Listening. Learning. Leading.®

Research Report
ETS RR–15-18

# Estimating Conditional Distributions of Scores on an Alternate Form of a Test

Samuel A. Livingston

Haiwen H. Chen

December 2015

# ETS Research Report Series



Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

# Estimating Conditional Distributions of Scores on an Alternate Form of a Test

Samuel A. Livingston & Haiwen H. Chen

Educational Testing Service, Princeton, NJ

Quantitative information about test score reliability can be presented in terms of the distribution of equated scores on an alternate form of the test for test takers with a given score on the form taken. In this paper, we describe a procedure for estimating that distribution, for any specified score on the test form taken, by estimating the joint distribution of equated scores on two forms of the test. The procedure requires as input the score distribution and the reliability coefficient, estimated for the test-taker population. In tryouts, the procedure performed well for score distributions like those commonly encountered in large-scale testing situations.

**Keywords** reliability; conditional distributions; alternate form; consistency

doi:10.1002/ets2.12066

How can testing agencies provide test users with a meaningful quantitative description of the reliability of the test scores? A test user knows what the test taker's reported score is. What the test user does not know — not even probabilistically — is how different that score would have been if the test taker had happened to take a different form of the test. The reliability coefficient does not provide that kind of information. The standard error of measurement — especially, the conditional standard error of measurement — does provide that kind of information but not in terms that are easily understood by test users without statistical or psychometric training. The test information curve provides reliability information in terms of a quantity that is even more difficult for test users to understand.

In this paper we describe an alternative procedure for describing the reliability of a test-taker's score: estimating a probability distribution for the score the test taker would have earned on another form of the test. Figure 1 shows one way to communicate this distribution to the test user in the form of a familiar graph. To understand the graph, the test user does not have to understand the concepts of true score or error of measurement. The graph shows how likely the test taker's score on an alternate form of the test would be to differ from the score the test taker actually received by any specified amount in either direction. Like the conditional standard error of measurement, this conditional distribution will differ from one part of the score range to another. But unlike the conditional standard error of measurement, it conditions on a quantity that the test user actually knows. It also shows the asymmetry that results from regression effects, which can be large for test takers with scores near the bottom or the top of the distribution.

Our procedure is intended for scores that are equated for difficulty in the test-taker population. Equating in a population implies that the distribution of the equated scores in that population is the same on the two forms (see Braun & Holland, 1982) — or as nearly so as possible, given the discreteness of the scores. However, our procedure requires a stronger assumption: that the equating of forms is correct for the test takers at each ability level (i.e., each true-score level). We do not expect that this assumption will be strictly true for any actual tests. We do expect that it will be approximately true and that the approximation will be close enough for practical purposes. In the words of George E. P. Box, "All models are wrong, but some are useful" (Box & Draper, 1987, p. 424).

Our procedure places no restrictions on the format or the scoring of the test; it is designed to work for any test on which it is possible to estimate the alternate-forms reliability of the scores. The procedure requires, as input, the score distribution and the estimated reliability coefficient in the population of test takers. The alternate form of the test is assumed to differ from the form taken only in those ways that are included as sources of error variance in the reliability estimate. In most large-scale testing programs, those would be the selection of specific items and, if the scoring involves judgment, the selection of the scorers who score a test taker's responses.

*Corresponding author*: S. Livingston, E-mail: SLivingston@ets.org

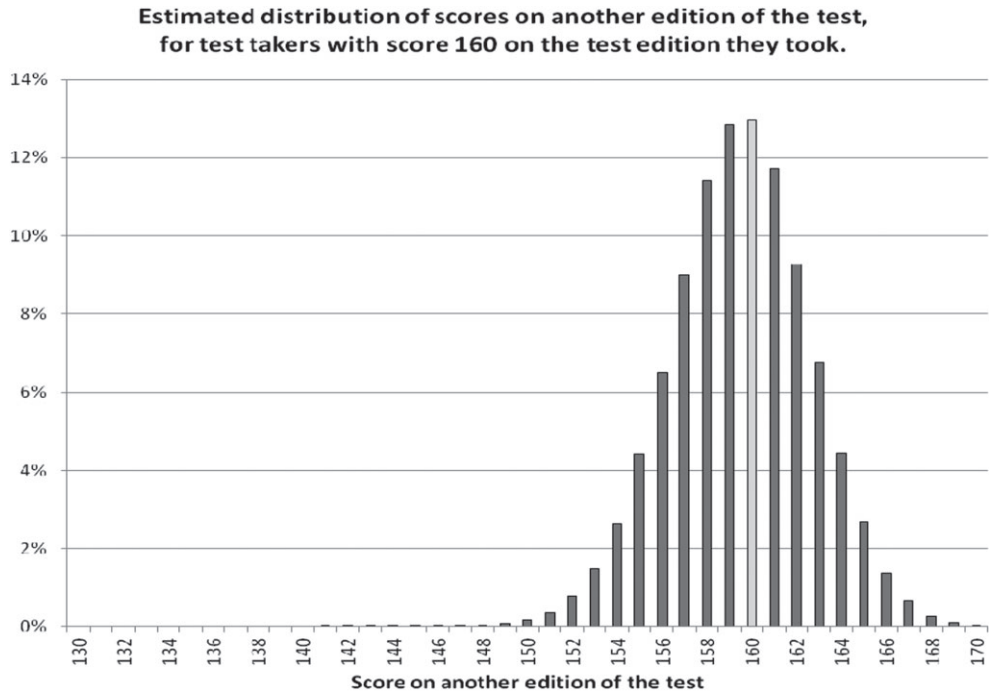


**Figure 1** Histogram of the conditional distribution.

## The Estimation Procedure

Our objective is to estimate the conditional distribution of scores on an alternate form for test takers with any given score on the form they took. For each possible score on the test, there is a conditional alternate-form score distribution to be estimated. Our approach to this task is to estimate all those conditional distributions at once by estimating the bivariate distribution of scores on two forms of the test. Fortunately, we were able to make use of a substantial amount of previous work. Our approach parallels that used by Livingston and Lewis (1995) to estimate the joint distribution of classifications on two parallel forms. To estimate the true-score distribution, we used a procedure developed by Lord (1965).

### Overview of the Procedure

Our approach to the problem requires that the true-score variable be partitioned into many small intervals. We use a partition of 100 intervals of equal size. For the test takers in each true-score interval, we estimate the joint distribution of scores on two forms of the test. Then we sum over the estimated true-score distribution to estimate the joint distribution of scores on two forms of the test in the full test-taker population.

To estimate the joint distribution of scores on two forms for the test takers in a single small true-score interval, we make two assumptions. First, because these test takers all have (very nearly) the same true score, we assume that their scores on the two forms are independent of each other. (This assumption is the basic assumption of classical test theory—that errors of measurement are independent.) Second, because the scores on the two forms are equated, we assume that at every true-score level, the score distributions on the two forms will be the same. In effect, we assume that the equating relationship in the full population will apply to the test takers at every true-score level. With these assumptions, at any single true-score level, the distribution of scores on one form provides all the information we need to determine the joint distribution of scores on two forms.

The procedure can be described in more detail as a sequence of steps:

1. **True-score distribution.** Estimate the true-score distribution as a four-parameter beta distribution using Lord's (1965) method. Partition the estimated distribution into 100 intervals of size .01.
2. **Effective test length.** The effective length of the test is the length of a hypothetical test having the same reliability as the actual test but consisting entirely of independent, equally difficult, 0/1 items. Estimate it by substituting sample

estimates for the parameters in the formula

$$n = \frac{\left(\mu_x - X_{\min}\right)\left(X_{\max} - \mu_x\right) - r\sigma_x^2}{\sigma_x^2\left(1 - r\right)},$$

rounded to the nearest integer, where $\mu_x$ and $\sigma_x^2$ are the mean and variance of the scores, $X_{\max}$ and $X_{\min}$ are the maximum and minimum possible scores, and $r$ is the reliability coefficient (Livingston & Lewis, 1995, pp. 186–187).

3. **Conditional score distributions on the hypothetical test.** At each true-score level (i.e., each interval of the partition in Step 1), generate the distribution of scores on the hypothetical test. This distribution is binomial $(n, p)$, where $n$ is the number of items in the hypothetical test (determined in Step 2), and $p$ is the midpoint of the true-score interval.
4. **Total group score distribution on the hypothetical test.** Estimate the distribution of scores on the $n$-item hypothetical test in the group of all test takers. This distribution is a weighted sum of the binomial distributions at the 100 true-score levels, weighted according to the true-score distribution estimated in Step 1. (Do not discard the separate binomial distributions for the 100 true-score levels. They will be used in a later step of the procedure.)
5. **Continuizing the total-group score distribution on the hypothetical test.** Continuize the score distribution produced in Step 4—the distribution of scores on the $n$-item hypothetical test in the full group—by distributing the frequency at each discrete score uniformly over the 1-unit interval centered on that score.
6. **Cut points for the continuized distribution.** Partition the continuized score scale for the $n$-item hypothetical test into intervals that include the same percentages of the test takers as the score levels on the actual test. There will be a cut point for each score on the actual test (except the highest). To find the cut point for a given score on the actual test, compute its cumulative frequency in the full group of test takers. Then find the point on the score scale of the $n$-item hypothetical test having exactly that cumulative frequency in the continuized distribution from Step 5. That point becomes the upper bound of the interval corresponding to the given score level on the actual test. These cut points, determined for the full group of test takers, can then be applied to any other group of test takers to convert their score distribution on the hypothetical test into an estimated score distribution on the actual test. In particular, the cut points can be applied to the group of test takers at each of the 100 true-score levels.
7. **Applying the cut points at each true-score level.** Return to the set of 100 binomial distributions generated in Step 3—the distribution of scores on the $n$-item hypothetical test for test takers at each of the 100 true-score levels. Continuize each of these 100 distributions by the same procedure used in Step 5, replacing the frequency at each discrete point with a uniform distribution over the 1-unit interval centered on that point. Then partition each of these continuized binomial distributions using the cut points determined in Step 6. The result will be an estimate of the distribution of scores on one form of the actual test for test takers at that true-score level.
8. **A bivariate distribution for the actual test at each true-score level.** Still working separately at each of the 100 true-score levels, use the distribution estimated in Step 7, for one form of the actual test, to estimate the bivariate distribution of equated scores on two forms of the actual test. The key assumption is that for test takers at that single true-score level, the equated scores on two forms of the actual test are statistically independent and identically distributed. Therefore, each cell frequency is the product of the marginal frequencies. This step produces 100 discrete bivariate distributions. Each of those distributions is an estimate of the joint distribution of scores on two forms of the actual test for test takers at one of the 100 true-score levels.
9. **The total-group bivariate distribution.** Sum the 100 bivariate distributions created in Step 8, weighting them according to the true-score distribution estimated in Step 1. The result is an estimate of the bivariate distribution of equated scores on two forms of the actual test in the full group of test takers.

Notice that the partition determined in Step 6 is based on the cumulative frequencies from the distribution of scores on the actual test. Any irregularities in this distribution—particularly, the irregularities arising from sampling variability—will be carried over into the estimated conditional distributions. For this reason, we strongly recommend presmoothing the observed score distribution before using it as input to the estimation procedure.

## Testing the Accuracy of the Estimates

To determine how accurately our procedure estimates the conditional alternate-form score distributions, we needed to use it in situations where those distributions were known. We created three such situations. In each case, we began with item

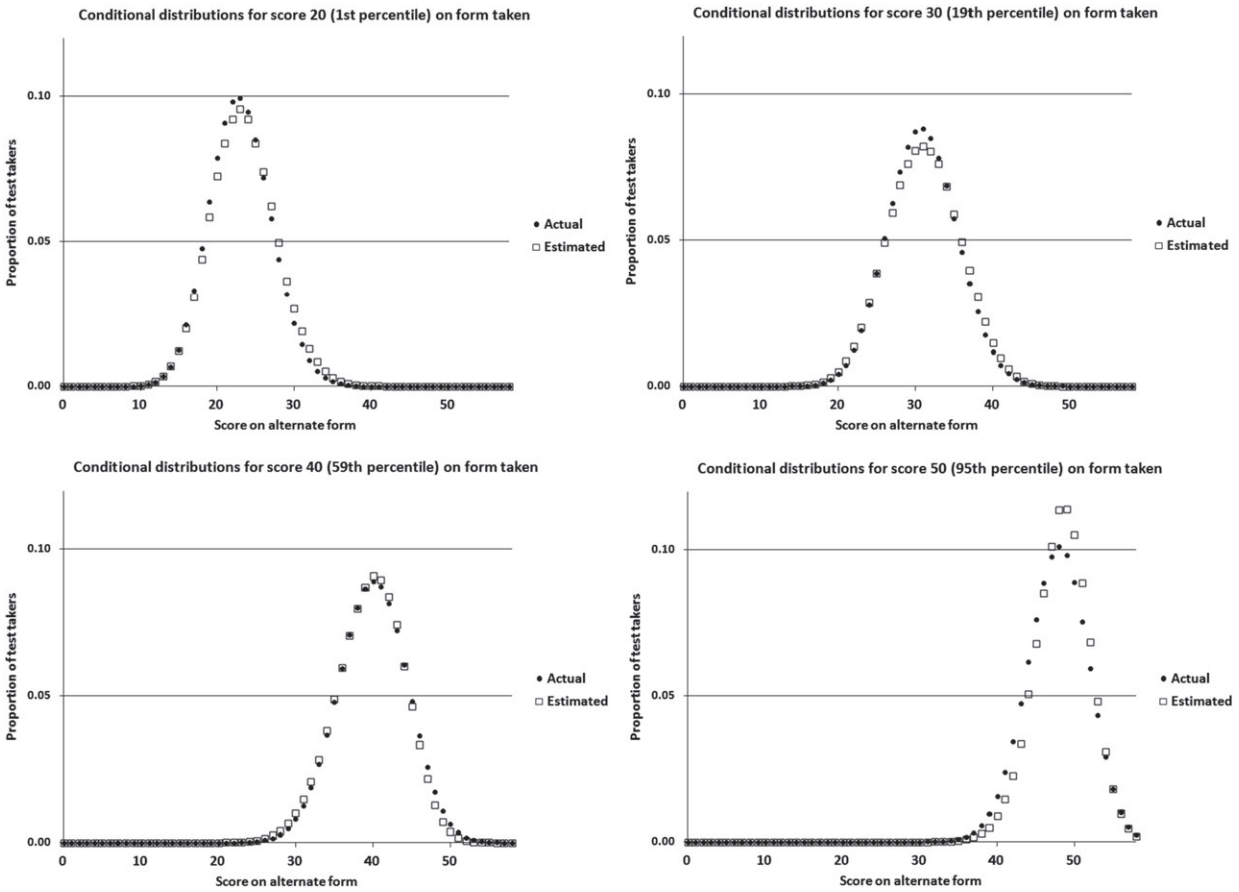


**Figure 2** Actual and estimated conditional distributions for Test MC58.

response data from a test taken by more than 3,000 test takers. These data were not simulated; they were the responses of real test takers to real test questions in a real testing situation. We divided the test into two half-length research forms, similar in content and as nearly equal in difficulty as possible. We computed each test taker's score on Research Form 1 and on Research Form 2 and computed the bivariate distribution of those two scores. This step gave us a bivariate distribution of scores of the same test takers on two forms of a test, but that distribution was not appropriate for testing our estimation procedure for two reasons.

First, the conditional distributions were not smooth. Although the full bivariate distribution included more than 3,000 test takers, the number of test takers in each conditional distribution (i.e., the number at each score level) was much smaller. Consequently, the conditional distributions contained irregularities that would not generalize to another group of test takers nor to the scores of the same test takers on a different pair of test forms. Our estimation procedure is not intended to reproduce these irregularities. To smooth all the conditional distributions in a single operation, we applied a bivariate log-linear smoothing procedure (Holland & Thayer, 1987) to the full bivariate distribution. This smoothing procedure preserves the general shape of the distribution while smoothing away the irregularities.

Second, the bivariate distribution was not symmetric. Our procedure is intended to estimate conditional distributions of equated scores on an alternate form of the test. We assume that any form-to-form differences in difficulty will be removed by equating.[1] For a valid test of our procedure, we needed the bivariate distribution to be symmetric. To create a symmetric bivariate distribution, we transposed the smoothed distribution, creating a mirror-image distribution, and added it to the original distribution. The resulting bivariate distribution represented the joint distribution of equated scores on two forms of the test. The conditional distributions in this bivariate distribution were the distributions to be estimated. For that reason, we refer to them as *actual distributions*.

The input information required by our estimation procedure came from this same smooth, symmetric bivariate distribution. For the observed distribution of scores on the form taken, we used the marginal distribution (which was the

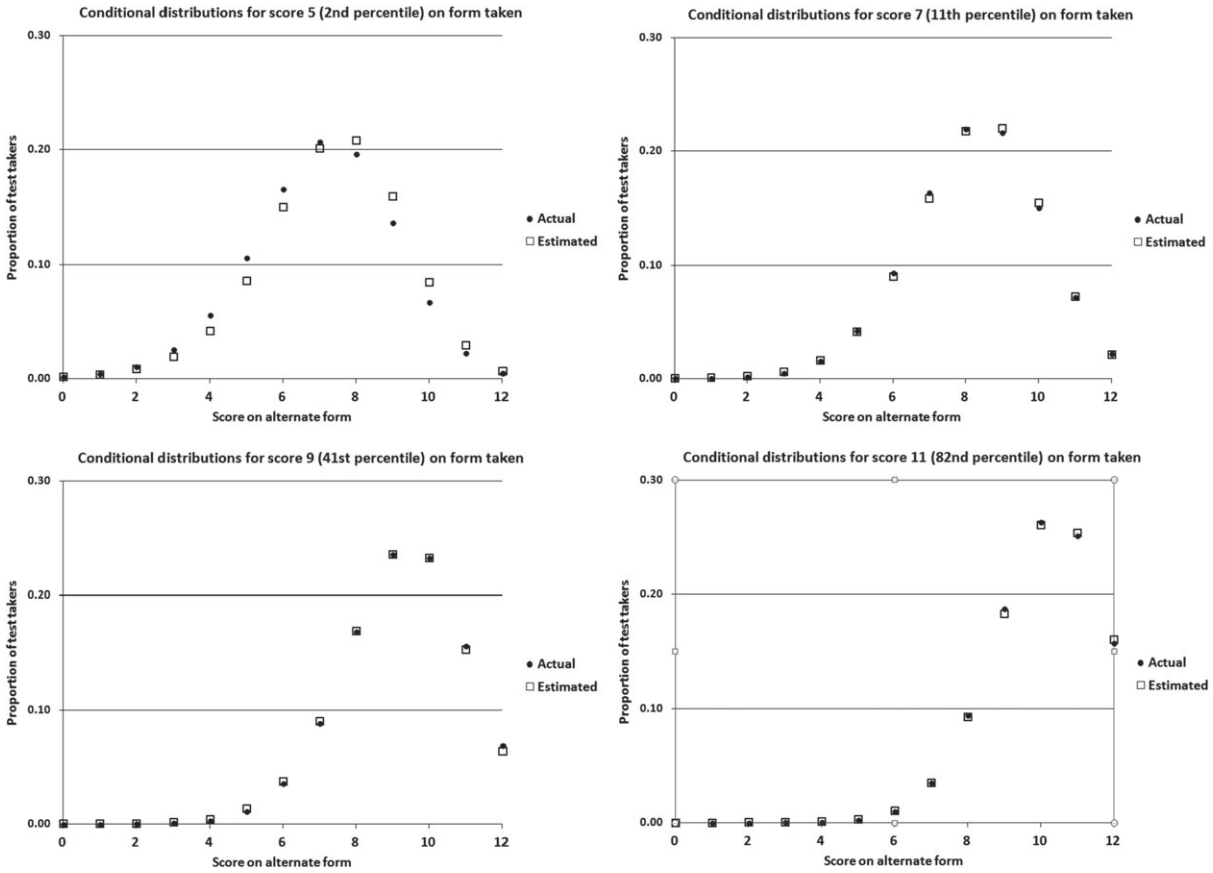


**Figure 3** Actual and estimated conditional distributions for Test CR12.

same for both scores in the bivariate distribution). For the reliability coefficient, we used the correlation in the bivariate distribution. We entered this information into our estimation procedure to estimate the conditional alternate-form score distributions for several score levels on the form taken. We compared these estimated conditional distributions with the actual conditional distributions computed directly from the bivariate distribution.

We conducted this experiment with three very different tests.

1. Test MC58 contained 58 multiple-choice items in each form. The reliability of the scores was .84.
2. Test CR12 contained six constructed-response items in each form. Each item was scored 0, 1, or 2; the possible range of total scores was 0 to 12. The reliability of the scores was .46.
3. Test MCCR36 contained 24 multiple choice items and six constructed-response items in each form. Each constructed-response item was scored 0, 1, or 2; the possible range of the total scores was 0 to 36. The reliability of the scores was .53.

As might be expected, the accuracy of the estimated conditional distribution changed gradually from one score level to the next. Figures 2 – 4 show the results for a few widely spaced score levels on each test.

Figure 2 shows the actual and estimated alternate-form score distributions on Test MC58 (reliability .84) for test takers with four different scores on the test form taken. At score 20 on the form taken (the 1st percentile), the estimated alternate-form score distribution was fairly accurate but showed slightly more regression to the mean than the actual distribution did. The largest difference between the cumulative distribution functions (CDFs)[2] was .039. At score 30 on the form taken (the 19th percentile), the estimated alternate-form score distribution showed slightly too much spread; the largest difference between the CDFs was .028. At score 40 on the form taken (the 59th percentile), the estimated alternate-form score distribution was fairly accurate but showed slightly too much regression to the mean; the largest difference between the CDFs was .024. At score 50 on the form taken (the 95th percentile), the estimated alternate-form score distribution was too peaked and did not show enough regression to the mean; the largest difference between the CDFs was .075.

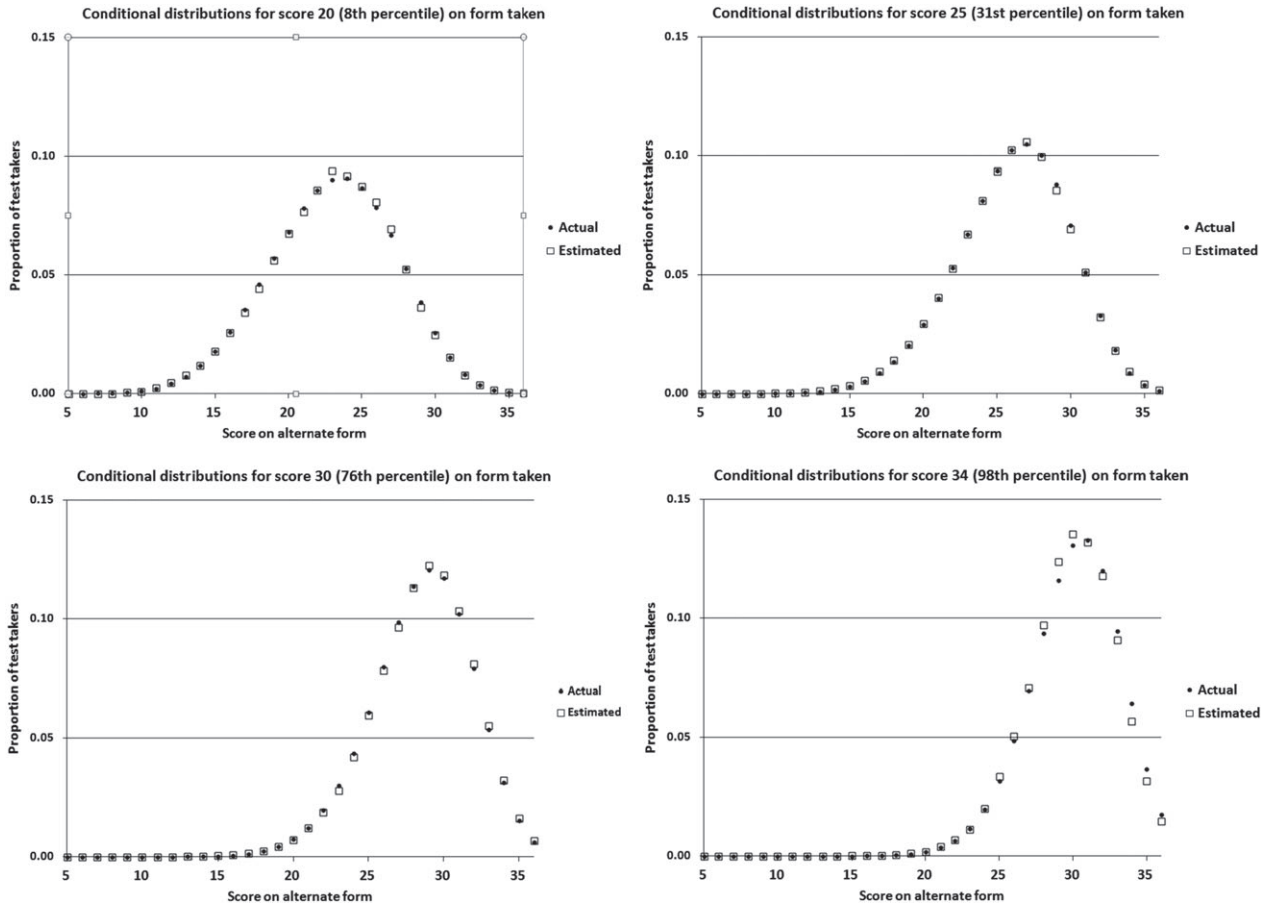


**Figure 4** Actual and estimated conditional distributions for Test MCCR36.

Figure 3 shows the actual and estimated alternate-form score distributions on Test CR12 (reliability .46), for test takers with four different scores on the test form taken. At score 5 on the form taken (the 2nd percentile), the estimated alternate-form score distribution showed more regression to the mean than the actual distribution did; the largest difference between the CDFs was .062. At scores 7 (the 11th percentile), 9 (the 41st percentile), and 11 (the 82nd percentile) on the form taken, the estimated alternate-form score distributions were quite accurate. At each of these score levels, the largest difference between the CDFs of the estimated and actual distributions was less than .01. At score 12 (not included in Figure 3), which was the 95th percentile and also the highest possible score, the estimated alternate-form score distribution did not show enough regression to the mean; the largest difference between the actual and estimated CDFs was .025.

Figure 4 shows the actual and estimated alternate-form score distributions on Test MCCR36 (reliability .53) for test takers with four different scores on the test form taken. At scores 20 (the 8th percentile), 25 (the 31st percentile), and 30 (the 76th percentile), the estimated alternate-form score distributions were quite accurate; the largest difference between the actual and estimated CDFs was .01 or less. For score 34 on the form taken (the 98th percentile), the estimated alternate-form score distribution showed slightly too much regression to the mean; the largest difference between the CDFs was .023.

## Discussion

Some test users use test scores as part of a formal decision process, creating and then applying a decision rule. Others use the scores as information about each test taker, making decisions without a formal decision rule. In either case, it is useful to know how well a test taker's score is likely to generalize beyond the edition of the test that the test taker happened to take.

From the test user's point of view, the essential question of test score reliability is, "How similarly would this test taker have performed on another edition of the test, containing different questions testing the same types of knowledge and skills?" This question can only be answered probabilistically. The relevant probabilities are those that make up the

conditional distribution of scores on an alternate form of the test given the test taker's score on the form taken. The procedure described in this paper provides a way to estimate those probabilities.

We tested the procedure by creating bivariate score distributions for two nonoverlapping half-length test forms, using the responses of real test takers to the items on a single full-length test form. We conducted this experiment with three different tests having different formats and different degrees of reliability. The univariate distributions were similar to those encountered in most large-scale testing situations, although the reliability coefficients for these half-length tests were lower. The results indicated that the procedure worked well for these data sets.

To probe the limits of the procedure, we also tried the procedure with some artificial data sets created to present unusual score distributions. We found that the procedure did not work well in situations where the true-score distribution was either extremely peaked or bimodal. Its poor performance under those conditions may be the result of the mathematical model used to estimate the true-score distribution (from Lord, 1965). A more flexible mathematical model for the true-score distribution (e.g., Lord, 1969) might produce better results in these unusual cases.

The current procedure appears to work well enough to justify its use with unimodal distributions when there is no reason to suspect that a large percentage of the test takers have true scores in a small portion of the score range. The estimated conditional distributions will make it possible for testing agencies to describe the reliability of the scores in terms of simple probability statements about a test taker's performance on an alternate form of the test.

## Notes

1 Score equating can make forms equally difficult for a group of test takers but not for each individual test taker (unless the reliability of the alternate forms is 1.00).

2 This quantity is the D-statistic in the Kolmogorov–Smirnov test (see, e.g., Siegel & Castellan, 1988, p. 145).

## References

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York, NY: Wiley.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). Academic Press: New York, NY.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Research Report No. RR-87-31). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00235.x

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, *30*, 239–270.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). *Psychometrika*, *34*, 259–299.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York, NY: McGraw-Hill.

**Suggested citation:**

Livingston, S. A., & Chen, H. H. (2015). *Estimating conditional distributions of scores on an alternate form of a test* (ETS Research Report No. RR-15-18). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12066

**Action Editor:** James Carlson

**Reviewers:** Steven Holtzman and Edward Kulick