would not be partially speeded, its difficulty would be the same as *j*). For this item *k*, *A%* of examinees are answering the item and *G%* are randomly guessing the answer (*G* = 1 – *A*). The proportion of correct responses for this item is a weighted average of the proportions of correct answers for the unspeeded item *j* and the completely speeded item *g*:

$$p_k = Ap_j + Gp_g .$$

Similarly, the joint proportion of examinees answering both items *i* and *k* correctly is a weighted average of the joint proportions of correct answers for items *i* and *j* and for items *i* and *g*:

$$p_{ik} = Ap_{ij} + Gp_{ig} = Ap_{ij} + Gp_i p_g \, .$$

As was previously indicated, the joint proportion of correct answers to items *i* and *g* is equal to the product of the marginal proportions for these items. The correlation between items *i* and *k* is given by:

$$\phi_{ik} = \frac{p_{ik} - p_i p_k}{\sqrt{p_i q_i p_k q_k}} = \frac{Ap_{ij} + Gp_i p_g - p_i(Ap_j + Gp_g)}{\sqrt{p_i q_i p_k q_k}} = \frac{A(p_{ij} - p_i p_j)}{\sqrt{p_i q_i p_k q_k}} \, .$$

Finally, in most cases the variance of item *k* ($p_k q_k$) can be approximated with that of item *j* ($p_j q_j$) because the effect of guessing on the variance of item *k* is small and inconsistent. For example, if G = .2 and *j* is an easy item (p_j = .8), the variance of item *j* is equal to .16 and the variance of item *k* is equal to .22. If *j* is a moderately difficult item (p_j = .6), the variance of item *j* is equal to .24 and the variance of item *k* is equal to .25. If *j* is a difficult item (p_j = .4), the variance of item *j* is still equal to .24 and the variance of item *k* is equal to .23. In summary, except for very easy items the variance of item *k* is similar to the variance of item *j* and consequently, the correlation between item *i* and item *k* is approximately equal to *A* times the correlation of item *i* and item *j*:

$$\phi_{ik} = \frac{A(p_{ij} - p_i p_j)}{\sqrt{p_i q_i p_k q_k}} \cong \frac{A(p_{ij} - p_i p_j)}{\sqrt{p_i q_i p_j q_j}} \cong A\phi_{ij} \, .$$

If the original item *j* is easy (high percentage correct) then this approximation is actually an overestimate, whereas for a difficult item this is a slight underestimate. As an example, the case where p_{ij} is equal to 40% and both p_i and p_j are equal to 50% results in a ϕ_{ij} of 0.20. When some of the examinees are randomly responding to item *j*, transforming it to a partially speeded item k, ϕ_{ik} is decreased almost linearly. Figure 1 shows the actual values of ϕ_{ik} together with the linear decrease line.

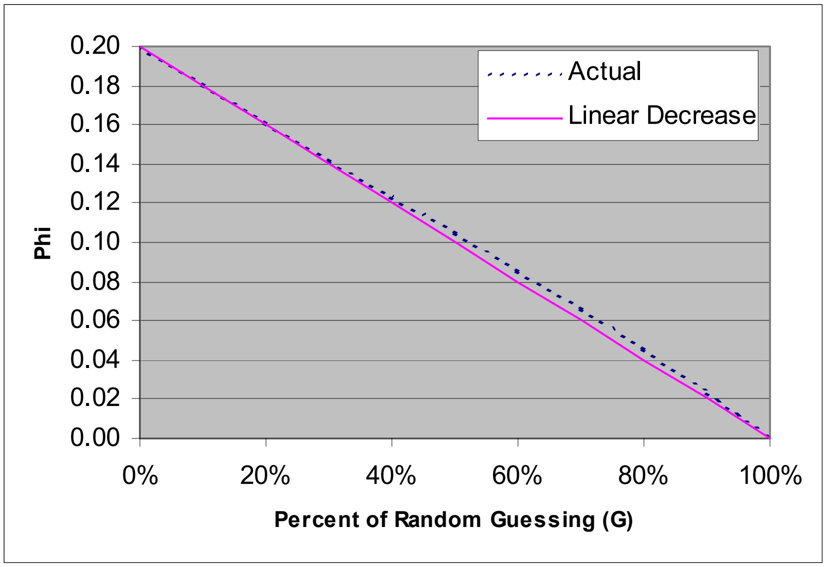

Figure 1. **ϕ_{ik} as a function of percentage of random guessing.**

This approximate linear decrease in the correlation of a speeded item with another item will have a similar effect on the (point-biserial) correlation of the partially speeded item with the total test score, because an item's point-biserial is closely related to its inter-item correlations. Gulliksen (1950, p. 376) shows that the sum of the terms in any column (or row) of the inter-item variance-covariance matrix is the covariance between that item and the total test score. For every item i,

$$\rho_{ix}\sigma_i\sigma_x = \sum_{j=1}^{N} \rho_{ij}\sigma_i\sigma_j \qquad (\rho_{ii} = 1),$$

where ρ_{ix} is the point-biserial between item i and total score x, and σ_x is the standard deviation of test scores. The point-biserial of item i is then given by

$$\rho_{ix} = \frac{\sum_{j=1}^{N} \rho_{ij} \sigma_i \sigma_j}{\sigma_i \sigma_x} .$$

Inspection of this last equation and the previous discussion suggests that increasing the speededness of item *i* will primarily have an effect on inter-item correlations (approximately linearly), and will have little effect on the standard deviation of the speeded item or of the test scores (especially if the number of test items is reasonably large). Consequently, the effect of speededness on the point-biserial of a partially speeded item can also be assumed to be approximately linear.

Adding an Item Without Increasing Reliability

These effects, in turn, will have a negative effect on the internal consistency measures of reliability for the test. When a single speeded item is added to a test, this effect may be very small and depends on many factors. However, another interesting question could be answered in more general terms: What is the percentage of random guessing for the added item that will make the new reliability estimate, with the added item, the same as the reliability of the test without the speeded item? In other words, what is the equilibrium point of the guessing rate where the noise introduced by guessing on an item cancels the effect of adding the item to a test?

Coefficient alpha for a test with *n* items depends on the number of items, the sum of item variances, and score variance:

$$\alpha_N = \frac{n}{n-1}(1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_N^2}) = \frac{n}{n-1}(1 - \frac{n\overline{\sigma_i^2}}{\sigma_{x(N)}^2}) .$$

Score variance, in turn, is equal to the sum of the elements in the variance-covariance matrix:

$$\sigma_{x(N)}^2 = n\overline{\sigma_i^2} + n(n-1)\overline{\sigma_{ij}} .$$

The first term is the sum of the item variances on the main diagonal, and the second term is the sum of all covariances (σ_{ij}) between any two different items *i* and *j* in the matrix. When a partially speeded item is added to the test the *new* test variance is approximately equal to:

$$\sigma^2_{x(N+1)} \cong (n+1)\overline{\sigma^2_i} + n(n-1)\overline{\sigma_{ij}} + 2nA\overline{\sigma_{ij}} .$$

The approximation is based on the results of the previous section. The first term is the sum of the item variances, and it is assumed that the new item variance is equal to the average of all previous items. The second term is the sum of the old item covariances (here nothing is assumed). The third term is the sum of the new 2*n* covariances associated with the new item (the last row and column of the new covariance matrix, without the element on the main diagonal). The result from the previous section can be used here: that the correlation between the new partially speeded item and any other item is decreased by a factor of *A*, the percentage of examinees answering (and not guessing) the last item. To summarize, the derivation of the new score variance assumes that the new item has about the same item variance as other items and that its correlation with other items is reduced to about *A* times the typical correlation between any two items.

Coefficient alpha for the new test is given by:

$$\alpha_{N+1} = \frac{n+1}{n}(1-\frac{\sum_{i=1}^{n+1}\sigma^2_i}{\sigma^2_{x(N+1)}}) = \frac{n+1}{n}(1-\frac{(n+1)\overline{\sigma^2_i}}{(n+1)\overline{\sigma^2_i} + n(n-1)\overline{\sigma_{ij}} + 2nA\overline{\sigma_{ij}}}) .$$

The question is now whether the relation between coefficient alpha for *n* items and coefficient alpha for *n*+1 items can be expressed in terms of *A*, the proportion of examinees answering, and not guessing, the new item. With some algebra it can be shown that if the two ratios between the sum of item variances and score variance are equal, then the following equation about *A* also holds:

$$\frac{(n+1)\overline{\sigma_i^2}}{(n+1)\overline{\sigma_i^2}+n(n-1)\overline{\sigma_{ij}}+2nA\overline{\sigma_{ij}}}=\frac{n\overline{\sigma_i^2}}{n\overline{\sigma_i^2}+n(n-1)\overline{\sigma_{ij}}} \qquad \Rightarrow \qquad A=\frac{n-1}{2n}.$$

And, following this, if the two coefficient alphas are equal then the following inequalities about *A* and *G* holds, too:

$$\alpha_N=\alpha_{N+1} \qquad \Rightarrow \qquad A>\frac{n-1}{2n} \qquad \Rightarrow \qquad G<\frac{n+1}{2n}.$$

The meaning of this inequality is that a little more than a 50% rate of guessing for the added item is, at most, the degree of speededness for which the coefficient alpha of the test including the speeded item will be equal to the coefficient alpha without that item (for *n* = 20 this rate is 55%, and for *n* = 50 this rate is 51%). In other words, when an item is added to a test and more than 50% of the examinees are forced to randomly guess the answer to this item because they do not have enough time left, the addition of that item may *lower* the internal consistency measure of reliability for this test.

Table 1 shows several computed examples of the equilibrium point of rate of guessing (*G*) of the new added item for which the new alpha for *n*+1 items will be equal to the old alpha with *n* items. Several factors were manipulated to observe their effect on the equilibrium *G*. For example (first row in Table 1), for a test with 20 items, average item difficulty of 40%, five choices per item, and average interitem correlation of .04, the coefficient alpha will be .45. If a new item is added to this test and the guessing rate is .55 or higher then the new coefficient alpha will be equal or lower than the original alpha of .45.

The table shows that the most influential factor is the average difficulty of test items—the difference in the equilibrium *G* between difficult tests (.40) and easy tests (.70) was 13%–15% with easy tests having higher *G* values. The second most influential factor was average item correlations—the difference in equilibrium *G* between low item correlations (.04) and high item correlations (.16) was 6%–7%, with low correlations having higher equilibrium *G* values. Both number of items and the number of choices per item had small effects on *G*.

Except for very difficult test items and very low item correlations (and reliabilities), equilibrium *G* did not exceed 50%. For a reasonably difficult test of .6, average interitem correlations of .12–.16, with 20–50 items, and a p_g of .20 or .25, the equilibrium *G* is .41–.44.

Table 1

Guessing Rate G of the n+1 Item Needed to Achieve Unchanged Alpha

Statistics for the test with *n* items						
Number of items	Average difficulty (p_i)	Number of choices	Percentage correct in guessing (p_g)	Average ρ_{ij}	α_N	Equilibrium *G*
20	.4	5	.20	.04	.45	.55
50	.4	4	.25	.16	.90	.46
20	.5	5	.20	.04	.45	.53
50	.5	4	.25	.16	.90	.45
20	.6	5	.20	.04	.45	.48
50	.6	4	.25	.16	.90	.41
20	.7	5	.20	.04	.45	.40
50	.7	4	.25	.16	.90	.34

The Effect of Speed Consistency on Alternate-form Reliability

The preceding discussion focused on internal consistency measures of reliability. With respect to alternate-form reliability, there are two different sources for the decrease of reliability due to the speed factor. In addition to the noise that is introduced to scores of each form as a result of guessing, the inconsistency in examinee response speed across different forms can lower the alternate-form reliability beyond the random guessing factor. If the reliability of response speed is perfect, that is, examinees complete the same number of items on both forms, then internal consistency measures of reliability will accurately estimate alternate-form reliability. But in the case of less than perfect speed consistency, the alternate-form reliability will be lower than single-form estimates. This effect will be demonstrated with a simulation.

The simulation was designed to show the effect of speed consistency on alternate-form reliability and its relation to internal consistency reliability. Setting the correlation between completion rates on two equivalent test forms operationalized speed consistency. Each simulated examinee was given two test forms composed of 50 equal-difficulty five-choice items drawn from a large pool of items. For each simulated examinee, a true score was defined as the proportion of items in the pool that the examinee can answer correctly. The true score of 100,000 examinees was sampled from a logit-normal distribution that was based on a normal distribution with parameters $\mu = 0$ and $\sigma = .7$ (the logit-normal transformation is $[1 + \exp(x)]^{-1}$). The mean of this distribution is .5, its standard deviation is approximately .16, and its shape is similar to the shape of a normal distribution with the same mean and standard deviation, except that its values are restricted to the [0,1] range. The range of this distribution was covered by setting the true score of the i^{th} examinee to be the $(i - .5) / 1000$ percentile of the distribution. The choice of the logit-normal distribution for the true score (and other) distributions was motivated by the ease of simulating such bivariate distributions with specified correlations.

In addition to the true score of each examinee, the simulation randomly determined the percentage of completed items for the two forms from a logit-normal distribution that was based on a normal distribution with parameters $\mu = -2.5$ and $\sigma = 1.0$. This is a skewed distribution with an approximate mean of .90 and a standard deviation of .10, so that we can expect on average a 90% completion rate (after the value of the completion rate was set it was rounded to the nearest number of items). Figure 2 shows the distribution of completed items.

The percentage of completed items for the second form was randomly determined from a preset correlation value between completion rates and the distribution of completion rates conditioned on the first form completion rate (Johnson, 1987).

For each examinee the first items of each 50-item form were completed and the last items were randomly guessed. For each of the completed items, the correctness of the examinee's response was randomly determined by generating a Bernoulli trial with the true score of the examinee as the probability of success. For each of the uncompleted items the correctness of the examinee's response was randomly determined by generating a Bernoulli trial with 20% (one over the number of options) as the probability of success.

As a reference point, the results of this simulation should be compared to the values of Cronbach alpha and alternate form reliability for completely unspeeded tests. These values were obtained in a separate simulation (where all items were completed by all examinees) and both were equal to .847.

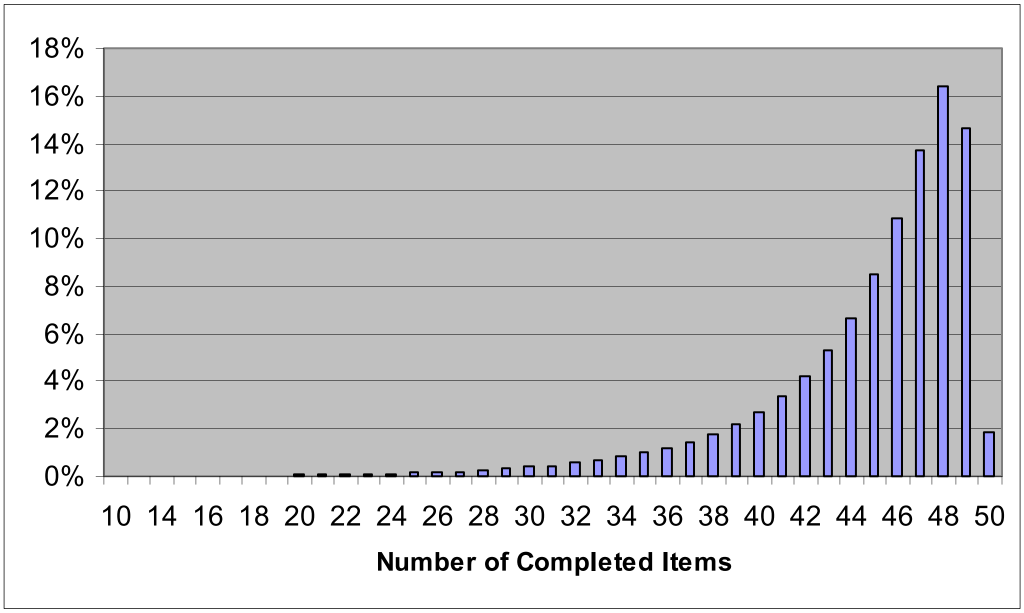

Figure 2. **Percentage of completed items from the logit-normal distribution based on parameters μ = -2.5 and σ = 1.0.**

Table 2 shows the Cronbach alpha (for the first form) and the alternate form score correlation as a function of speed consistency. As expected, the internal consistency reliability estimate was not affected by the speed consistency and its value was around .824 for all correlation values between completed items. Applying the Spearman-Brown formula to the speeded and unspeeded coefficient alpha values reveals that this decrease in reliability, compared to the value for the unspeeded 50-item test (.847), corresponds to about 8 items in test length. However, as expected, Table 2 also shows that the alternate form reliability is lowered even further when speed consistency decreases. This decrease is significant—from an estimated number of 44 unspeeded items to 33 items.

These results show that, although internal consistency measures of reliability for speeded MC tests reflect the decrease in internal consistency due to noise introduced by guessing, they do not reflect the further decline in reliability that results from the inconsistency of this noise across occasions.

Table 2

Internal Consistency and Alternate Form Reliability as a Function of Speed Consistency

Correlation between completed items	Cronbach alpha for first form	Alternate form correlation	No. of unspeeded items with same alternate form correlation
1.0	.824	.829	44
0.8	.823	.820	41
0.6	.824	.808	38
0.4	.823	.801	36
0.2	.823	.795	35
0.0	.824	.787	33

Relation Between Speed and Performance and the Effect of Speededness on Reliability

The preceding discussion assumes that there is no relation between individual speed and performance on the test. This standard assumption for power tests may not hold in reality; however, past research on this issue generally found that, in tests that require reasoning, the correlation between response time and performance is typically *positive* (for a review, see Schnipke & Scrams, 2002). Scrams and Schnipke (1997), for example, found that higher-ability examinees on the GRE® General Test (a nonadaptive computerized version) took *more* time to respond than low-ability examinees for the verbal- and quantitative-reasoning subtests (r^2 = .39 and .33, respectively, for the relation between estimated ability and slowness examinee parameters) and no relation was found for the analytical subtest (r^2 = .00). This state of affairs (positive speed-performance correlations) will translate into larger effects of speededness on test reliability, because higher ability examinees will tend to experience more time-related pressure, they will be forced to guess the answers of more items, and thus the noise due to the speededness factor will reduce the differences in ability between examinees. Negative correlations between response speed and performance will have an opposite influence and may cancel the effects of speededness on reliability that were shown above.

The following simulation examined the sensitivity of the previous results to positive correlations between performance on the test and number of items completed. The simulation was similar to the one described in the previous section, except that after the true score for the performance of the examinee was generated, a *true score* for the percentage of completed items was randomly determined from a preset correlation value between true score completion rates and true score performance rates. This correlation determined the degree of relation between speed and performance. Then the actual completion rates for the two forms were randomly generated from the true completion rate based on a relatively high completion rate reliability of .8.

Table 3 presents the results of this simulation for true performance-completion correlations between .0 and .6 (a positive correlation between performance and *completion rate* corresponds to a negative correlation between performance and *response speed*). As in the previous simulation, the results in Table 3 should be compared to the .847 reliability of the unspeeded test (where all examinees complete all items). The table shows that positive correlations (corresponding to negative correlations between performance and speed) indeed result in higher reliabilities compared to the case of zero speed-performance correlation; however, even for a relatively high correlation of .6, the reliabilities did not reach the original .847 reliability of the unspeeded test.

Table 3

Internal Consistency and Alternate Form Reliability as a Function the Correlation Between True Performance and True Completion Rate (With Completion Reliability of .8)

Correlation between true performance and true completion rate	Cronbach alpha for first form	Alternate form correlation
0.6	.844	.841
0.4	.839	.834
0.2	.833	.825
0.0	.823	.814

Conclusion

The purpose of this paper is to show that, in MC tests that do not penalize guessing, reliability measures are lowered by speededness in the tests. Because the examinees who run out of time guess the remaining answers instead of omitting them, the speededness produces noise in examinees' responses, lowering the reliability of the test.

The amount of random guessing in an item is associated with the item's correlation with scores for another item or a test (and, particularly, with the item's point-biserial). Furthermore, when an item is added to a test and no more than one-half of the examinees will have enough time to solve it, the internal consistency reliability of the test is likely to decrease. The surprising implication is that it is possible to reduce the number of items in a speeded MC test and still retain the same level of reliability or even increase reliability. The effects of speededness on alternate-form reliability depend in addition on the reliability of examinees' response speed across test forms. When examinees do not complete the same number of items on different forms, the alternate-form reliability may be lower than internal consistency reliability for a single form. Even when there is a positive correlation between performance and completion rate the reduction in reliability associated with speededness is not completely erased.

In conclusion, this paper shows that the presence of speededness on multiple-choice tests has adverse consequences for their psychometric properties.

References

Anastasi, A., & Urbina, S. (1997). Reliability of speeded tests. In A. Anastasi & S. Urbina (Eds.), *Psychological testing* (7th ed., pp. 102–105). Upper Saddle River, NJ: Prentice Hall.

Crocker, L., & Algina, J. (1986). Factors that affect reliability coefficients. In L. Crocker & J. Algina (Eds.), *Introduction to classical and modern test theory* (pp. 143–146). Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons.

Johnson, M. E. (1987). *Multivariate statistical simulation.* New York: John Wiley & Sons.

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

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–443). Washington, DC: American Council on Education.

Traub, R. E. (1994). Factors affecting the reliability coefficient. In R. E. Traub, *Reliability for the social sciences: Theory and applications* (pp. 98–114). Thousand Oaks, CA: Sage.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213–232.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing* (pp. 237–266). Mahwah, NJ: Lawrence Erlbaum.

Scrams, D. J., & Schnipke, D. L. (1997). *Making use of response times in standardized tests: Are accuracy and speed measuring the same thing?* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (ETS RR-95-02). Princeton, NJ: ETS.