

Test Design and Speededness

Wim J. van der Linden

CTB/McGraw-Hill

A critical component of test speededness is the distribution of the test taker's total time on the test. A simple set of constraints on the item parameters in the lognormal model for response times is derived that can be used to control the distribution when assembling a new test form. As the constraints are linear in the item parameters, they can easily be included in a mixed integer programming model for test assembly. The use of the constraints is demonstrated for the problems of assembling a new test form to be equally speeded as a reference form, test assembly in which the impact of a change in the content specifications on speededness is to be neutralized, and the assembly of test forms with a revised level of speededness.

Introduction

The notion of speededness in testing refers to an interaction between three important factors: the cognitive speed at which the test taker works during the test, the amount of labor required by the items, and the time limit on the test. A test is more speeded when a test taker has to work faster, answering the items requires more labor, and/or the time limit is tightened.

As the speed of the test taker is one of these factors, it actually is incorrect to refer to the speededness of a *test*. For a given time limit, the same test can demonstrate different degrees of speededness for different test takers. A case in point is a test taker with a disability not related to the proficiency that is measured, which forces him/her to work slower than others; for instance, an impaired vision that requires taking the test in braille. An important problem when trying to accommodate such test takers is how to find the adjustment of the time limit necessary to make the test equally speeded as for regular test takers (e.g., Stretch & Osborne, 2005).

Also, although it is common parlance to refer to testing as being either speeded or unspeeded, speededness always is a matter of degree. In fact, attempts to quantify the degree of speededness have a long history, dating back as early as the work by Gulliksen (1950) and Cronbach and Warrington (1951), who used the number of unattempted items and the number-correct scores for versions of the same test form with and without the time limit as input for their measures, respectively. Interestingly, the tradition begun by the former still continues in the criterion of test speededness currently used in the testing industry, which qualifies a test as unspeeded when at least 80% of the test takers complete all items and all test takers complete at least 75% of the items. For a more extensive review of the earlier history of the attempts to quantify test speededness, see Rindler (1979).

The three basic factors that drive test speededness are all present in the definition of the degree of test speededness as the *probability* of a test taker running out of time

before completing the test proposed in van der Linden (2011). The calculation of the probability requires a statistical model for the response-time distributions on the test items. I first review the definition of the probability and then discuss the model used in the current research.

Let T_i be the random variable for the response time by a test taker on items $i = 1, \dots, n$ in the test. The total time for the test taker is the sum of the response times on the items; that is, $T_{\text{tot}} = \sum_{i=1}^n T_i$. We use $F_{T_{\text{tot}}}(t)$ to denote the cumulative distribution function of T_{tot} . For a given time limit t_{lim} , the probability π of the test taker running out of time is

$$\begin{aligned} \pi &= \Pr \left\{ \sum_{i=1}^n T_i > t_{\text{lim}} \mid \tau, \alpha, \beta \right\} \\ &= 1 - F_{T_{\text{tot}}}(t_{\text{lim}} \mid \tau, \alpha, \beta), \end{aligned} \quad (1)$$

where τ is the test taker's speed, $\beta = (\beta_1, \dots, \beta_n)$ is the vector of item parameters for the amount of labor required by the items, and $\alpha = (\alpha_1, \dots, \alpha_n)$ is the vector of the item parameters that controls the variances of the response-time distributions on the items in the model below.

In statistics, expressions as in (1) are known as survival functions. However, to avoid the labeling of the event of a test taker running out of time as a "survival", I will treat (1) as an expression for the risk π of running out of time as a function of the test taker's speed τ for a test with item parameters (α, β) and a time limit t_{lim} , and refer to it as the risk function for the test. Throughout this paper, I will assume that the item parameters have been estimated with enough precision to treat them as known. As indicated below, for computerized tests with automatically logged response times, the estimation involves only a minor extension of regular item pretesting and calibration.

Depending on which quantities are known and unknown, risk functions can be used in three different ways. First, obviously, they can be used to evaluate the risk π for a given test with item parameters (α, β) and time limit t_{lim} ; that is, evaluate its degree of speededness. One possible application is an empirical check on a subjective claim that a test taker ran out of time because the test was too speeded. The check could involve estimating the speed parameter for the test taker from the portion that was completed. The estimate could then be used, for instance, for a comparison between the test taker's speed and the speed of relevant other test takers or, along with the item parameters for the full test, to project the time the test taker would have needed to complete the test. The ML estimate of τ has a simple, closed form (see below). Another possible application is to evaluate the risk before the test is administered for a range of speed levels known to be reasonable for the testing program. But if the goal is to avoid undesirable levels of speededness, the following two types of use of (1) are actually more attractive.

Second, risk functions can be used to calculate the time limit t_{lim} required for a given test with item parameters (α, β) to guarantee a predetermined level of risk π . The only thing necessary is reformulating (1) into t_{lim} as a function of $1 - \pi$; that is,

as the quantile function

$$t_{\text{lim}} = F_{T_{\text{tot}}}^{-1}(1 - \pi \mid \tau, \alpha, \beta) \quad (2)$$

for the total-time distribution. As a population of test takers typically has a common time limit, the limit has to be calculated for a minimum acceptable speed level, τ_0 . When speed is not intended to be measured by the test (i.e., τ is to be treated as a nuisance parameter), τ_0 should be chosen low enough to include all test takers (except for a few possible outliers). When τ is an intentional parameter, however, the choice becomes judgmental; τ_0 should then represent intended behavior on the test. For a few methods for choosing τ_0 for tests with speed as an intentional parameter as well as empirical examples of the quantile functions that should be used, see van der Linden (2011).

Finally, we may select a new test form to realize a desired level of risk π for a given time limit t_{lim} . This actually is the most practical use of (1) because it enables a testing agency to maintain a fixed degree of speededness for a fixed time slot across its test forms. An application of the idea to use item selection to control the speededness of an adaptive test for a given time slot is reported in van der Linden (2009a). The application capitalizes on the opportunity provided by adaptive testing to update an estimate of the test taker's speed during testing and use the updates to select the remaining items for the test to meet the time limit.

The current research was motivated by the wish to extend the idea of assembling a test form with a given degree of speededness to fixed test forms. For fixed forms, there is no opportunity to control speededness through item selection during the test. However, as shown below, there actually is no need to. The response-time model we will use has the attractive possibility of controlling the speededness at all possible levels of speed simultaneously when selecting the items for the test form.

Time Distributions

A key element in (1) is the distribution of the total time on the test. I focus on the distribution that is obtained for the lognormal model for the response time by a test taker on an item.

Item Model

The distribution of the response time for a test taker with speed $\tau \in [- \infty, \infty]$ on item i is modeled as the lognormal density

$$f(t_i; \tau, \alpha_i, \beta_i) = \frac{\alpha_i}{t_i \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} [\alpha_i (\ln t_i - (\beta_i - \tau))]^2 \right\}, \quad (3)$$

with parameters $\beta_i \in [- \infty, \infty]$ for the time intensity or amount of labor required by item i and $\alpha_i \in (0, \infty]$ for its discriminating power. The model follows directly from a fundamental equation for response-time modeling derived from the definition of speed on a test as the rate of change in the amount of labor performed on the items with respect to time (van der Linden, 2009b). Bayesian estimation of

the parameters with Gibbs sampling from their posterior distributions is described in van der Linden (2006). But rather than calibrating the items separately under the model in (3), it is statistically more efficient to calibrate them jointly with respect to all parameters in the response and response-time models using the hierarchical modeling framework in Klein Entink, Fox, and van der Linden (2009) and van der Linden (2007).

For a test with calibrated items, the ML estimate of the speed parameter in the model can be shown to be equal to

$$\hat{\tau} = \frac{\sum_{i=1}^n \alpha_i^2 (\beta_i - \ln t_i)}{\sum_{i=1}^n \alpha_i^2}, \quad (4)$$

which has the nice interpretation of $\hat{\tau}$ as the precision-weighted average of the differences between the time intensities of the items and the logarithm of the test taker's response times on them.

The fit of the lognormal model can be checked, for instance, using techniques based on Bayesian predictive checks, the deviance information criterion (DIC), Bayes factors, and Bayesian residuals. For the application of these techniques to the model, see Klein Entink, Fox, and van der Linden (2009), and van der Linden (2006). The techniques have been used to check the fit of the model to the empirical response times for a variety of tests, including the *Armed Services Vocational Aptitude Test Battery* (ASVAB) (van der Linden, 2006), the *GMAT Quantitative Test* (van der Linden & Guo, 2008), Raven's *Advanced Progressive Matrices* (APM) test (Goldhammer, 2010), another figural matrix test (Klein Entink, Kuhn, Hornke, & Fox, 2009), a test of Dutch as a foreign language, a test of quantitative and scientific reasoning proficiencies for college students (NAW-8), and a neurosis scale in a personality questionnaire (Klein Entink, Fox, & van der Linden, 2009), as well as data sets from the *CPA Uniform Examination* (Chuah & van der Linden, 2008; Finger & Chuah, 2009; van der Linden, Breithaupt, Chuah, & Zhang, 2007). For each of the applications, except for an occasional item, the checks showed fit that was excellent for all practical purposes.

Total-Time Distribution

The density of the total time on a test with response-time distributions on the items described by the lognormal model is the n -fold convolution of the densities in (3). Its derivation requires the calculation of a different integral for each item in the test, but even for the case of two items only approximate solutions are known (e.g., Naus, 1969). For larger test lengths, the convolution integrals become completely intractable. However, all we need for the test-assembly methods proposed below are the first three cumulants of the distribution (mean, variance, and an expression related to its skewness), along with an approximation to its density that enables us to inspect the results graphically. Because they have already been derived in van der Linden (2011), I review the results only briefly.

From a general expression for the moments of the standard lognormal distribution in Equation 14 (e.g., Kotz & Johnson, 1985, pp. 134–136), the first three cumulants of the response-time distribution on item i for the model in (3) can be derived as

$$\mathcal{E}(T_i) = \exp(-\tau) \exp(\beta_i + \alpha^{-2}/2), \quad (5)$$

$$\mathcal{E}[T_i - \mathcal{E}(T_i)]^2 = \exp(-2\tau) \exp(2\beta_i + \alpha^{-2})[\exp(\alpha^{-2}) - 1], \quad (6)$$

$$\mathcal{E}[T_i - \mathcal{E}(T_i)]^3 = \exp(-3\tau) \exp(3\beta_i + 3\alpha^{-2}/2)[\exp(3\alpha_i^{-2}) - 3 \exp(\alpha_i^{-2}) + 2]. \quad (7)$$

Each of these expressions contains a factor depending only on the test taker and another depending only on the item. This feature allows us to define new item parameters

$$q_i \equiv \exp(\beta_i + \alpha_i^{-2}/2), \quad (8)$$

$$r_i \equiv \exp(2\beta_i + \alpha_i^{-2})[\exp(\alpha_i^{-2}) - 1], \quad (9)$$

$$s_i \equiv \exp(3\beta_i + 3\alpha^{-2}/2)[\exp(3\alpha_i^{-2}) - 3 \exp(\alpha_i^{-2}) + 2]. \quad (10)$$

As the response times on the items are assumed to be conditionally independent given τ (“local independence”) and the cumulants are known to be additive for independent random variables, we can use these new item parameters to derive the following simple expressions for the first three cumulants of the total-time distribution on a test:

$$\mathcal{E}(T_{\text{tot}}) = \exp(-\tau) \sum_{i=1}^n q_i, \quad (11)$$

$$\mathcal{E}[T_{\text{tot}} - \mathcal{E}(T_{\text{tot}})]^2 = \exp(-2\tau) \sum_{i=1}^n r_i, \quad (12)$$

$$\mathcal{E}[T_{\text{tot}} - \mathcal{E}(T_{\text{tot}})]^3 = \exp(-3\tau) \sum_{i=1}^n s_i. \quad (13)$$

Although the cumulants are easy to calculate for the lognormal response-time model, the actual shape of the total-time distribution is unknown. In order to inspect the

results for the test-assembly method presented in this paper graphically, I approximated the unknown shape by a member from the standard lognormal family, which has density

$$f(x; \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{\ln x - \mu}{\sigma} \right)^2 \right\}, \mu, \sigma > 0. \quad (14)$$

The approximation is based on Fenton’s (1960) method of cumulant matching. For two versions of the method—one for the first two cumulants and another for the second and third—the approximating standard lognormal distributions are derived in van der Linden (2011). Both versions perform equally well, but the former has closed-form expressions for the parameters μ and σ^2 , which follow directly from the item parameters introduced in (8)–(9) as

$$\mu = -\tau + \ln \left(\sum_{i=1}^n q_i \right) - \ln \left(\frac{\sum_{i=1}^n r_i}{\left[\sum_{i=1}^n q_i \right]^2} + 1 \right) / 2 \quad (15)$$

and

$$\sigma^2 = \ln \left(\frac{\sum_{i=1}^n r_i}{\left[\sum_{i=1}^n q_i \right]^2} + 1 \right). \quad (16)$$

An example of the fit of the approximation is provided in Figure 1, which shows both an empirical histogram for the distribution of T_{tot} for a test taker with speed $\tau = .0$ on the 78-item reference test used in the empirical examples below and the lognormal in (14)–(16) fitted to the histogram. The histogram is the result of 100,000 simulated test takers working at $\tau = .0$, with the sampled response times on the items added to obtain their total time on the test. The sample size was large enough to adequately represent the true distribution; for instance, its mean was equal to 5,875 seconds whereas the true mean in (11) for the simulated item parameters was equal to 5,876 seconds. The two variances were exactly equal (490). The fit of the approximating lognormal has been shown to be quite good under different conditions; for the important upper tail of the distribution, it has even been excellent.

The example also illustrates another important fact for the lognormal distribution—only a few lower cumulants are required to control the critical features of its shape with remarkable precision. As is well known, for a distribution skewed to the right, the mean is always larger than the median. The standard lognormal in (14) has median equal to μ and mean equal to $\exp(\mu + \sigma^2/2)$ (e.g., Kotz & Johnson, 1985, pp. 134–136). It thus meets this condition for any (positive) μ and σ . In fact, the difference between the mean and the median is strongly controlled by the size of

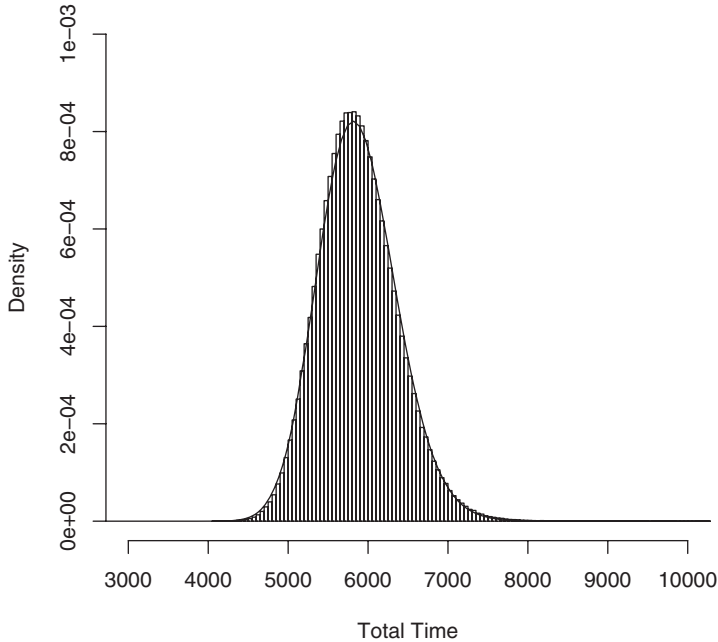


Figure 1. Histogram of the total-time distribution for a sample of $N = 100,000$ test takers working at average speed ($\tau = .0$) on the reference test of 78 items along with the standard lognormal density fitted to the histogram.

σ^2 . As a result, together these two parameters already account for the location, variance, and much of the characteristic skewness of response-time distributions.

Controlling Speededness Through Test Design

The factorization of the three cumulants of the response-time distributions in (5)–(7) into independent factors for the test taker and the item is another extremely convenient feature of the lognormal response-time model. In combination with the conditional independence of the response times, it leads to the expressions for the total-time cumulants in (11)–(13), which not only are simple but, for any given speed τ , also additive in the items. This means that, whatever the lognormal parameters of items added to a test, the mean total time for a given test taker is completely determined by the sum of the q_i parameters. Likewise, the variance and third cumulant are determined by the sums of the r_i and s_i parameters, respectively.

This feature allows us to build a new test form just by adding and removing items until the sums of these three parameters have a desired value. For instance, we could start with a desired mean and variance for a given level of τ , divide their values by $\exp(\tau)$ and $\exp(\tau^2)$, respectively, and use the results as target values for the sums of the q_i and r_i parameters for the new form. Because of the additivity, constraints on the sums of q_i , r_i , and s_i for a new form can easily be implemented through the use of a mixed integer programming model for test assembly.

For a testing program with new test forms to be built parallel to a reference form, we do not even need to bother about the factors with τ in (11)–(13) at all. The only thing required is sums of item parameters q_i , r_i , and s_i equal to those for the reference test. The total-time distribution for the new form is then identical to the one for the reference test for *any* possible speed level. Consequently, their marginal total-time distributions are automatically identical for any population of test takers.

Test forms with identical total-time distributions also have identical risk functions in (1). The claim of test forms having the same degree of speededness can then be made in the strongest possible sense—that is, without having to specify a minimum acceptable level of speed and maximum acceptable risk for which it should hold. Control of speededness through the design of the test is thus much more powerful than through the two alternative forms discussed earlier (manipulation of the time limit or post hoc evaluation and subsequent adjustment of the test).

The method is also quite effective. I will demonstrate this for three types of examples of test-form assembly from an item pool for a set of test specifications from a real-world testing program. The first type is a new form assembled to be parallel to a reference form. The second type of example shows the robustness of the new response-time constraints to a few dramatic changes in the test specifications for the testing program. The final example addresses the hypothetical case of the program having to revise the degree of speededness of its forms but wanting everything else to remain identical. This type of change may arise, for instance, when test takers' complaints about the degree of speededness of the test appear to be serious, or a test used for the selection of applicants for a job has to be made more speeded as the result of a new criterion study. But before presenting the examples, the test-assembly model that was used is explained.

Test-Assembly Model

As the empirical examples involved large numbers of test specifications in addition to the response-time constraints, mixed integer-programming modeling was used to represent the complete set of specifications. A key step in this approach to automated test assembly is the definition of 0–1 decision variables x_i , $i = 1, \dots, I$, for the items in the pool, which take the value $x_i = 1$ if item i is selected for the new test form and $x_i = 0$ if it is not. The variables are used to model the constraints as well as the objective function for the test. The type of modeling is extremely flexible and can accommodate any of the typical constraints met in the testing practice. Examples are: (i) quantitative constraints that impose bounds on the sums or means of numeric item attributes (e.g., item parameters, word counts, item-information values, exposure rates) for the entire form or a section in it; (ii) categorical constraints with bounds on the numbers of items from categories such as content classes, cognitive levels, and item type or format; and (iii) logical constraints to control the structure of the test (item sets with common stimuli, conditional selection of one type of item given another, anchor structure). Solutions for mixed integer-programming test-assembly models are easily calculated using the enormous power available in the solvers in modern linear programming software packages. The examples in this paper were calculated using *CPLEX 9.0* (ILOG, Inc., 2003). For a review of the current

possibilities of mixed integer-programming modeling for automated test assembly, see van der Linden (2005).

Examples

The item pool in the examples was a previous pool of 756 items from the *Law School Admission Test* (LSAT) calibrated under the 3-parameter logistic (3PL) model. In each of the examples, a test of 78 items consisting of the three sections of the LSAT (Logical Reasoning, Reading Comprehension, and Analytical Reasoning) was assembled. (The actual LSAT has two versions of one of these sections.) In addition to the time constraints introduced below, 436 constraints were necessary to deal with the various item types, content classifications, answer key distributions, word counts, etc., in force for the LSAT. As objective, the distance between the test-information function and the target for the LSAT was minimized. One form of 78 items was pre-assembled to serve as the reference form in each of the examples. All other forms were assembled from the pool of the remaining 678 items.

As no item parameters α_i and β_i in the response-time model in (3) were available for the LSAT items, I sampled them conditionally on the values of their IRT parameters. The conditional distribution was a multivariate normal distribution obtained from the *Arithmetic Reasoning* test in the adaptive version of the ASVAB. The distribution was estimated recalibrating the ASVAB items for the data set in the examples in van der Linden (2006) using a hierarchical framework with the 3PL model and the lognormal response-time model as first-level models and a multivariate normal distribution for their item parameter as second-level model (van der Linden, 2007). (Both the LSAT and the ASVAB have the 3PL model as their operational response model.) The recalibration was done using the Bayesian *R* software package `cirt` described in Fox, Klein Entink, and van der Linden (2007). In addition to the item parameters in the two first-level models, the package gives an estimate of the covariance matrix for the multivariate normal distribution between these parameters. Subsequently, the conditional distribution given the parameters in the 3PL model was used to sample the response-time parameters for the LSAT items conditional on their values for the parameters in the 3PL model. For the given joint distribution, the conditional distribution is also multivariate normal with means determined by the parameters of the items for the 3PL model and a covariance matrix that follows from the matrix for the joint distribution. The result was an item pool with all content attributes and IRT parameters for the LSAT as well as response-time parameters compatible with their IRT parameters according to the conditional distribution estimated from the ASVAB test.

The average estimate of the τ parameters was set equal to zero for identifiability reasons. In order to understand the empirical range of the speed parameters in the examples, it is helpful to know that σ_τ was estimated to be equal to .24.

New Form Equally Speeded as Reference Form

The values of the sums of item parameters in (11)–(13) for the reference form were used as target values and the new test form was assembled to meet these values. In order to evaluate the effect of the number of cumulants used to control the new

form, the example was repeated for constraints based on the first one, two, and three cumulants.

Let $j = 1, \dots, n$ denote the items in the reference form. The target values were defined as

$$\mathcal{T}_q \equiv \sum_{j=1}^n q_j, \quad (17)$$

$$\mathcal{T}_r \equiv \sum_{j=1}^n r_j, \quad (18)$$

$$\mathcal{T}_s \equiv \sum_{j=1}^n s_j. \quad (19)$$

The types of constraints that were used had the general form

$$\sum_{i=1}^I q_i x_i \leq \mathcal{T}_q + \delta_q, \quad (20)$$

$$\sum_{i=1}^I q_i x_i \geq \mathcal{T}_q - \delta_q, \quad (21)$$

$$\sum_{i=1}^I r_i x_i \leq \mathcal{T}_r + \delta_r, \quad (22)$$

$$\sum_{i=1}^I r_i x_i \geq \mathcal{T}_r - \delta_r, \quad (23)$$

$$\sum_{i=1}^I s_i x_i \leq \mathcal{T}_s + \delta_s, \quad (24)$$

$$\sum_{i=1}^I s_i x_i \geq \mathcal{T}_s - \delta_s, \quad (25)$$

where δ_q , δ_r , and δ_s are tolerances that define small intervals about the target values \mathcal{T}_q , \mathcal{T}_r , and \mathcal{T}_s , respectively. The constraints were added to the test-assembly model for $\delta_q = \delta_r = \delta_s = .01$.

Figure 2 shows the results for the three runs for a test taker operating at average speed ($\tau = 0$). The curves in each plot are the lognormal densities for the total-time distribution in (14)–(16) for the reference form and the new form assembled to have the same distribution. Even for the constraints on the first cumulant only, the match is already satisfactory. But for the constraints on the first two and three cumulants, the fit is perfect for all practical purposes. Especially noteworthy is the fit for the upper tail, which is the critical part of the distribution when the focus is on the speededness of the new form. Runs for alternative values of τ yielded results that were entirely comparable. Finally, remember that, in addition to the results demonstrated in Figure 2, the new test forms were also parallel with respect to the 436 other constraints in force for the LSAT and met the target for the test information function.

Observe that the new form does not need to have the same length as the reference test to meet its target values. The same holds, in fact, for any of the other specifications—which takes us to the next set of examples.

Changing Test Specifications but Maintaining the Degree of Speededness

Three different changes in the specifications for the LSAT were introduced but the target values in (17)–(19) were left intact. More specifically,

1. The new form was made much more difficult by shifting the target for the information function over a distance of 1.2 to the right.
2. All constraints related to the item-set structures for two of the sections were dropped.
3. The new form was shortened from 78 to 69 items (adjusting all bounds in the constraints of the test assembly model by approximately 10% to allow for the reduction).

Of course, such changes are not realistic; no testing agency would ever consider making one of its testing programs more difficult by an amount that corresponds to more than one standard deviation of the ability distribution, removing a large subset of its content constraints as the one associated with the attributes for the common stimuli in the item sets for the LSAT, or shortening its test forms without profiting from a shorter testing time. However, the purpose of these changes was not to make the example more realistic, but to challenge the effectiveness of the new constraints for the control of the total-time distribution on a test form through rather dramatic changes in other parts of the total set of constraints to be imposed on the items.

The result for the first change for the case of the first three cumulant and a test taker with average speed ($\tau = 0$) is displayed in Figure 3. The total-time distributions hardly differed from the distribution on the reference test. The result for the second change was entirely comparable and is not shown here for economy of space. For the shorter form of 69 items, the difference became noticeable but still was too small to have much practical meaning (Figure 4). We also tried shorter versions of the same test, but 10% reduction of the test length appeared to be right on the edge of what was possible for the current item pool without making the differences too large.

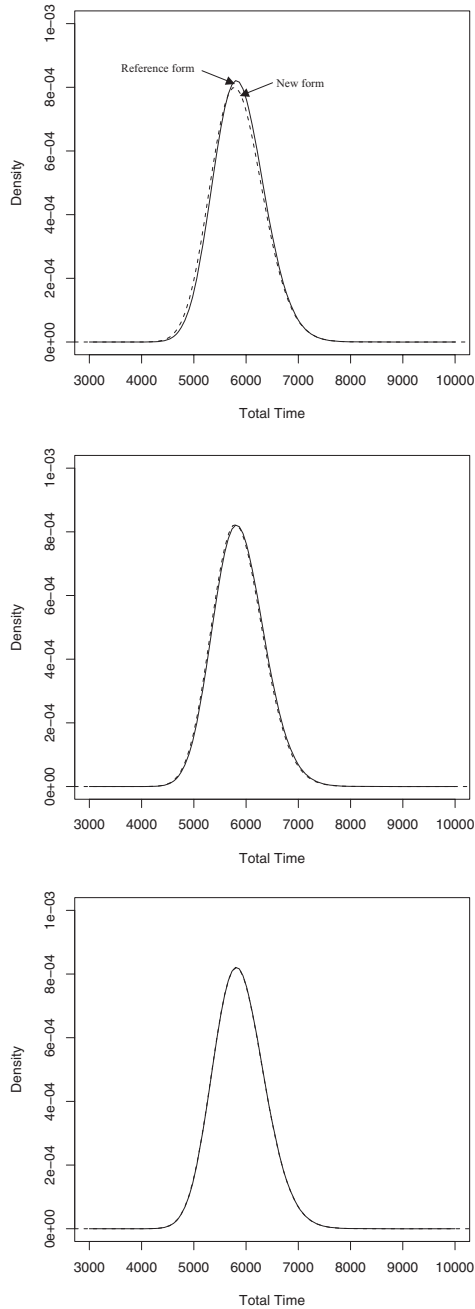


Figure 2. Total-time distributions for a test taker working at average speed ($\tau = .0$) on the reference form (solid curve) and the new form (dashed curve) assembled with the constraints based on the first one (top panel), two (middle panel), and three (bottom panel) cumulants.

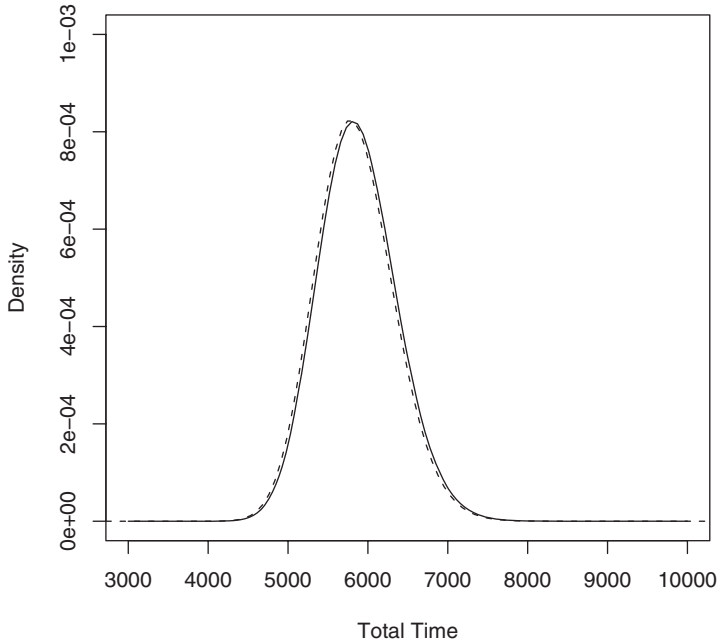


Figure 3. Total-time distributions for a test taker working at average speed ($\tau = 0$) on the reference form (solid curve) and the new form (dashed curve) assembled with the target for its information function shifted to the right over a distance of 1.2.

Changing the Degree of Speededness

The final example is for the case of a necessary change in the degree of speededness of the test with all other specifications left unchanged, including the time slot available for the test. The only possible way to realize this goal is by adjusting the time intensities of the items in the new form.

The adjustment can be implemented through the following series of steps: First, a speed level is selected to implement the change, for example, a minimum acceptable level of speed in use for the testing program or the average estimate for a few selected candidates who completed a recent form in the program just before the time limit. Let τ^{old} denote this level. Second, the risk function in (1) is used to evaluate the risk at τ^{old} for the given time limit. The risk is denoted as π^{old} . Third, the new level of risk, π^{new} , is set relatively to π^{old} . Fourth, the speed level associated with π^{new} is determined, using (1) again. Finally, the desired change

$$\Delta = \tau^{\text{new}} - \tau^{\text{old}} \quad (26)$$

is absorbed in the targets in (17)–(19) for the future test forms.

For example, the new target $\mathcal{T}_q^{\text{new}}$ follows from the expression for the mean total time on the test in (11) upon rewriting it as

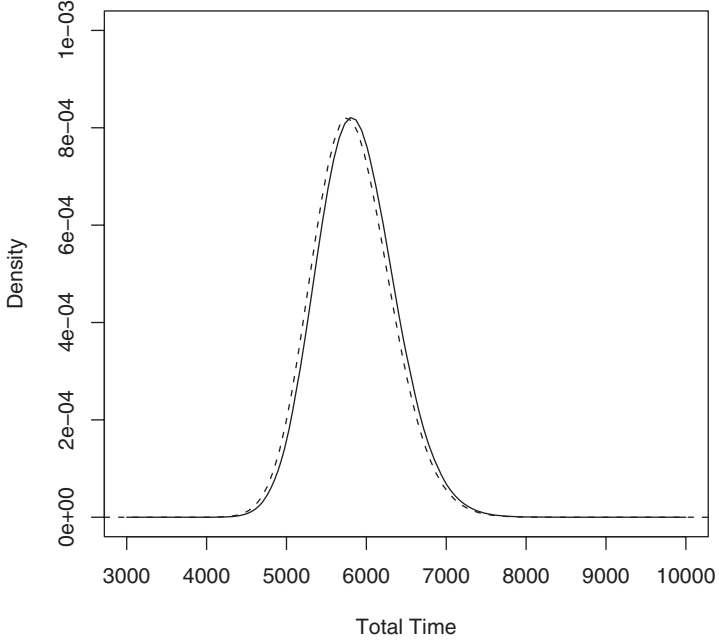


Figure 4. Total-time distributions for a test taker working at average speed ($\tau = 0$) on the reference form (solid curve) and the new form (dashed curve) with the length of the new form reduced from 78 to 69 items.

$$\begin{aligned}
 \mathcal{E}(T_{\text{tot}}) &= \exp(-\tau^{\text{old}}) \sum_{i=1}^n q_i \\
 &= \exp(-\tau^{\text{new}}) \left[\frac{\exp(-\tau^{\text{old}})}{\exp(-\tau^{\text{new}})} \sum_{i=1}^n q_i \right] \\
 &= \exp(-\tau^{\text{new}}) \left[\Delta \sum_{i=1}^n q_i \right] \\
 &= \exp(-\tau^{\text{new}}) \mathcal{T}_q^{\text{new}}.
 \end{aligned} \tag{27}$$

For the entire set of new targets it thus holds that

$$\mathcal{T}_q^{\text{new}} \equiv \Delta \sum_{j=1}^n q_j, \tag{28}$$

$$\mathcal{T}_r^{\text{new}} \equiv 2\Delta \sum_{j=1}^n r_j, \tag{29}$$

$$\mathcal{T}_s^{\text{new}} \equiv 3\Delta \sum_{j=1}^n s_j. \tag{30}$$

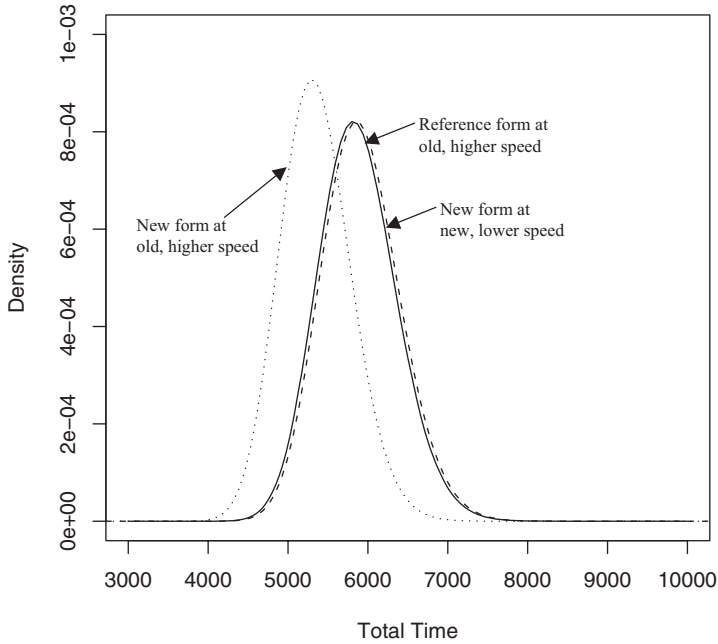


Figure 5. Total-time distributions for a test taker working at average speed ($\tau = 0$) on the reference form (solid curve) and a speed changed by $\Delta = -.1$ on the new form (dashed curve). The dotted curve is for the total-time distribution that would be obtained if the test taker had worked at the old speed ($\tau = 0$) on the new test form.

Figure 5 shows the result for a new form assembled from the LSAT pool to meet the new targets for a change equal to $\Delta = -.1$. The change may look small, but remember that the standard deviation for the speed parameter in the data set σ_τ was estimated to be .24. It thus amounted to a decrease of the speededness of the test by some 40% of the standard deviation. In spite of the change, the total-time distribution for the new form in Figure 5 is still close to the distribution for the reference form.

It is important to interpret the distributions in Figure 5 correctly: The distribution for the reference test is for the test taker working at the old average speed. The one for the new form with the less time-intensive items is for a test taker who accepted the new degree of speededness and slowed down to the new level of speed required to realize the change. Just for comparison, Figure 5 also shows the total-time distribution for the new form that would have been obtained if the test taker had not accepted the change and kept working at the old, higher level of speed. Not surprisingly, this distribution is more toward the lower end of the time scale than the other two.

Concluding Comments

As demonstrated by the examples, the only thing required to make a new form equally speeded as a reference form is to match the target values for the sums of the item parameters q_i , r_i , and s_i . It is thus not necessary to match these parameters item

by item, let alone do so directly for the item parameters α_i and β_i in the response-time model. Both relaxations give us much leeway when assembling a new form from an item pool.

The constraints in (20)–(25) can also be included in a model for the simultaneous assembly of a set of test forms to match a reference form. Each of the new forms then has the same old degree of speededness. The major change required for multiple-form assembly is the replacement of the decision variables x_i in the test-assembly model by variables x_{if} for the assignment of item i to form f in the set that is to be assembled. In addition, new constraints are necessary to control item overlap between the forms (van der Linden, 2005, chap. 6).

These possibilities hold regardless of all other constraints imposed on the assembly of new test forms. As long as these forms meet their target values in (17)–(19), features such as their content distribution or even their length do not play any role. In fact, the new and reference form do not even need to be for a test for the same ability. Exactly the same approach can be used to replace an existing test by an entirely new test—or, for that matter, a new battery of short tests—for a given time slot. The only thing required are items for the new test calibrated under the response-time model with the same empirical identifiability constraint as for the old test; for instance, a subset of test takers known to have worked at the same average speed, or one common item with its β_i parameter fixed. (Of course, we would have to eliminate the common item when scoring test takers for a different ability.) The actual linking procedure is simpler than those in use for parameter linking in IRT models (van der Linden, 2010).

The reader surprised by this conclusion should recall that there exist no “dimensionality problems” for speed. The person and item parameters in the lognormal model in (3) are for response times measured in natural units (e.g., seconds). The size of the units does not depend in any way on the ability measured by the test. Obviously, test takers may be inclined to work more slowly on one test than another, and tests with different types of items are likely to differ in their time intensity. But a change in the values of person and item parameters is not the same thing as a change of dimension.

Acknowledgments

Earlier work on this topic received funding from the Law School Admissions Council (LSAC). The opinions and conclusions contained in this paper are those of the author and do not necessarily reflect the policy and position of LSAC. The author is indebted to Rinke H. Klein Entink for computational support.

References

- Chuah, S. C., & van der Linden, W. J. (2008, March). *Detection of aberrant candidate responses: Improving detection by combining response patterns and response-time data*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.
- Cronbach, L. J., & Warrington, W. G. (1951). Time limit tests: Estimating their reliability and degree of speeding. *Psychometrika*, *14*, 167–188.

- Fenton, L. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communication Systems*, 8, 57–67.
- Finger, M., & Chuah, S. C. (2009, April). *Response-time model estimation via confirmatory factor analysis*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.
- Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package `cirt`. *Journal of Statistical Software*, 20(7), 1–14.
- Goldhammer, F. (2010, May). *Application of response-time modeling: Speed in reasoning tasks and its distinctness to reasoning ability*. Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.
- Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.
- ILOG, Inc. (2003). *CPLX 9.0* [computer program and manual]. Incline Village, NV: Author.
- Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74, 21–48.
- Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, 14, 54–75.
- Kotz, S., & Johnson, N. L. (1985). *Encyclopedia of statistical science* (Vol. 5). New York, NY: Wiley.
- Naus, J. I. (1969). The distribution of the logarithm of the sum of two log-normal variates. *Journal of the American Statistical Society*, 64, 655–659.
- Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, 16, 261–270.
- Stretch, L. S., & Osborne, J. W. (2005). Extended time test accommodation: Directions for future research and practice. *Practical Assessment, Research & Evaluation*, 10(8), 1–8.
- van der Linden, W. J. (2005). *Linear models for optimal test assembly*. New York, NY: Springer.
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31, 181–204.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308.
- van der Linden, W. J. (2009a). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, 33, 25–41.
- van der Linden, W. J. (2009b). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46, 247–272.
- van der Linden, W. J. (2010). Linking response-time parameters onto a common scale. *Journal of Educational Measurement*, 47, 1–23.
- van der Linden, W. J. (2011). Setting time limits on tests. *Applied Psychological Measurement*, 35. In press.
- van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, 44, 117–130.
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73, 365–384.

Author

WIM J. VAN DER LINDEN is Chief Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; wim_vanderlinden@ctb.com. His primary research interests include test theory, applied statistics, and research methods.