A Practical Approach to Balancing Time Demands across Test Forms[1]

David J. Scrams and Russell W. Smith, Alpine Testing Solutions

Four simple algorithms based on descriptive item response time statistics are evaluated for pre-equating test forms in terms of time requirements. Results show that pre-equating forms based on item response time means or medians performs as well as a more complex distribution-based form-construction algorithm. While the ease of implementation supports operational use, the realities of shallow item pools and overarching importance of traditional item characteristics is discussed.

Test sponsors often administer multiple forms of an exam either concurrently or consecutively. In these cases, industry standards require that candidates administered different forms receive comparable and fair treatment (AERA, APA, & NCME, 1999). Psychometricians and test development practitioners operationalize this standard in part by ensuring that forms are constructed to be as parallel as possible based on available response data and item characteristics. With the advent and proliferation of computer-based testing, response-time data has become more readily available for operational uses, and the possibility of including them in form construction seems both logical and desirable.

While literature on response times in testing is by no means extensive, a fair amount of work has focused on establishing theoretical models to describe the relationship between response time and examinee ability (Roskam, 1997; Thissen, 1983; van der Linden, 2006; Verhelst, Verstralen, & Jansen, 1997). These theoretical efforts laid the foundation for applied work in areas such as speededness and adaptive item selection (Schnipke & Scrams, 1997; van der Linden, Scrams, & Schnipke, 1999; van der Linden, 2008). Drawing heavily on the mathematical models established in the purely theoretical work, these efforts maintained a theoretical feel while approaching practical problems.

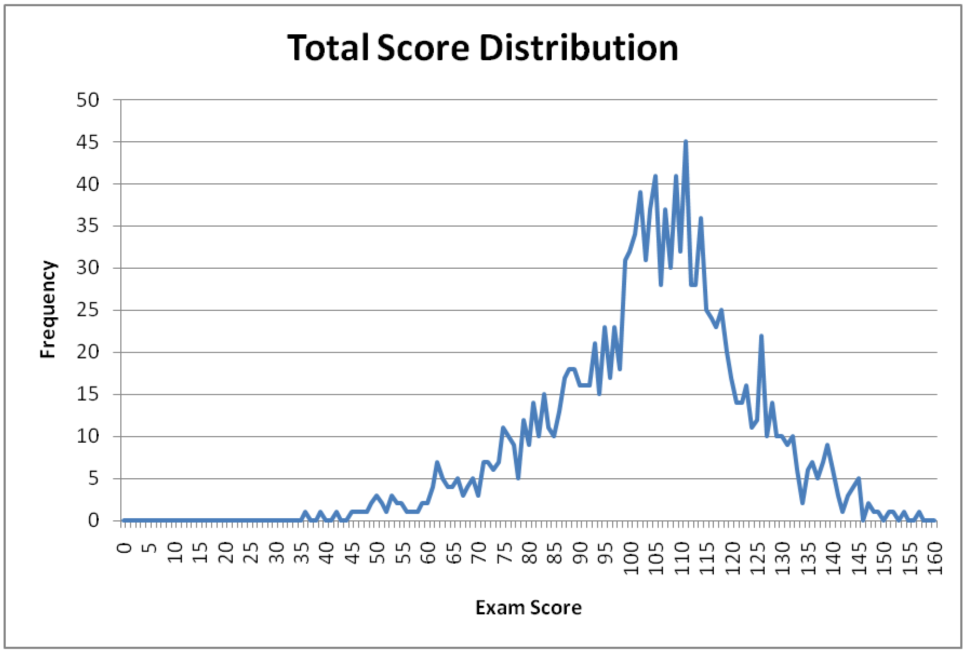
Over the same period of time, computerized test administration has become increasingly more common throughout the testing industry. Applied psychometricians are now commonly in possession of response-time data and interested in straightforward methods for incorporating these data into standard test-development practices. The current work is firmly focused on providing tools that fit this niche. The specific goal is to compare easily implemented algorithms for pre-equating time constraints in the construction of parallel exam forms. In this context, “easily implemented” was operationalized as limiting the algorithms to those based on simple descriptive statistics such as item response time means, medians, and standard deviations. Algorithms requiring unique calibrations such as those explored by Thissen (1983) and others (Roskam, 1997; van der Linden, 2006; Verhelst, Verstralen, & Jansen, 1997) were not considered.

[1] Paper presented at the 2010 annual meeting of the National Council on Measurement in Education, Denver, CO.

Method

Operational data from a large-scale professional licensure exam were used. This exam was administered as 2 160-item forms, but data from only one form was used in the present work. Data for this form were available from 1355 candidates who earned an average exam score of 103.67 (SD=18.92) with an average total test time of 108.6 minutes (SD=29.1). Exam score and time distributions are provided in Figures 1 and 2. There was no indication of exam score being related to total test time as shown in Figure 3.

Figure 1. Total Score Distribution



Data Handling

The operational data were manipulated to simulate a typical pre-test scenario. The 200 candidates with the earliest administration dates were selected from the pool of 1355. Data from these candidates were set aside and treated as a pre-test sample who received a single 160-item form. The remaining 1155 candidates were treated as operational candidates and used to evaluate the performance of the various form-construction algorithms.

Before applying any form-construction algorithms, exploratory analyses of the 200 pre-test candidates were performed. Descriptive statistics showed similar characteristics to the overall pool in terms of test score (Mean=104.16, SD=17.11) and total time (Mean=112.00, SD=29.71). Several outliers were also identified who took substantively less time to answer items. After considering the overall distributions, candidates who answered any item in less than 10 seconds were removed from further analyses. This reduced the pre-test sample to 189.

Figure 2. Total Exam Time Distribution

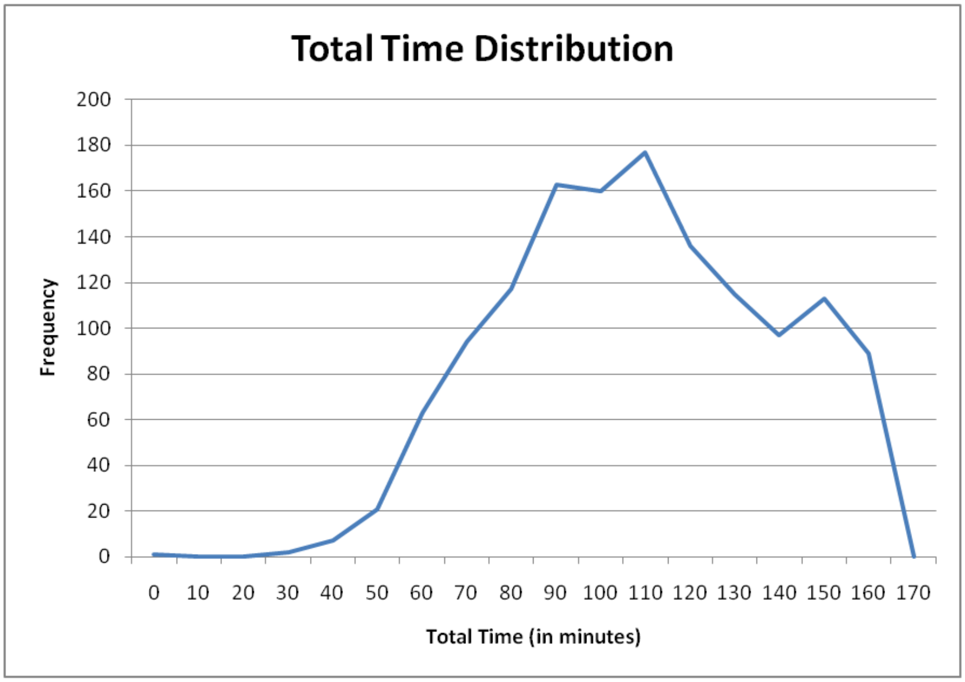
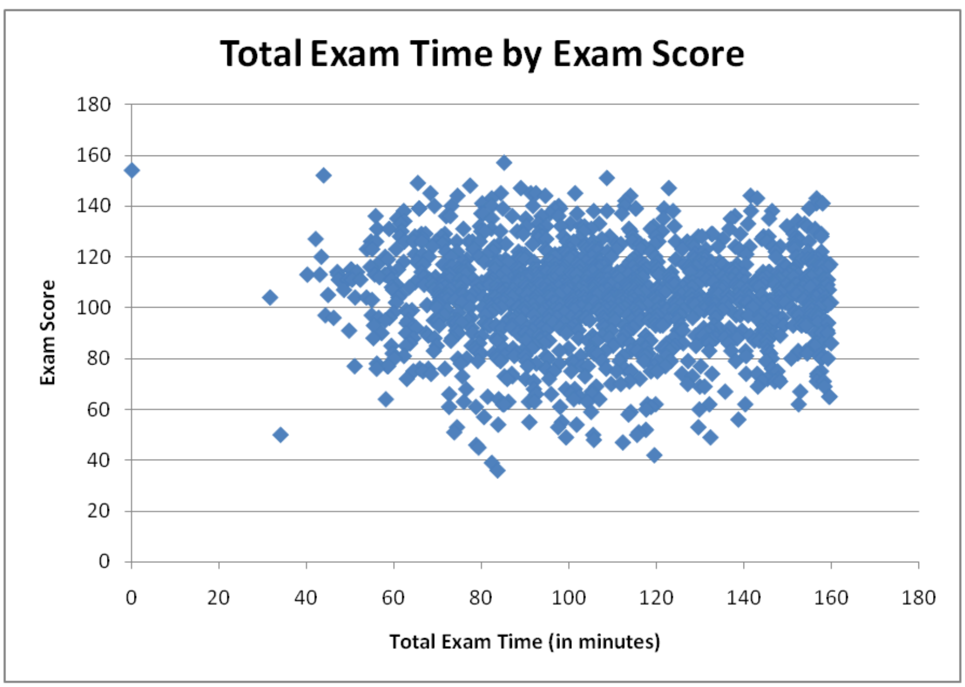


Figure 3. Total Exam Time by Exam Score



Algorithms

Four simple algorithms based on descriptive item statistics were considered. The simplest two algorithms select items based purely on measures of central tendency: mean and median. The remaining two algorithms add standard deviation as a second consideration, so the four algorithms are mean only, median only, mean plus standard deviation, and median plus standard deviation. Items are matched across forms based on Euclidean distance which simplifies to the square root of the squared difference in means (medians) in the simpler two algorithms. In other words, as the first form was being built, a corresponding item was selected for the second form such that the item was most similar in terms of the relevant descriptive statistic or statistics.

Two additional algorithms were applied for reference purposes. To set a baseline, a random algorithm was used that simply selected an item at random with no consideration of response time data. Second, a distribution-based algorithm was used. This algorithm selected an item based on the root mean squared difference between the cumulative distribution functions of the two items under consideration. The difference was evaluated across 41 points along the cumulative distribution functions. This algorithm should arguably be more accurate than the ones based on simple statistics given the larger amount of information used in its calculation, but it's more complicated than desired by most practitioners. The distribution-based algorithm is included here as a reference point against which the performance of the simpler algorithms can be evaluated.

Although the focus for this work is on pre-equating time requirements across forms, a simple set of content constraints were included. This involved 5 content categories and associated coverage specifications. Forms were not built to meet difficulty or other psychometric constraints although the extension to that more typical situation is addressed in the discussion.

Simulations

For each of the 6 algorithms (4 algorithms of interest and 2 reference algorithms), the pre-test data were sampled 1000 times according to the following process:

1. Select a random item for the reference form.
2. Search the remaining items for the item that most closely matches the reference item according to the relevant algorithm.
3. Repeat process with a new random item selected for the reference form until an entire 2-form, 50-item exam has been constructed.

This process was used to construct 1000 random 2-form exams for each algorithm. There is no overlap between the forms as overlap would increase the comparability of the response-time characteristics. The simulated builds can be considered the non-overlapping item sets for two overlapping forms with no loss of generality.

Results

Performance of each algorithm was evaluated for each set of simulated forms by comparing the cumulative distribution functions of total test time for the "operational" candidates. A root mean squared difference was calculated across 51 points along the cumulative distribution functions, although there were no differences past the 45th point. The distributions of the resulting root mean squared differences are shown in Figure 4. As expected, the random algorithm performed less well than all other algorithms. Unexpectedly, however, performance of the simpler algorithms is indistinguishable from one another and from that of the more complex distribution-based selection algorithm. Descriptive statistics for the overall performance of each algorithm are provided in Table 1.

Figure 4. Root Mean Squared Difference Distributions

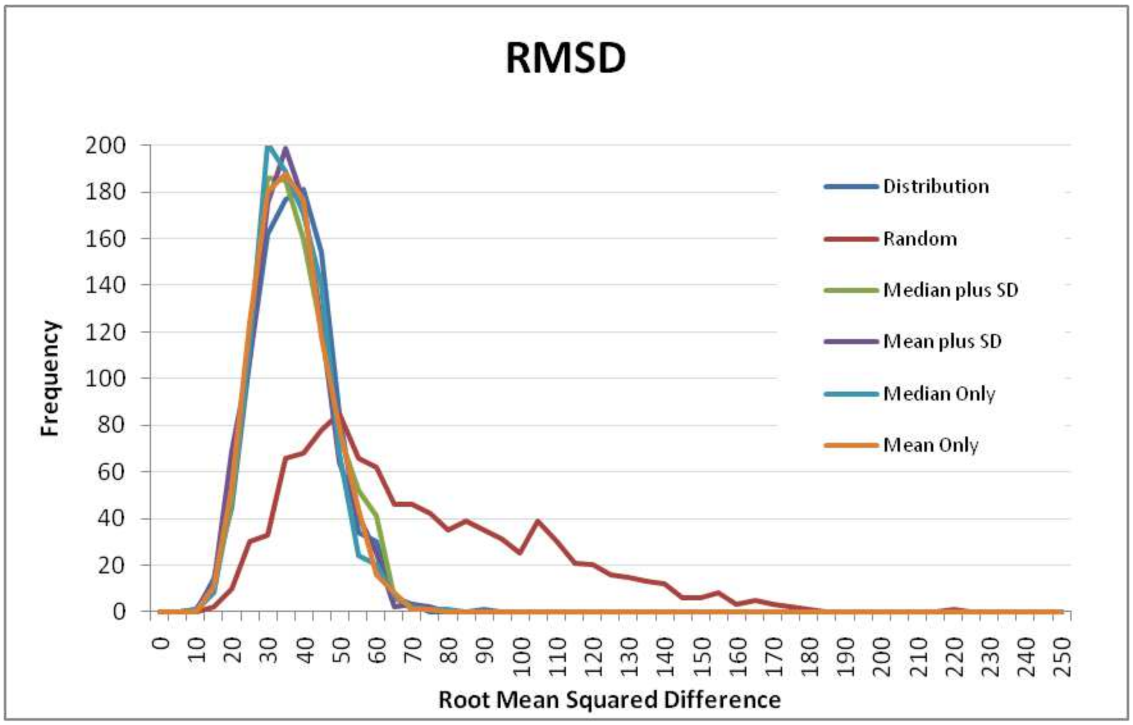
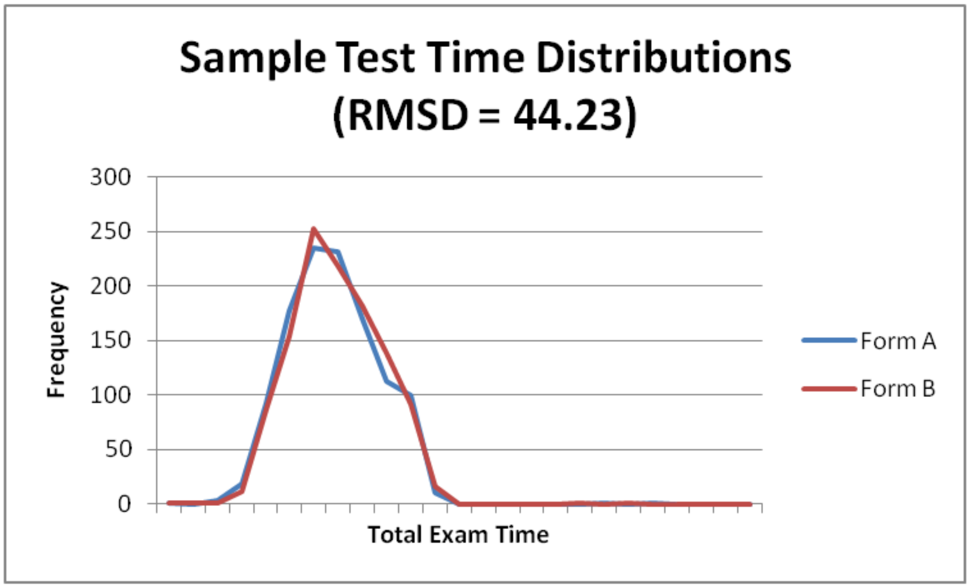


Table 1. Descriptive Statistics for Root Mean Squared Differences

Algorithm	Mean RMSD	Median RMSD	SD of RMSD
Distribution	35.22	34.75	10.22
Random	68.88	59.98	33.87
Median plus SD	34.97	33.91	10.55
Mean plus SD	33.98	33.17	10.18
Median only	34.04	33.08	9.74
Mean only	34.17	33.50	9.99

In order to provide some level of context for the size of RMSDs, Figure 5 shows a random solution from the Mean Only simulations. This solution resulted in a slightly higher than average RMSD (44.23).

Figure 5. Sample Test Time Distributions from the Mean Only Simulations (RMSD=44.23)



Discussion

The current work is a straightforward approach to a simple problem facing operational psychometricians and test-development practitioners. With item response times readily available, why not incorporate them into form-construction algorithms in order to provide comparable time requirements to examinees receiving different forms? The essential issue has been addressed in the more complex situation of adaptive item-selection algorithms primarily by van der Linden (2008; van der Linden, Scrams, & Schnipke, 1999), but the current goal was to provide a simpler solution for the simpler case of fixed-form test administration.

The results were not entirely expected. Selecting items based on equivalence of the entire response-time distribution was expected to outperform simpler algorithms focused only on one or two item statistics, but this was not borne out by the current data. The simpler algorithms performed equally well. This result is promising.

That said, there are numerous limitations to the current work. First, the data were not actually obtained as separate pre-test and operational samples, so differences between pre-test and operational administrations are not reflected in the data. That is, to the extent that a program's operational administrations vary from their pre-test administrations or to the extent that candidates behave differently in the two situations, the current data might not completely mirror operational results. While this is a limitation of the current work, it's arguable that any such impacts would likely be program specific, so applicability of pre-test response-time data as estimates of operational response-time data would need to be evaluated specifically for any program interested in pursuing the approach.

In the same vein, the current work did not incorporate traditional psychometric build constraints such as item difficulty or form reliability. This is an important limitation and should not be underestimated, but much like the previous limitation, this one is heavily dependent on the specific item pool. A highly constrained item pool may be so limited in terms of meeting parallelism in item difficulty, that response-time constraints cannot also be balanced. On the other hand, a deep item pool may easily support additional constraints. While a reasonable argument can be made for item difficulty and other similar psychometric considerations to trump parity in expected time requirements, there is likely still value in considering the probable time discrepancy among potential forms in a constrained situation.

Even apart from these important considerations, there is cause for caution in interpreting the current results too broadly. In the same way that an item pool may be shallow in terms of meeting parallelism in item difficulty, a pool may be shallow in meeting response-time parallelism. The absolute quality of parallelism obtained in the current work should not be the focus of interpretation. What's of importance, rather, is the relative quality of the simple algorithms compared to the more complex distribution-based algorithm and the baseline random-selection algorithm. There is no evidence in the current data to suggest that the more complex algorithm is required. The simpler algorithms resulted in comparable solutions.

The approach is simple, but applying the algorithms to additional datasets is highly encouraged. Programs differ from one another in myriad ways, and some program-specific characteristics are very likely to impact the results and limit their applicability. It is possible that a more realistic dataset might tease apart differences in algorithm performance.

References

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. J., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In Mills, C. N., Potenze, M. T., Fremer, J. J., & Ward, W. C. (eds.). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates.

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer-Verlag.

Thissen, D. (1983). Timed testing: An approach using item response testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179–203). New York: Academic Press.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Beahvioral Statistics*, *31*, 181–204.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*; *23*; 195–210.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–186). New York: Springer.