Linking Response-Time Parameters onto a Common Scale

Wim J. van der Linden
CTB/McGraw-Hill

Although response times on test items are recorded on a natural scale, the scale for some of the parameters in the lognormal response-time model (van der Linden, 2006) is not fixed. As a result, when the model is used to periodically calibrate new items in a testing program, the parameter are not automatically mapped onto a common scale. Several combinations of linking designs and procedures for the lognormal model are examined that do map parameter estimates onto a common scale. For each of the designs, the standard error of linking is derived. The results are illustrated using examples with simulated data.

Computerized testing programs automatically record the response times (RTs) of their test takers. The information in these RTs can be used to check the quality of the items, improve the design of the tests, monitor their quality during test administrations, diagnose the response behavior of the test takers, and increase the efficiency of test scoring. However, to fully exploit the information, a statistical model for the distribution of the RTs is required.

The model addressed in this research is the lognormal model proposed in van der Linden (2006). This model, which is briefly reviewed below, has separate parameters for the person and item effects on RT distributions, very much like a regular item response theory (IRT) model for the response distributions on test items. Due to the separation of these effects it becomes possible, for example, to diagnose the degree of speededness of a test (van der Linden, Breithaupt, Chuah, & Zhang, 2007), set the degree of speededness to a predetermined level both for a linear (van der Linden, submitted) and an adaptive test (van der Linden, 2009b), test hypotheses about the joint impact of attributes of the items on their difficulties and time intensities (Klein Entink, Fox, & van der Linden, 2009; Klein Entink, Kuhn, Hornke, & Fox, 2009), improve IRT item calibration (van der Linden, Klein Entink, & Fox, in press) and item selection in adaptive testing (van der Linden, 2008), and check test behavior for possible aberrances (van der Linden & Guo, 2008).

In order to use these applications in an operational testing program, it is necessary to periodically calibrate new test items under the RT model. This can simply be done as part of regular item pretesting and calibration. The extra step of estimating the RT parameters of the items does not require any substantial additional work or time; in fact, useful (Bayesian) procedures for estimating both the response and RT parameters are now available which do so in one simultaneous run (Fox, Klein Entink, & van der Linden, 2007; Klein Entink, Fox, & van der Linden, 2009; van der Linden, 2007).

The research in this paper addresses a more practical issue involved in calibrating new test items with respect to their time parameters, namely that of linking the
parameter estimates from a new calibration to the scale already established for the program. The same problem exists in regular IRT calibration (e.g., Kolen & Brennan, 2004, chap. 6), where different procedures have been developed to deal with the issue, such as the popular Stocking-Lord (1983) procedure. Because RTs are always recorded on a scale with a natural zero and a fixed unit (e.g., seconds), one might be tempted to think that the problem would not exist for the calibration of items with respect to their RT parameters. However, as will become clear below, this impression is incorrect.

In order to emphasize the analogies and differences between the linking of response and RT parameters from different calibrations onto a common scale, our treatment will be somewhat different from the traditional approach to parameter linking in IRT, which sometimes seems inclined to view it just as an IRT version of the older problem of observed-score equating in testing. Instead, we will treat linking solely as a consequence of an identifiability problem in response and RT modeling, which entails the necessity to impose additional restrictions on the model parameters to produce statistical estimates. The arbitrary character of these restrictions creates the linking problem. But before doing so, we will first review the lognormal RT model used in this research.

**RT Model**

The lognormal model for RTs was initially motivated by the wish to have a flexible model for RT distributions on the different types of items used in computerized testing. Its parameter structure may remind one of that of the two-parameter logistic (2PL) response model (see below), with a person parameter to represent the speed at which the test takers operate on the items and item parameters for their time intensity and discriminating power; but, as will be discussed below, the analogy is incorrect. In addition, because RTs have a natural lower bound, it is not necessary to introduce an extra parameter for a lower asymptote as in the 3PL response model. For a more detailed motivation along these lines, see van der Linden (2006).

Recently, however, the model has been derived directly from the definition of the speed at which test takers operate on test items in combination with two simple operations to adjust for the random nature of RTs as well as their tendency to skewed distributions (van der Linden, 2009a). Let $t_{ij}$ denote the response time for test taker $j$ recorded on item $i$, $\tau_j \in (-\infty, \infty)$ the speed at which $j$ works, and $\beta_i \in (-\infty, \infty)$ the amount of labor required by item $i$. The model then follows as

$$\ln T_{ij} = \beta_i - \tau_j + \epsilon_i, \quad \epsilon_i \sim N(0, \alpha_i^{-2}),$$

that is, as a normal density of the logtime with mean $\beta_i - \tau_j$ and discrimination $\alpha_i$ for item $i$ (reciprocal of the standard deviation). The actual values of the $\tau_j$ and $\beta_i$ parameters in an empirical application depend on the type of items as well as an identifiability restriction that has to be chosen (see below). Because a normal density for the logtime is the same as a lognormal density for the time on the natural scale (Johnson & Kotz, 1970, chap. 14), we can write (1) equivalently as a lognormal density for the distribution of $T_{ij}$:
The logarithmic transformation is due to the skewed nature of the RTs. Also, it is important to note the distinction between time and speed that underlies the model. Basically, speed parameter $\tau_j$ is a latent parameter for the amount of work per unit of time performed by the test taker when solving the item. The parameter has to be estimated from the observed RTs, just as labor intensity parameter $\beta_i$. Unlike time, speed is therefore not measured in seconds. Finally, observe that speed $\tau_j$ and labor intensity $\beta_i$ are assumed to be constant parameters whereas the RTs vary randomly about $\beta_i - \tau_j$. Observe the analogy between these assumptions and those of constant person and item parameters but random responses in IRT modeling.

Procedures and software for estimating the parameters in the model and assessing its fit to empirical RTs have been presented elsewhere (Fox et al., 2007; Klein Entink, Fox, & van der Linden, 2009; van der Linden, 2006, 2007) and are not reviewed here. In several applications to real-world test data, the model has been shown to yield a remarkably good fit (see the references in the introductory section).

Although estimation of the model is rather straightforward, we should be aware of one technical issue that has to be addressed while doing so: the indeterminacy of some of the parameters in the model. This issue, which is more generally known as a problem of identifiability in statistical modeling, and may cause the parameters in different calibration studies to be on different scales, will be discussed after we have reviewed an analogous identifiability problem and its role in parameter linking in regular IRT calibration.

Response Model

Let $U_{ij}$ be the response by test taker $j = 1, \ldots, N$ on item $i = 1, \ldots, n$. One of the mainstream models for dichotomous response variables is the three-parameter logistic (3PL) model:

$$\Pr(U_{ij} = 1) = p_i(\theta_j; a_i, b_i, c_i) \equiv c_i + (1 - c_i)\Psi(\theta_j),$$

with

$$\Psi(\theta_j) \equiv \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}.$$

In this model, parameter $\theta_j \in (-\infty, \infty)$ represents the ability of test taker $j$, and parameters $b_i \in (-\infty, \infty)$, $a_i > 0$, and $c_i \in [0, 1]$ can be interpreted as the difficulty, discriminating power, and guessing probability for item $i$, respectively (Lord, 1980). If $c_i = 0$, the model becomes the two-parameter (2PL) model, which is just the logistic function in (4).

Observe that the lognormal model in Equation 2 specifies a density function for the distribution of the random variable of interest ($T_{ij}$) but that this does not hold for the representation of the response model in Equations 3 to 4, which is a response
function for the correct response on the item \((U_{ij} = 1)\). However, the distribution of \(U_{ij}\) is Bernoulli with a probability function that follows directly from (3):

\[
f(u_{ij}; \theta_j; a_i, b_i, c_i) \equiv p_i(\theta_j; a_i, b_i, c_i)^{u_{ij}} [1 - p_i(\theta_j; a_i, b_i, c_i)]^{1-u_{ij}}. \tag{5}
\]

Therefore, although the parameter structure of the lognormal model reminds one of that of the representation of the 2PL model in (4), a direct comparison between the two models should be made between their density and probability functions in (2) and (5). The difference \(\beta_i - \tau_j\) in (2) can be interpreted as the mean of the distribution of \(\ln T_{ij}\) and the reciprocal of the discrimination parameter, \(a_i^{-1}\), as its standard deviation. However, the distribution of response variable \(U_{ij}\) in (5) does not have mean and standard deviation \(b_i\) and \(a_i^{-1}\) but \(p_i(\theta_j; a_i, b_i, c_i)\) and \(\{p_i(\theta_j; a_i, b_i, c_i)[1 - p_i(\theta_j; a_i, b_i, c_i)]\}^{1/2}\), respectively. (Parameter \(a_i\) does have a direct impact on the standard deviation of \(U_{ij}\), though; see the graphical example in van der Linden, 2006, fig. 1. It should therefore still be interpreted as a discrimination parameter.). For the current research, however, the most important difference between the two models is that the key expression \(a_i[\ln t_{ij} - (\beta_i - \tau_j)]\) in the RT model is linear in the observed variable \(\ln t_{ij}\) whereas expression \(a_i(\theta_j - b_i)\) in the response model does not contain \(u_{ij}\) at all. Instead, the formal position of \(\ln t_{ij}\) has now been taken by the unknown ability variable \(\theta_j\).

### Identifiability of the 3PL Model

As is well known, the response model in Equations 3 to 4 has an indeterminacy of the origin and unit of scale of \(\theta\): that is, the response function does not change if the origin is changed (the same constant is added to all \(\theta_j\) and \(b_i\)) or the same occurs to the unit of scale (all \(\theta_j\) and \(b_i\) are multiplied by the same constant and \(a_i\) is divided by it). This lack of determinacy is an instance of the more general problems of identifiability in statistical modeling (e.g., Casella & Berger, 2002, section 11.2).

A parameter in a probability model is identifiable if different values of it correspond to different distributions of the random variable. A probability model is identifiable if each of its parameters is identifiable. Identifiability problems do not exist for the standard distributions in the statistical textbooks, such as the regular versions of the lognormal or Bernoulli distribution above. For these, a change of any of their parameters has a well-defined effect on their distribution, e.g., their location or skewness. But they easily arise if their densities are given more complex parameter structures to account for person and item effects, such as in (2) and (5).

The following argument pinpoints the identifiability problem for the 3PL model. First, the Bernoulli probability function in (5) has only one parameter—success parameter \(p_i(\theta_j; a_i, b_i, c_i)\). Thus, the model is identifiable when \(p_i(\theta_j; a_i, b_i, c_i)\) is. Second, because of a one-to-one relation between parameter \(p_i(\theta_j; a_i, b_i, c_i)\) and guessing parameter \(c_i\), the former is identifiable only when the logistic function \(\Psi(\theta_j)\) in (4) is. Third, as \(\Psi(\cdot)\) is monotone, \(\Psi(\theta_j)\) is identifiable on the condition that the parameter structure \(a_i(\theta_j - b_i)\) is. But this is not the case because of the indeterminacy discussed above.
Model identifiability is a necessary condition for parameter estimation. The requirement follows directly from the role played by the likelihood function in it. The function exists of a product of the factors in (5), one for each combination of test taker and item. Obviously, it is impossible to infer unique parameter estimates from a likelihood function when these factors are not identifiable. (This conclusion does not imply that we necessarily have good estimates as soon as the model is identifiable. Although the likelihood then changes with each parameter, the rate of change for some of them can be minute, and we then remain quite uncertain as to their true value. For obvious reasons, this condition is usually referred to as poor or weak identifiability.)

Problems of identifiability can be solved by putting more restrictions on the parameters. The fact that, for the 2PL and 3PL models, the problem of identifiability boils down to an indeterminacy of the origin and unit of scale immediately suggests additional restrictions on these quantities. A popular practice is to set the mean and standard deviation of the parameters $\theta_j$ equal to zero and one across the test takers in the sample, respectively.

Although popular, these restrictions are nothing but conventions. For a single item calibration study, any other set of two linear restrictions on the parameters, for instance, on the mean and standard deviation of the $b_i$ parameters, or the values of two of these parameters (with their size reflecting their actual order), would yield equally good estimates. However, the choice becomes critical when in later calibration studies new items for the same testing program have to be put on the same scale. We are then faced with a problem of parameter linking.

The traditional way of linking calibrations in IRT-based testing programs to a scale that has already been established is to run a computer program with its default identifiability restrictions and adjust afterwards. The adjustment is required because formally identical restrictions are not necessarily empirically identical; that is, they may actually refer to different levels of achievement. For example, if we impose the restrictions $\mu_\theta = 0$ and $\sigma_\theta = 1$ on two different calibrations and the second group of test takers is more able on average and/or shows larger variation in ability, actually two different sets of empirical restrictions have been imposed. Linking transformations are thus required to adjust parameter estimates from different calibration studies for the empirical differences between identifiability restrictions.

Identifiability of the Lognormal RT Model

Because $\ln t_{ij}$ is measured in fixed units and has a natural zero, it follows that a version of the lognormal RT model with a single location parameter, that is, with the substitution of

$$\mu_{ij} \equiv \beta_i - \tau_j$$

into (2), is identifiable. In this version, $\mu_{ij}$ is the mean of $\ln T_{ij}$ and $\alpha_i$ the reciprocal of its standard deviation. For $m$ realizations of $\ln T_{ij}$ (that is, independent observations of the RTs of test taker $j$ on item $i$), the maximum-likelihood (ML) estimates of the two parameters are defined by the following simple expressions:
\[ \tau_j = \beta_i - \mu_{ij} \]

Figure 1. Graphical illustration of the unidentifiability of speed parameter \( \tau_j \) and time intensity parameter \( \beta_i \). The model constrains both parameters only to be on the line \( \tau_j = \beta_i - \mu_{ij} \).

\[
\hat{\mu}_{ij} = m^{-1} \sum \ln t_{ij} \tag{7}
\]

\[
\hat{\alpha}_i = m \left( \sum \ln t_{ij} - \hat{\mu}_{ij} \right)^{-1/2} \tag{8}
\]

This argument shows that the RT model in (2) can only have an identifiability problem for its time intensity parameters \( \beta_i \) and speed parameters \( \tau_j \). That it does have this problem is demonstrated by the fact that, for any value of \( \varepsilon \), the distribution of \( \ln T_{ij} \) remains the same if we replace \( \beta_i \) and \( \tau_j \) by \( \beta_i - \delta \) and \( \tau_j - \delta \).

The lack of identifiability can also be shown graphically. Because \( \mu_{ij} \) is identifiable, we can draw the line \( \tau_j = \beta_i - \mu_{ij} \) for a fixed value of \( \mu_{ij} \) in the space of possible values of \( (\beta_i, \tau_j) \) in Figure 1. Without any additional restrictions, we only know that \( \beta_i \) and \( \tau_j \) are on this line but have no unique values for these parameters, let alone knowing whether such values could be estimated.

It is interesting to observe that, unlike discrimination parameter \( \alpha_i \) in the 2PL or 3PL model, \( \alpha_i \) is always identifiable. This point is illustrated empirically in Figure 2, where the estimates of \( \alpha_i \) for the same 50-item test from two different calibration samples are plotted against one another. One calibration was based on a sample of 500 test takers from a uniform population with \( \tau \sim U(-2, 2) \); the other on a sample of the same size from a normal population \( \tau \sim N(1, .5) \). The identifiability restrictions used in the two calibrations were unrelated. Neither did we use any explicit form of linking. Nevertheless, except for random estimation error, the pairs of independent estimates for each item lie on the identity line.
Convenient Identifiability Restrictions

For the lognormal RT model to become fully identifiable, only one extra linear restriction on the time intensity and/or speed parameters is required. The choice of restriction is practically unlimited and can be guided by convenience only. Three obvious types of restrictions are presented.

First, we can fix one of the time intensity or speed parameters at an arbitrary constant $c$. This type of restriction amounts to the addition of a line $\beta_i = c$ or $\tau_j = c$ to the space of $(\beta_i, \tau_j)$ in Figure 1, whereupon the other parameter is uniquely determined by the intersection of the line with $\tau_j = \beta_i - \mu_{ij}$.

It is easy to show that, given a simple condition on the calibration design, if one of the time intensity or speed parameters is identifiable, the parameters for all item and person parameters in the calibration are identifiable. The argument runs as follows: Suppose $\beta_i$ is identifiable. Then $\tau_j$ is also identifiable. The same then holds for the speed parameter $\tau_{j_0}$ of any other test taker $j_0$ who responds to item $i$. The claim is true because $\mu_{ij_0}$ is identifiable, and, hence, so is $\tau_{j_0} = \mu_{ij_0} - \beta_i$ for all $j_0 \neq j$. But then it immediately follows that the time intensity parameters $\beta_{i_0}$ of all other items $i_0 \neq i$ to which any of these test takers responds are identifiable. The only condition required to obtain identifiability of all parameters is connectivity of the calibration.
design; that is, there should be a path along common item or person parameters from any given item-person combination to any other combination. (Note that this condition of connectivity of the calibration design is not the same as that of the data matrix required for the existence of unique estimates in the Rasch model in Fischer (1981), although the former is necessary for the latter to hold.)

This type of identifiability restriction would be helpful in combination with a fixed reference item. A reference item should have a task that is typical of the content domain but insensitive to memory or learning effects. If such items are carefully calibrated using an extremely large sample, parameter linking could be established automatically by inserting a reference item in each new item calibration and fixing its parameters to the known values. Admittedly, the number of domains for which such items are possible is small but testing of psychomotor skills might be an appropriate candidate.

Second, alternatively, we could fix the mean of all speed parameters at a constant $c$. For $c = 0$ the choice leads to

$$\mu_\tau = 0,$$

which is a restriction analogous to $\mu_\theta = 0$ for the 2PL and 3PL response models above.

Interestingly, the choice can be interpreted directly in terms of the RTs on the items. From (6),

$$\mathcal{E}(\ln T) = \mu_\beta - \mu_\tau,$$

where the expectation is taken over replications, test takers, and items. Thus, the identifiability constraint in (9) implies

$$\mu_\beta = \mathcal{E}(\ln T).$$

This consequence enables us to interpret the values of the time-intensity parameters $\beta_i$ as deviations from the expected logtime for a random test taker from the population on a random item in the test. In addition, it shows that in an actual calibration study, (9) sets the average estimate of the time intensity parameters equal to the average logtime in the data set; that is, the restriction leads to

$$\hat{\beta} = (nN)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N} \ln t_{ij}.$$  

Third, analogous to the preceding type of restriction, we could also choose

$$\mu_\beta = 0.$$
Obviously, this restriction implies

\[ \mu_\tau = \mathcal{E}(\ln T), \]

which now enables us to interpret the speed parameters as deviations from the average logtime for the population of test takers on the test.

Finally, it is not necessary to define the restrictions in (9) or (13) on all speed parameters or time intensity parameters in the calibration study. They can also be defined on a subset of them. This option becomes convenient when we have to link RT parameters in different calibration studies through a smaller set of common items or test takers.

**Linking Procedures**

The choice of linking procedure depends on the structure of the two calibration designs as well as the relationship between them. We will refer to the union of two calibration designs as a **linking design**. It should not come as a surprise that a necessary condition for a linking design to be effective is the same feature of connectivity as in the discussion of the first type of identification restriction above. If a design would miss this, we may still be able to map one part of the parameters onto the previous scale but not all of them.

As already noted, connectivity requires a linking design to have common persons and/or items. The same condition holds for the sample designs in studies of observed-score equating; for a review, see e.g., von Davier, Holland & Thayer (2005, chap. 2). Although linking problems are of a different nature, the terminology for these designs is well established, and, where meaningful, we can use the same terminology. We will therefore group our presentation of the linking procedures around the following designs: (a) single-group design, (b) randomly equivalent groups design, (c) anchor-item design, and (d) concurrent-calibration design.

In principle, two different kinds of linking procedures are possible:

1. The nature of the linking design may enable us to impose exactly the same identifiability restriction on both calibrations. As a result, the second set of parameters are automatically mapped onto the same scale as the first set, and no adjustment afterwards is necessary. We will refer to this procedure as **implicit linking**.
2. If both calibrations have different identifiability restrictions, a linking transformation has to be established that maps the second parameters onto the scale for the first. Because both scales always have equal units due to the presence of \( \ln t_{ij} \) in the model (see the earlier discussion on the identifiability of \( \alpha_i \)), the transformation can be written as

\[ \tau^{(1)} = \tau^{(2)} + v. \]

We will refer to this type of linking as **post hoc linking**.
Post hoc procedures can also be used as a check on possible implementation problems for a linking design with implicit linking. Examples of implementation problems are unanticipated changes in the testing conditions (e.g., a change in the degree of speededness of the test) or in the population of test takers. The same problems are a threat to parameter linking in IRT calibration (e.g., Oshima, 1994). The treatment of such problems is beyond the scope of this paper; we only observe that, except for random estimation error, a well-implemented design with implicit linking should yield an estimate of the linking constant equal to $v = 0$.

**Single-Group Design**

In a single-group design for parameter linking, the second set of items is administered to the same group of test takers under identical conditions. This feature admits a simple version of implicit linking using a restriction on the mean of the $\tau$s. Let $\hat{\mu}_\tau(1)$ be the average estimate of the $\tau$ parameters of the group of test takers used in the calibration of the first set of items. During the calibration of the second set, the following identifiability restriction is imposed:

$$N^{-1} \sum_{j=1}^{N} \tau_j^{(2)} = \hat{\mu}_\tau(1).$$  \hspace{1cm} (16)

Post hoc linking with a single-group design involves estimating the unknown constant $v$ in the linking transformation in (15). Averaging the transformation over the common test takers gives

$$v = \hat{\mu}_\tau(1) - \hat{\mu}_\tau(2).$$  \hspace{1cm} (17)

Thus, the post hoc linking transformation can be written as

$$\tau^{(1)} = \tau^{(2)} + \hat{\mu}_\tau(1) - \hat{\mu}_\tau(2).$$  \hspace{1cm} (18)

Of course, the transformation has to be applied both to the estimates of the $\tau_j$ and $\beta_i$ parameters.

It is interesting to note that the estimate of the linking constant in (17) is a least-squares solution to the problem of fitting a line with unit slope and unknown intercept to the plot of the two sets of parameter estimates. This claim is proven in the Appendix.

It is not necessary that all test takers in the second calibration are common. Both procedures can also be applied to a common subgroup of test takers.

The impact of estimation error on these procedures will be discussed later. For now, we only observe that the implicit and post hoc linking procedures are equally sensitive to estimation error during the two calibrations, and that convenience is the only criterion for choosing between them. For post hoc linking, the sensitivity to the errors in the second calibration follows directly from (18). In implicit linking, the sensitivity arises because the identifiability restriction in the second calibration is actually imposed at the level of the parameter estimates during the calibration. (We
are unable to impose any empirical restriction directly on the values of unknown parameters."

**Randomly Equivalent Groups Design**

In this type of linking design, the sets of items administered in the first and second calibration are administered to two random samples of test takers from the same population. The means in (16) and (18) are now estimated from these two samples; otherwise, the procedures are entirely identical to those for linking with a single-group design.

Again, the procedures can also be applied if the two calibrations only have a randomly equivalent subgroup. When it is doubted whether the two samples are from the same population, the samples might be improved through poststratification on a set of well-chosen conditioning variables. In fact, when speed and ability correlate, we could even poststratify using the ability scores.

**Anchor-Item Design**

When the two calibrations have common items, the linking procedures can be based on their parameters. In principle, one common item could suffice but then, to reduce the impact of estimation error, large samples of test takers would be required.

Implicit linking is achieved through the imposition of the following restriction on the second calibration:

\[
K^{-1} \sum_{k=1}^{K} \beta_{j}^{(2)} = \hat{\mu}_{\beta}^{(1)},
\]

where \(k = 1, \ldots, K\) are the common items and \(\hat{\mu}_{\beta}^{(1)}\) is their average estimate from the first calibration.

Alternatively, analogous to (18), if the second set of items has already been calibrated, *post hoc* linking using the transformation

\[
\beta^{(1)} = \beta^{(2)} + \hat{\mu}_{\beta}^{(1)} - \hat{\mu}_{\beta}^{(2)}
\]

is still possible. Although the transformation is calculated from the parameter estimates for the common items only, it should also be applied to the \(\beta\) parameters of all items as well the speed parameters \(\tau_j\).

**Concurrent Calibration Design**

This last type of linking typically arises in pre-equating studies where data from several groups of test takers on different sets of items are “patched together” for a single calibration. We then have a more general design with structurally missing data which does not have the clear-cut pattern of a single-group or an anchor-test design. If the design is (a) still connected and (b) contains some of the items and/or test takers from an earlier calibration in which the scale of the parameters was
established, all new items can be calibrated concurrently using an identifiability restriction derived from the earlier calibration.

This implicit linking requires a version of the restriction in (16) or (19) with the average taken over the earlier items or test takers. As shown by the standard errors of linking in the next section, when there is a choice between earlier items and test takers one of the main factors that should guide the choice is the size of the link. An advantage of concurrent calibration is that it prevents potentially complicated chains of post hoc linking through the transformations in (18) and/or (20), which could easily lead to an undesirable propagation of linking error.

**Standard Errors of Linking**

Identifiability restrictions and linking transformations are subject to estimation error in the $\tau_j$ and $\beta_i$ parameters. In addition, the randomly equivalent groups design involves error due to the sampling of test takers. The size of all these errors is summarized in the standard error of linking. The derivation of the (approximate) standard errors in this section shows that, for the conditions typically met in real-world testing programs, the impact of linking errors would not be too substantial.

From the standard theory for the normal distribution (e.g., Lehmann, 1999, ex. 7.1.4), for known parameters $\alpha$ and $\beta$, Fisher’s information about $\tau_j$ in the RT on an item is equal to $I(\tau) = \alpha^2$. Therefore, the (asymptotic) variance of the ML estimator of $\tau_j$ from the items $i = 1, \ldots, n_1$ in the first calibration is equal to

$$I(\tau)^{-1} = \left( \sum_{i=1}^{n_1} \alpha_i^2 \right)^{-1}.$$  \hfill (21)

This expression assumes local independence between the RTs, which seems reasonable for a RT model for test takers operating at a constant speed (van der Linden, 2007, 2009a). For a statistical test of the assumption, see van der Linden and Glas (in press).

Observe that the variance is independent of the true value of $\tau_j$. Consequently, for a group of $N$ test takers, the sampling variance of the estimator of the mean $\mu_\tau$ is asymptotically equal to

$$N^{-1} \left( \sum_{i=1}^{n_1} \alpha_i^2 \right)^{-1}.$$  \hfill (22)

A single-group design involves the estimation of $\mu_\tau$ for the same group of test takers in two independent calibrations. Hence, the standard error of linking for this design can be approximated as

$$\sigma_{SG} = N^{-1/2} \left[ \left( \sum_{i=1}^{n_1} \alpha_i^2 \right)^{-1} + \left( \sum_{i=1}^{n_2} \alpha_i^2 \right)^{-1} \right]^{1/2}.$$  \hfill (23)

where $n_2$ denotes the number of items in the second calibration.
For the randomly equivalent groups design, the standard error also involves the sampling error for the two groups of \( N_1 \) and \( N_2 \) test takers from the common population. Let \( \sigma_\tau \) be the standard deviation of \( \tau \) in this population. Because the two types of error are independent, an approximate standard error of linking for this type of design is

\[
\sigma_{\text{REG}} = \left[ (N_1 + N_2)^{-1} \sigma_\tau^2 + N_1^{-1} \left( \sum_{i=1}^{n_1} \alpha_i^2 \right)^{-1} + N_2^{-1} \left( \sum_{i=1}^{n_2} \alpha_i^2 \right)^{-1} \right]^{1/2}.
\]

(24)

The derivation of the standard error for the anchor-item design runs analogous to that of (23). For known parameters \( \alpha \) and \( \tau \), it holds that \( I(\beta) = \alpha^2 \). In the first calibration, the parameters \( \beta_k \) of anchor items \( k = 1, \ldots, K \) are estimated from the RTs of test takers \( j = 1, \ldots, N_1 \). The estimate of \( \beta_k \) is inferred from \( N_1 \) RTs on the same item \( k \); therefore, its sampling variance is equal to \( (N_1 \alpha_k^2)^{-1} \). It follows that the estimator of \( \mu_\beta \) for the entire set of anchor items has asymptotic variance

\[
(K N_1)^{-1} \sum_{k=1}^{K} \alpha_k^{-2}.
\]

(25)

As \( \mu_\beta \) is also estimated (independently) in the second calibration, the standard error of linking for an anchor-item design can be approximated as

\[
\sigma_{\text{AI}} = K^{-1/2} \left[ (N_1^{-1} + N_2^{-1}) \sum_{k=1}^{K} \alpha_k^{-2} \right]^{1/2}.
\]

(26)

All three standard errors decrease with the numbers of persons and sums of the discrimination parameters for the items from which the linking constants are estimated. For the single-group and randomly equivalent groups designs, the linking constant is the difference between the means of the \( \tau \)s estimated from the sets of items in the two calibrations. Hence, the sums of the discrimination parameters in (23) and (24) are over all these items. For the anchor test design, the linking constant is the difference between the estimates of the means of the \( \beta \)s from the two administrations of the anchor items; therefore, the sum is over the set of anchor items.

For calibration samples of equal size, the size of the sums of discrimination parameters explains the main difference between the standard errors for the single-group and randomly equivalent groups designs and that for the anchor-item design. Because the number of items to be calibrated is typically larger than the number of anchor items, the first two types of design have smaller standard errors of linking. Besides, the randomly equivalent groups design is the only design sensitive to sampling error. It can thus be concluded that the single-group design is the most efficient linking design.

This conclusion is confirmed by the plots in Figure 3, which show the three standard errors as a function of the size of the calibration samples. Each of these plots
Figure 3. Standard errors of linking for the single-group, randomly equivalent groups, and anchor-item designs as a function of the size of the calibration samples.

is for a design in an empirical linking study reported in the next section. All tests in this study had \( n = 50 \) items. The number of anchor items in the anchor-item design was \( K = 10 \). The standard deviation of \( \tau \) in the population of test taker for the randomly equivalent groups design was set equal to \( \sigma_\tau = .116 \). This value was found for an earlier empirical data set on which the standard test in the empirical study was based. Further details of the linking study are given below. For convenience, the plots are based on a common size \( N \) for the two calibration samples. For any value of \( N \), the single-group design is superior. The only difference between the standard errors for this design and the randomly equivalent groups design is due to the role of \( \sigma_\tau \), which controls the height of the horizontal asymptote in the plot of the latter. Because of the asymptote, for sample sizes larger than, say, \( N = 100 \), the anchor-test design already outperforms the randomly equivalent groups design. So, for real-world sample sizes the anchor-item design might be second best. Of course, these comparisons rely directly on the size of the discrimination parameters of the sets of items that are calibrated as well as the population standard deviation \( \sigma_\tau \). In this example, the discrimination parameters in the second calibration were somewhat
larger for the single-group design than the other two designs (see below). Also, in some applications, we may be able to manipulate the standard deviation.

Although the analysis of the linking errors for the three types of design yields a statistical preference for the single-group design, nonstatistical considerations should also play a role. As a rule, designs with common items are more robust with respect to implementation errors than designs with common or randomly equivalent test takers. Typical implementation errors for the latter two are due to changes in the test takers between test administrations (learning, forgetting, change of population, etc). If such errors occur, they lead to linking bias in addition to the usual random error due to parameter estimation and/or sampling of test takers. When designing linking studies, we are thus faced with the same dilemma between bias and efficiency typical of nearly all statistical optimization problems.

**Examples**

The goal of these examples is to illustrate the different linking designs and transformations for tests with realistic ranges for the RT parameters. Because we wanted to evaluate the results of each linking relative to a fixed scale, a standard test of \( n = 50 \) items was adopted. Parameters \( \alpha_i \) and \( \beta_i \), \( i = 1, \ldots, 50 \), for the items in this test were sampled from uniform distributions \( U(1, 3) \) and \( U(3, 5) \), respectively. The items were calibrated using RTs generated for \( N = 500 \) person with parameters \( \tau_j \) sampled from \( U(-2, 2) \). The standard restriction in this calibration was \( \mu_\tau = 0 \). The ranges of these parameters were chosen to match those found in an earlier calibration of an item pool for the Arithmetic Reasoning test from the Armed Services Vocational Aptitude Test Battery (ASVAB) using a large empirical data set (van der Linden, 2006).

All parameters were estimated using the Markov-chain Monte Carlo (MCMC) procedure (Gibbs sampler) for the lognormal RT model presented in the same reference. For this procedure, it is easy to impose identifiability restrictions on the \( \tau_j \) or \( \beta_i \) parameters. The operation required is a renorming of the draws from the conditional posterior distributions of the pertinent parameters after their iteration step in the procedure. (For instance, if the restriction is on the mean of a subset of the \( \beta_i \) parameters, the mean for this subset is subtracted from the sampled value for each of the \( \beta_i \)s.) Because all other parameters are sampled conditional on the values for the renormed parameters, the scale defined by the restrictions is automatically imposed on all relevant parameters.

The MCMC procedure in the current study had an identical implementation as in the earlier study with the ASVAB test. Particularly, to get estimates that could be considered as MLEs, the prior distributions were chosen to be virtually noninformative (e.g., the priors for the \( \tau_j \) parameters were normals with a standard deviation equal to 1,000); for more details, see van der Linden (2006).

Each implicit or post hoc linking was evaluated through a comparison between (a) the estimates of the \( \tau_j \) or \( \beta_i \) parameters in the first calibration, and (b) the estimates in the second calibration linked to the scale for the first calibration. As demonstrated in Figure 2, it was not necessary to evaluate the estimates of the \( \alpha_i \) parameters; they are identifiable and always on the same fixed scale.
Parameter Estimation Error

When evaluating the linking results, it is helpful to have a general sense of the estimation error in the RT parameters. The plots in Figure 4 show the true and estimated parameters for the standard test adopted in this study. Generally, the estimates were quite accurate. The estimates of the $\tau_j$ parameters had little error, even though each of them was based on the RTs on 50 items only. The estimates of the $\beta_i$ parameters were based on 500 test takers and matched their true values quite closely. For the 3PL response model, it is generally difficult to get accurate estimates of the discrimination parameter for this sample size. But the estimates of the $\alpha_i$ parameters in the RT model were quite satisfactory, the reason being their shared scale with the observed logtimes. Because of this, these parameters are less susceptible to poor identifiability than the discrimination parameters in the 3PL model, which easily suffers from a tradeoff between their estimates and those of the unknown $\theta$s.

Single-Group Design

The first calibration in the study of this linking design was for the items in the standard test with identifiability restriction $\mu_\tau = 0$. The second set of 50 items was
generated with parameters randomly drawn from $\alpha \sim U(1, 2)$ and $\beta \sim U(4, 5)$. (The only exception was the first item, for which the parameter values were maintained; see below). Observe that the items in this second set were thus considerably more time intensive and had less discriminating power. The RTs on the second set of items were generated for the same 500 test takers as for the standard test.

Two different second calibration runs were performed. The first run used implicit linking. The mean estimate of the $\tau$s from the calibration of the standard test was $\hat{\mu}_\tau^{(1)} = .000$ (which was as expected given the choice of population, $\tau \sim U(-2, 2)$, and the size of the sample). The identifiability restriction for the second calibration was therefore also $\mu_\tau = .000$. The first plot in Figure 5 shows that although they were estimated from different items, except for estimation error the two sets of estimates of the $\tau$s were on the same scale.

For the second run, the arbitrary choice $\mu_\tau = 1$ was made for the identifiability restriction. The mean estimate of the $\tau$s was now $\hat{\mu}_\tau^{(2)} = .999$. The estimates from this run were mapped back onto the scale of the standard test using the linking transformation $\tau^{(1)} = \tau^{(2)} - .999$ (see equation 18). The second plot in Figure 5 shows that for this type of linking the second estimates of $\tau$ were also on the same scale as the first estimates.

As an independent check on the quality of the linking, one item in the second calibration was actually kept identical to that in the standard test (Item 1 with true parameters $\alpha_1 = 2.375$ and $\beta_1 = 3.107$). The initial estimate of $\beta_1$ was equal to 3.080. For the second calibration with implicit linking an estimate of the same parameter equal to 3.107 was found. The calibration run with subsequent post hoc linking yielded an estimate equal to $v = 4.110 - .999 = 3.111$ on the scale for the standard test. The differences between these three estimates were all acceptable given the standard error of estimation for $\beta_1$, which was estimated to be equal to .045, .046, and .044 in the three different calibrations.
Table 1

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>First calibration</td>
<td>.000</td>
<td>1.166</td>
</tr>
<tr>
<td>Implicit linking ($N = 500$)</td>
<td>.000</td>
<td>1.140</td>
</tr>
<tr>
<td>Post hoc linking ($N = 500$)</td>
<td>.000</td>
<td>1.140</td>
</tr>
<tr>
<td>Post hoc linking ($N = 200$)</td>
<td>.000</td>
<td>1.163</td>
</tr>
</tbody>
</table>

**Randomly Equivalent Groups Design**

The setup of this linking study was entirely identical to that of the preceding case. The only difference was that the second calibrations were based on a new sample of $N = 500$ test takers from the same population $\tau \sim U(-2, 2)$. This modification enabled us to study the role of sampling error, which is the only extra source of random error for this type of linking design. To assess the impact of the size of the sample of test takers, we repeated the second calibration for another sample of $N = 200$ test takers from the same population with arbitrary identifiability restriction $\mu_\tau = 1$.

Unlike the single-group design, the results from this linking study cannot be evaluated using the estimates of the individual $\tau$ parameters; rather they are evaluated by comparing their distributions for the two samples of test takers. Table 1 gives the means and standard deviations of these distributions. The means were all equal to .000 because they were set equal to this number in the first calibration or made equal to it by implicit or post hoc linking in the second calibrations. However, the standard deviations were close enough to suggest distributions of $\tau$ estimates for test takers sampled from a population with identical variance. The reduction of the sample size from $N = 500$ to 200 should lead to larger deviations of the sample distributions of the $\tau$ estimates from the population distribution of $\tau$, which was uniform over $[-2, 2]$. The histograms in Figure 6 confirm this expectation.

Item 1 was also kept constant in this part of the study. The estimates of $\beta_1$ in the alternative second calibration runs were equal to 3.257, 3.256, and 3.252, respectively. This is somewhat larger than the estimate in the standard calibration (3.080). But this initial estimate was smaller than the true value $\beta_1 = 3.107$ of this parameter.

**Anchor-Item Design**

Two versions of the anchor-item design were studied. In the first version, the set of items in the second calibration shared the first five items with the standard tests. The other 45 items had new parameters sampled from $\alpha \sim U(1, 2)$ and $\beta \sim U(4, 5)$. Also, a new calibration sample of $N = 500$ test takers was drawn from $\tau \sim N(1, .5)$. Thus, the second calibration involved more time-intensive (but less discriminating) items in combinations with test takers from a population working at a higher speed. For the first five items in the standard test, the calibration yielded a mean estimate $\mu_{\hat{\beta}(1)} = 4.190$. Therefore, in the run with implicit linking, the identifiability restriction $\mu_{\beta} = 4.190$ was imposed on Items 1-5. The other run had the standard restriction $\mu_\tau = 0$ on all test takers and was followed by post hoc linking.
Figure 6. Distributions of the estimated values of $\tau_j$ for the independent samples of test takers in a linking study with a randomly equivalent groups design study. For the sample size of $N = 500$, the distributions of the estimates are from (a) the standard calibration, (b) a second calibration with implicit linking, and (c) a second calibration with *post hoc* linking. For the size of $N = 200$, the distribution is from (d) a second calibration with *post hoc* linking.

The second version of the anchor-item design was identical except for the fact that $K = 10$ items were common between the first and the second calibrations. The 10 items had a mean estimate $\mu_{\hat{\beta}(1)} = 3.880$ in the first calibration, and the same mean was imposed on the anchor items in the second calibration to guarantee implicit linking.

The linking constant was estimated to be equal to $v = 4.190 - 3.240 = .950$ for the design with $K = 5$ and $v = 3.880 - 2.908 = .972$ for the design with $K = 10$ anchor items. In Figure 7, the two sets of estimates of the time intensity parameters $\beta_i$ of the common items in the two versions of the linking design are compared both for the runs with implicit and *post hoc* linking. Again, each of the comparisons shows estimates that are close to identical.
**Concluding Comments**

As already noted, linking of RT parameters is susceptible to the same implementation problems as linking of response parameters. For instance, if test takers would be able to identify the anchor items in the tests and take them less seriously, the validity of both types of linking would be threatened. The same would occur when the introduction of a new time limit changed the degree of speededness of the tests.

An important difference between the linking of response and RT parameters is the potential of dimensionality problems. If the dimensionality of the test forms in a testing program changes over time (e.g., because a new generation of items appears to be more language dependent than before), attempts to link new items to an old ability scale become precarious. Such problems do not exist for RT models.

Unlike the latent abilities measured by test items, RT always remains the same univariate observed variable. It does, therefore, not make any sense to consider alternative “dimensions of speed,” let alone think of speed as a potential multidimensional
concept. Candidates may work faster or slower on different types of test items but this is not a dimensionality issue. Such changes would violate the assumption of constancy of speed that underlies the use of the RT model in (2) but never imply a change from one speed dimension to another. (The assumption of constancy of speed is equivalent to that of the constancy of ability in response modeling. Both are related through a speed-accuracy tradeoff; see van der Linden, 2009a.)

This issue of dimensionality reveals another important difference between the calibration of test items under the 3PL response model and the RT model. Earlier, we called attention to the fact that discrimination parameters in the RT model are less sensitive to poor identifiability and estimated more accurately than their counterparts in the 3PL model. Also, the simpler identifiability restrictions for the RT model lead to linking constants with more robust estimates (sample means) than constants such as in the Stocking-Lord procedure. In fact, as shown elsewhere (van der Linden, Klein Entink, & Fox, in press), when test items are calibrated jointly under the two models, that is, in a hierarchical framework with second-level models for the distributions of their item and person parameters, some of these advantages are imported in the estimation of the IRT parameters.

Appendix: Least-Squares Solution

Estimates of the linking constant in (17) can be found as least-squares solutions to the problem of fitting a line with unit slope and unknown intercept to a plot of the two sets of speed parameter estimates. The problem is depicted in Figure A1. Unlike a regression problem, the least-squares criterion should be applied to the distances between the individual points in the plot and their orthogonal projections onto the line with the unknown intercept.

![Figure A1](image-url)

*Figure A1.* Derivation of the linking constant as a least-squares solution to the problem of fitting a line with unit slope and unknown intercept to a plot of two sets of parameters estimates.
Let \((x_k, y_k)\) denote the two estimates of a parameter in a design with \(K\) common items or persons. The line we want to fit is

\[ y = x + v \]  

(A1)

for the best value of \(v\).

The line through \((x_k, y_k)\) orthogonal to (A1) has the equation

\[ y = -x + x_k + y_k. \]  

(A2)

It is easy to verify that the point of intersection of (A1) and (A2) is

\[ \left[ \frac{x_k + y_k - v}{2}, \frac{x_k + y_k + v}{2} \right]. \]  

(A3)

Using Pythagoras’ theorem, the squared distance between \((x_k, y_k)\) and its projection onto (A1) can be written as

\[ (x_k - y_k + v)^2/2. \]  

(A4)

The least-squares criterion finds \(v\) as the solution of

\[ \min_b \sum_{k=1}^{K} (x_k - y_k + v)^2. \]  

(A5)

Differentiating (A5) and setting the result equal to zero gives

\[ v = \mu_y - \mu_x. \]  

(A6)

Linking constant \(v\) is thus equal to the difference between the two average estimates.

**Acknowledgments**

An earlier part of this study 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 computational assistance of Wim M. M. Tielen is gratefully acknowledged.

**References**


**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.