
LSAC RESEARCH REPORT SERIES

■ **Linking Response-Time Parameters Onto a
Common Scale**

Wim J. van der Linden
University of Twente, Enschede, The Netherlands

■ **Law School Admission Council**
Research Report 08-02
March 2008

A publication of the Law School Admission Council



The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States, Canada, and Australia. Headquartered in Newtown, PA, USA, the Council was founded in 1947 to facilitate the law school admission process. The Council has grown to provide numerous products and services to law schools and to more than 85,000 law school applicants each year.

All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also members. Accredited law schools outside of the United States and Canada are eligible for membership at the discretion of the LSAC Board of Trustees.

© 2009 by Law School Admission Council, Inc.

All rights reserved. No part of this work, including information, data, or other portions of the work published in electronic form, may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, 662 Penn Street, Box 40, Newtown, PA 18940-0040.

This study is published and distributed by LSAC. The opinions and conclusions contained in this report are those of the author(s) and do not necessarily reflect the position or policy of LSAC.

Table of Contents

Executive Summary	1
Abstract	1
Introduction	1
Response-Time Model	2
Response Model	3
<i>Identifiability of the 3PL Model</i>	3
<i>Linking IRT Calibrations</i>	4
Identifiability of the Lognormal Response-Time Model	5
<i>Convenient Identifiability Restrictions</i>	6
Linking Procedures	7
<i>Single Group Design</i>	8
<i>Randomly Equivalent Groups Design</i>	8
<i>Anchor Item Design</i>	9
<i>Concurrent Calibration Design</i>	9
Standard Errors of Linking	9
Empirical Examples	11
<i>Parameter Estimation Error</i>	12
<i>Single Group Design</i>	13
<i>Randomly Equivalent Groups Design</i>	13
<i>Anchor Item Design</i>	14
Concluding Comments	15
References	16
Appendix: Least-Squares Solution	17

Executive Summary

In the analysis of data from the Law School Admission Test (LSAT) and similar standardized tests, a mathematical model called item response theory (IRT) is commonly used to estimate both the characteristics of the test questions (items) and the ability level of the test takers. Typical item-level statistics (called item parameters) estimated by an IRT model are difficulty, discrimination (i.e., the power of an item to distinguish between more able and less able test takers), and susceptibility to guessing. This research addresses the case of a testing program characterizing test items not only with respect to the IRT parameters, but also for a response-time model. The advantages of evaluating these factors simultaneously include the opportunity to check test items for dysfunction, to evaluate test forms with regard to speededness (i.e., the extent to which test takers run out of time before finishing the test), and to diagnose the test takers' RTs for possible aberrances (e.g., answer copying) during the test. In addition, as shown in an earlier report for the Law School Admission Council, calculating these statistics simultaneously results in more stable estimates.

Although RTs on test items are recorded on a natural scale (e.g., in seconds), the scale for some of the parameters in the lognormal RT model being applied is not fixed. As a result, when the model is used to estimate item parameters, the estimates from different samples have to be mapped onto a common scale. Such mappings are possible if the samples have a design with common items (an anchor test design) and/or involve common test takers (either a single group or partially overlapping groups) or randomly equivalent samples of test takers.

In this research, several combinations of such linking designs and linking procedures that map the parameter estimates onto a common scale are examined and the precision of the results evaluated. Linking designs with a single group tend to outperform the anchor test design and the randomly equivalent groups design. In addition, for larger samples, the anchor test design produces better linking than the randomly equivalent groups design.

Abstract

Although response times (RTs) on test items are recorded on a natural scale, the scale for some of the parameters in the lognormal RT model is not fixed. As a result, when the model is used to periodically calibrate new items in a testing program, the parameters are not automatically mapped onto a common scale. Several combinations of linking designs and procedures for the lognormal model that may map the parameter estimates onto a common scale are examined. For each of the designs, the standard error of linking is derived. The results are illustrated through the use of empirical examples.

Introduction

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. To exploit the information, however, 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 (2006a). This model, which is briefly reviewed below, has separate parameters for the person and item effects on RT distributions, very much like the regular item response theory (IRT) model for the response distributions of test items. The separation of these effects makes it 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 test (van der Linden, 2007, October) and an adaptive test (van der Linden, 2009); test hypotheses about the joint impact of attributes of the items on their difficulties and time intensities (Klein Entink, Fox, & van der Linden, 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).

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 a substantial amount of additional work or time; in fact, useful (Bayesian) procedures for estimating both the response and RT parameters are now available that 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 (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. As will become clear below, however, this impression is incorrect.

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 as an IRT version of the older problem of 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.

Response-Time 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 is reminiscent 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; however, as will be discussed below, the analogy is only superficial. 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 three-parameter logistic (3PL) response model. (For a more detailed discussion along these lines, see van der Linden [2006a]).

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, in press). Let t_{ij} denote the RT for test taker j recorded on item i , τ_j the speed at which j works, and β_i the amount of labor required by item i . The model then follows as

$$\ln T_{ij} = \beta_i - \tau_j + \epsilon_i, \quad \epsilon_i \sim N(0, \alpha_i^{-2}), \quad (1)$$

that is, as a normal density of the log RT with mean $\beta_i - \tau_j$ and discrimination α_i for item i (i.e., reciprocal of the standard deviation). 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} :

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) \equiv \frac{\alpha_i}{t_{ij} \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left[\alpha_i (\ln t_{ij} - (\beta_i - \tau_j)) \right]^2 \right\}. \quad (2)$$

Procedures and software for estimating the parameters in the model and assessing its fit to empirical RTs have been presented elsewhere (Fox, Klein Entink, & van der Linden, 2007; Klein Entink, Fox, van der Linden, 2009; van der Linden, 2006a, 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 Introduction, above).

Although estimation of the model is rather straightforward, one technical issue has to be addressed in 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 $j = 1, \dots, N$ on item $i = 1, \dots, n$. One of the mainstream models for dichotomous response variables is the 3PL model:

$$\Pr\{U_{ij} = 1\} = p_i(\theta_j; a_i, b_i, c_i) \equiv c_i + (1 - c_i)\Psi(\theta_j), \quad (3)$$

with

$$\Psi(\theta_j) \equiv \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}. \quad (4)$$

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, respectively, for item i (Lord, 1980). If $c_i = 0$, the model becomes the 2PL model, which is just the logistic function in (4).

Observe that the lognormal model in (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 (3) and (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):

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

Therefore, although the parameter structure of the lognormal model is reminiscent 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, α_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 $p_i(\theta; a_i, b_i)$ and $\{p_i(\theta; a_i, b_i)[1 - p_i(\theta; a_i, b_i)]\}^{1/2}$, respectively. (Discrimination parameter a_i does have a direct impact on the standard deviation of U_{ij} , though; see van der Linden, 2006a, Fig. 1). For the current research, however, the most important difference between the two models is that the key expression $\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))$ 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} . Instead, the formal position of $\ln t_{ij}$ has now been taken by the unknown ability variable θ_j .

Identifiability of the 3PL Model

The response model in (3) and (4) has an indeterminacy of the origin and unit of scale of θ ; that is, the response function does not change if the origin is changed (i.e., the same constant is added to all θ_j and b_i) or the same occurs to the unit of scale (i.e., all θ_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 (Casella & Berger, 2002, sect. 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 relationship 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 consists 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 if 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 θ_j equal to zero and one, respectively, across the test takers in the sample.

Although popular, these restrictions are nothing but conventions. For a single item calibration, 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 value of one of these parameters and the difference between two others) would yield equally good estimates. However, the choice becomes critical when later calibrations of new items for the same testing program have to be on a scale that has already been fixed. We are then faced with a problem of parameter linking.

Linking IRT Calibrations

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 for a possible change of the origin and/or unit of the scale for the new estimates. The adjustment is required because *formally identical restrictions are not necessarily empirically identical* (i.e., 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 if 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. A linking transformation should then be used to adjust one of the sets of parameter estimates for the differences between these restrictions.

A popular procedure for parameter linking in IRT is the Stocking–Lord (1983) procedure. The procedure assumes a set of common items (anchor items) and finds the linking transformation as the linear transformation that minimizes the differences between the sum of the response functions for the common items in the two calibrations at a grid of well-chosen θ values. More formally, suppose we have common items $k = 1, \dots, K$ with estimates $(a_k^{(1)}, b_k^{(1)}, c_k^{(1)})$ and $(a_k^{(2)}, b_k^{(2)}, c_k^{(2)})$ in the two calibrations. The goal is to find the transformation that maps the second set on scale for the first set; that is,

$$\theta^{(1)} = u\theta^{(2)} + v. \quad (6)$$

The Stocking–Lord procedure finds the constants u and v in (6) by minimizing the squared differences of sums

$$\left[\sum_{k=1}^K p_k(\theta; a_k^{(1)}, b_k^{(1)}, c_k^{(1)}) - \sum_{k=1}^K p_k\left(\theta; \frac{a_k^{(2)}}{u}, ub_k^{(2)} + v, c_k^{(2)}\right) \right]^2 \quad (7)$$

over a grid $\{\theta_1, \theta_2, \dots, \theta_m\}$.

An advantage of the Stocking–Lord procedure is its robustness with respect to small variations between item parameter values that hardly have any impact on the shape of the response functions of the items. Alternative procedures find the linking constants u and v , for example, through comparisons of the mean estimates of the parameters $a_i^{(1)}$ and $a_i^{(2)}$ and $b_i^{(1)}$ and $b_i^{(2)}$, respectively. For a review of several of these alternatives, see Kolen and Brennan (2004, sect. 6.3).

It is a good point to highlight the subtle differences between traditional observed-score equating and parameter linking in IRT applications alluded to earlier. In the former, we have to find the family of transformations between the observed scores on two different tests that adjusts for the differences between their items and the abilities of the persons to which they are administered (van der Linden, 2006b). In the latter, the IRT model already accounts for these differences, but we

now have to adjust for the differences in the scale of its parameters when two different sets of identifiability restrictions are imposed. If the restrictions are identical, no adjustment is necessary, even when, for instance, there are large differences between the ability distributions of the two groups of test takers. Examples of automatic parameter linking for the lognormal RT model through identical identifiability restrictions are given later in this paper.

Identifiability of the Lognormal Response-Time 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, i.e., with the substitution of

$$\mu_{ij} \equiv \beta_i - \tau_j \quad (8)$$

into (2), is identifiable. In this version, μ_{ij} is the mean of $\ln T_{ij}$, and α_i is the reciprocal of its standard deviation. For m realizations of $\ln T_{ij}$ (i.e., independent observations of the RTs of test taker j on item i), the maximum-likelihood estimates (MLEs) of the two parameters are defined by the following simple expressions:

$$\hat{\mu}_{ij} = m^{-1} \sum \ln t_{ij} \quad (9)$$

$$\hat{\alpha}_i = m(\sum \ln t_{ij} - \hat{\mu}_{ij})^{-1/2}. \quad (10)$$

This argument shows that the RT model in (4) can only have an identifiability problem for its time intensity parameters β_i and speed parameters τ_j . That it does have this problem is demonstrated by the fact that, for any value of ε , the distribution of $\ln T_{ij}$ remains the same if we replace β_i and τ_j by $\beta_i - \varepsilon$ and $\tau_j - \varepsilon$.

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

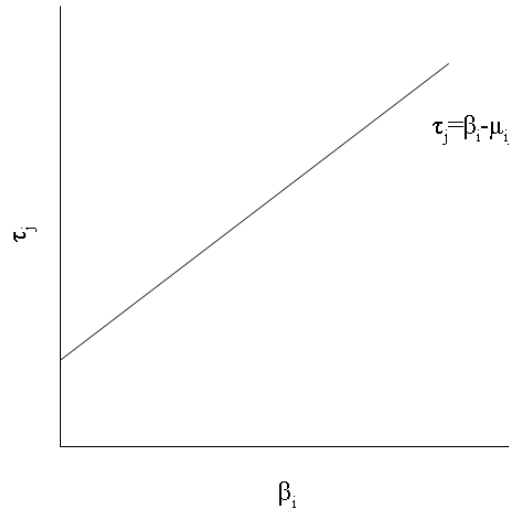


FIGURE 1. Graphical illustration of the unidentifiability of speed parameter τ_j and time intensity parameter β_i . The model constrains both parameters only to be on the line $\tau_j = \beta_i - \mu_{ij}$.

It is interesting to observe that, unlike discrimination parameter a_i in the 2PL or 3PL model, α_i is always identifiable. This point is illustrated empirically in Figure 2, where the estimates of α_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, and we did not 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.

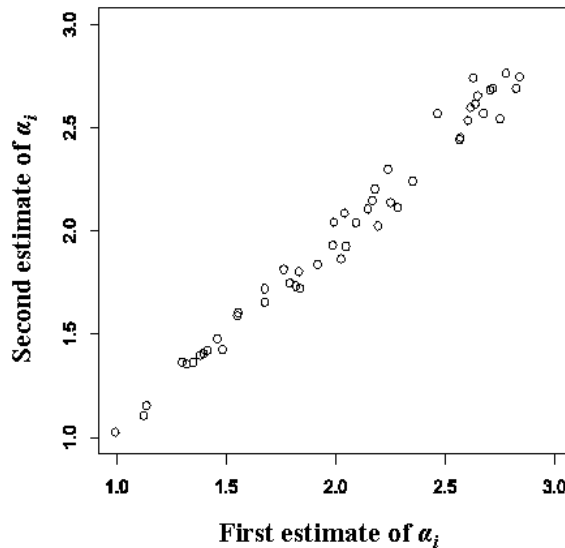


FIGURE 2. Empirical illustration of the identifiability of discrimination parameter α_i . Except for random error, the estimates of the discrimination parameters from two unrelated calibrations are automatically on the same scale.

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 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 (β_i, τ_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 β_i is identifiable. The same then holds for the speed parameter τ_{j_0} of any other test taker j_0 who responds to item i . The claim is true because μ_{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 β_{i_0} of all other items $i_0 \neq i$ to which a test taker 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 it is necessary for it 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. Alternatively, we could fix the mean of all speed parameters at a constant c . For $c = 0$ the choice leads to

$$\mu_\tau = 0, \quad (11)$$

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 (8),

$$\epsilon(\ln T) = \mu_\beta - \mu_\tau, \quad (12)$$

where the expectation is taken over replications, test takers, and items. Thus, the identifiability constraint in (11) implies

$$\mu_\beta = \epsilon(\ln T). \quad (13)$$

This consequence enables us to interpret the values of the time intensity parameters β_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, (11) sets the average estimate of the time intensity parameters equal to the average logtime in the dataset; that is, the restriction leads to

$$\widehat{\beta} = (nN)^{-1} \sum_{i=1}^n \sum_{j=1}^N \ln t_{ij}. \quad (14)$$

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

$$\mu_\beta = 0. \quad (15)$$

Obviously, this restriction implies

$$\mu_\tau = \epsilon(\ln T), \quad (16)$$

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 (11) or (15) 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 missed this, we might 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 (von Davier, Holland, & Thayer, 2005, chap. 2). Although linking problems are of a different nature, the terminology for these designs is well established; 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 is 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 have equal units, $u = 1$ in (6) and the transformation can be written as

$$\tau^{(1)} = \tau^{(2)} + v. \quad (17)$$

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 or in the population of test takers (e.g., a change in the degree of speededness of the test). The same problems are a threat to parameter linking in IRT calibration (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 close 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 τ s. Let $\mu_{\tau^{(1)}}$ be the average estimate of the τ 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:

$$N^{-1} \sum_{j=1}^N \tau_j^{(2)} = \mu_{\tau^{(1)}}. \quad (18)$$

Post hoc linking with a single group design involves estimating the unknown constant v in the linking transformation in (17). Averaging the transformation over the common test takers gives

$$v = \mu_{\tau^{(1)}} - \mu_{\tau^{(2)}}. \quad (19)$$

Thus, the post hoc linking transformation can be written as

$$\tau^{(1)} = \tau^{(2)} + \mu_{\tau^{(1)}} - \mu_{\tau^{(2)}}. \quad (20)$$

Of course, the transformation has to be applied to the estimates of both the τ_j and β_i parameters.

It is interesting to note that the estimate of the linking constant in (19) 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 proved in the Appendix.

It is not necessary for all test takers in the second calibration to be common. Both procedures can also be applied to a common subgroup among the 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 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 (20). 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 (18) and (20) 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:

$$K^{-1} \sum_{k=1}^K \beta_j^{(2)} = \mu_{\beta}^{(1)}, \quad (21)$$

where $k = 1, \dots, K$ are the common items and $\mu_{\beta}^{(1)}$ is their average estimate from the first calibration.

Alternatively, analogous to (20), if the second set of items has already been calibrated, post hoc linking using the transformation

$$\beta^{(1)} = \beta^{(2)} + \mu_{\beta}^{(1)} - \mu_{\beta}^{(2)} \quad (22)$$

is still possible. Although the transformation is calculated from the parameter estimates for the common items only, it should also be applied to the β parameters of all items (as well the speed parameters τ_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 that does not have the clear-cut pattern of a single group or anchor test design. If the design (a) is still connected and (b) contains some of the items and/or test takers from an earlier calibration in which the scale of the parameters is 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 (18) or (21), 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 (20) and/or (22), 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 τ_j and β_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 (Lehmann, 1999, ex. 7.1.4), for known parameters α and β , Fisher's information about τ_j in the logRTs on an item is equal to $I(\tau) = \alpha^2$. Therefore, the (asymptotic) variance of the maximum-likelihood estimator of τ_j from the items $i = 1, \dots, n_i$ in the first calibration is equal to

$$\left(\sum_{i=1}^{n_i} \alpha_i^2 \right)^{-1}. \quad (23)$$

This expression assumes local independence between the logRTs, which seems reasonable for an RT model for test takers operating at a constant speed (van der Linden, 2007).

Observe that the variance is independent of the true value of τ_j . Consequently, for a group of N test takers, the sampling variance of the estimator of the mean μ_τ is asymptotically equal to

$$N^{-1} \left(\sum_{i=1}^{n_1} \alpha_i^2 \right)^{-1}. \quad (24)$$

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

$$\sigma_{\text{SG}} = N^{-1/2} \left\{ \left(\sum_{i=1}^{n_1} \alpha_i^2 \right)^{-1} + \left(\sum_{i=1}^{n_2} \alpha_i^2 \right)^{-1} \right\}^{1/2}. \quad (25)$$

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 σ_τ be the standard deviation of τ in this population. Because the two types of errors are independent, an approximate standard error of linking for this type of design is

$$\sigma_{\text{REG}} = \left\{ (N_1 + N_2)^{-1} \sigma_\tau^2 + N_1^{-1} \left(\sum_{i=1}^{n_1} \alpha_i^2 \right)^{-1} + N_2^{-1} \left(\sum_{i=1}^{n_2} \alpha_i^2 \right)^{-1} \right\}^{1/2}. \quad (26)$$

The derivation of the standard error for the anchor item design runs analogous to that of (25). For known parameters α and τ , it holds that $I(\beta) = \alpha^2$. In the first calibration, the parameters β_k of anchor items $k = 1, \dots, K$ are estimated from the RTs of test takers $j = 1, \dots, N_1$. The estimate of β_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 estimator of μ_β for the entire set of anchor items has asymptotic variance

$$(KN_1)^{-1} \sum_{k=1}^K \alpha_k^{-2}. \quad (27)$$

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

$$\sigma_{\text{AI}} = K^{-1/2} \left\{ (N_1^{-1} + N_2^{-1}) \sum_{k=1}^K \alpha_k^{-2} \right\}^{1/2}. \quad (28)$$

All three standard errors decrease with the number 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 τ s estimated from the sets of items in the two calibrations. Hence, the sums of the discrimination parameters in (25) and (26) are over all these items. For the anchor test design, the linking constant is the difference between the estimates of the means of the β 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 design and the randomly equivalent groups design and the standard errors for the anchor item design. Because the number of items to be calibrated is always larger than the number of anchor items, the first designs 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 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 τ 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 dataset 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 in standard errors between this design and the randomly equivalent groups design is due to the role of σ_τ , 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 σ_τ . In this example, the discrimination parameters in the second calibration were somewhat larger for the single group design than for the other two designs (see below). Also, in some applications, we may be able to manipulate the standard deviation.

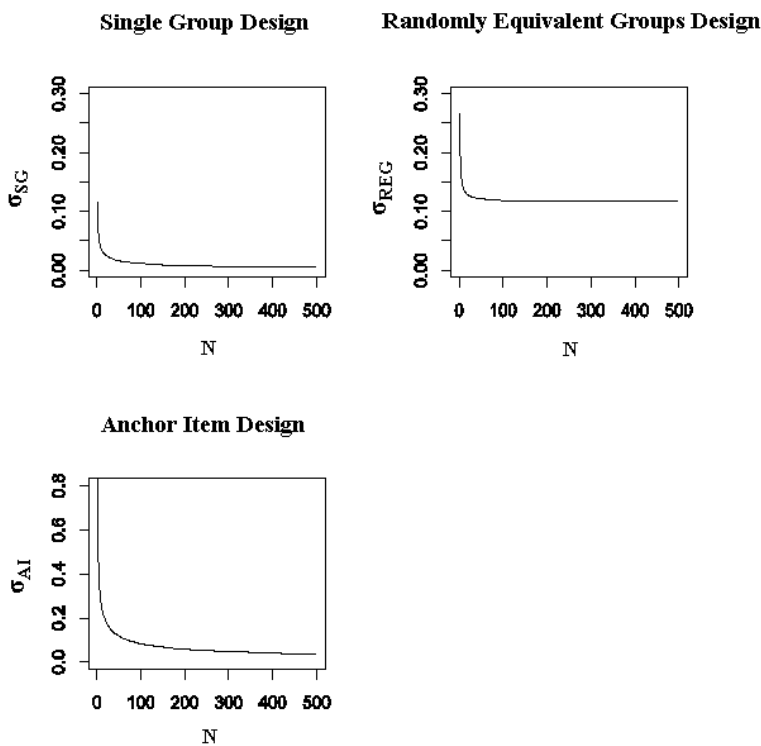


FIGURE 3. Standard errors of linking for the single group, randomly equivalent groups, and anchor item designs as a function of the size of the calibration samples

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 are due to changes in the test takers between test administrations. 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 that is typical of nearly all statistical optimization problems.

Empirical Examples

The goal of these examples is to illustrate the different linking designs and transformations for tests with realistic ranges of RT parameters. Because we wanted to evaluate the results of each linking relative to a fixed scale, we adopted a standard test of $n = 50$ items. Parameters α_i and β_i , $i = 1, \dots, 50$, 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 $N = 500$ test takers with parameters τ_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 dataset (van der Linden, 2006a).

All parameters were estimated using the Markov chain Monte Carlo (MCMC) procedure (Gibbs sampler) for the lognormal RT model presented in same reference. For this procedure, it is easy to impose identifiability restrictions on the τ_j or β_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 β_i parameters, the mean for this subset is subtracted from the sampled value for each of the β_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 MLEs, the prior distributions were chosen to be virtually noninformative (e.g., the priors for the τ_j parameters were normals with a standard deviation equal to 1,000); for more details, see van der Linden (2006a).

Each implicit or post hoc linking was evaluated through a comparison between (a) the estimates of the τ_j or β_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 α_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 τ_j parameters had little error, even though each of them was based on the RTs on 50 items only. The estimates of the β_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 α_i parameters in the RT model were quite satisfactory, the reason being their shared scale with the observed logRTs. 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 θ s.

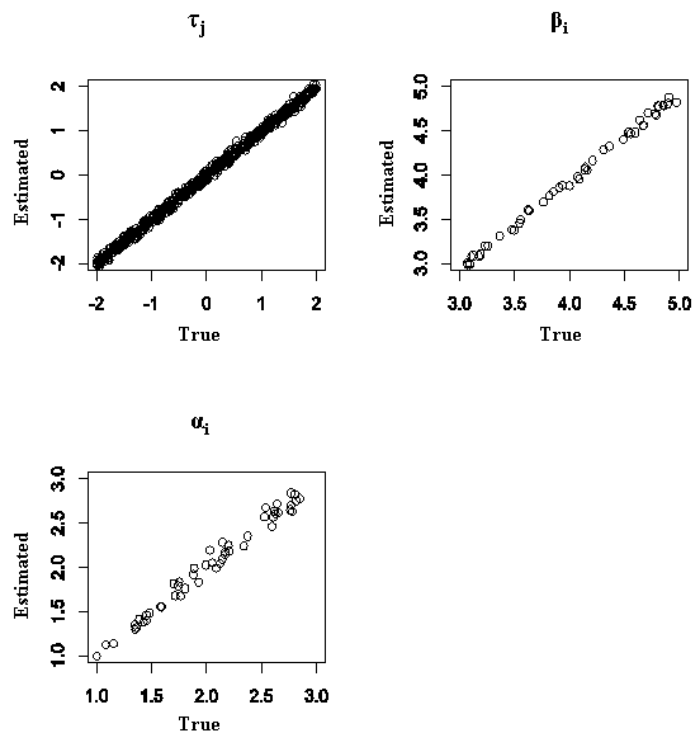


FIGURE 4. True and estimated values of τ_j , β_i , and α_i for the standard dataset in the empirical linking study ($N = 500$ test takers; $n = 50$ items)

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 τ s from the calibration of the standard test was $\mu_{\tau^{(1)}} = .000$ (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 τ s were on the same scale.

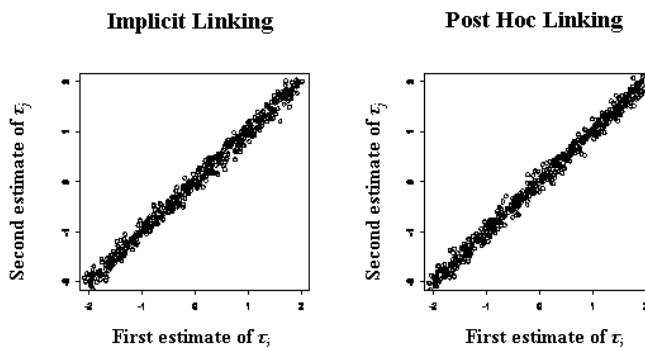


FIGURE 5. Comparison of the estimated values of τ_j for $N = 500$ test takers from two different tests in a linking study with a single group design study with (a) implicit linking and (b) post hoc linking

For the second run, the arbitrary choice $\mu_\tau = 1$ was made for the identifiability restriction. The mean estimate of the τ s was now $\mu_{\tau^{(2)}} = .999$. 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 [20]). The second plot in Figure 5 shows that, for this type of linking, the second estimates of τ 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 calibrations 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 β_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 β_1 , estimated to be equal to .045, .046, and .044, respectively, in the three different calibrations.

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 τ parameters, but only 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 τ estimates for test takers sampled from

a population with identical variance. The reduction of the sample size from $N = 500$ to $N = 200$ should lead to larger deviations of the sample distributions of the τ estimates from the population distribution of τ , which was uniform over $[-2, 2]$. The histograms in Figure 6 confirm this expectation.

TABLE 1
Mean and SDs of τ estimates

	Mean	SD
First calibration	.000	1.166
Implicit linking ($N = 500$)	.000	1.140
Post hoc linking ($N = 500$)	.000	1.140
Post hoc linking ($N = 200$)	.000	1.163

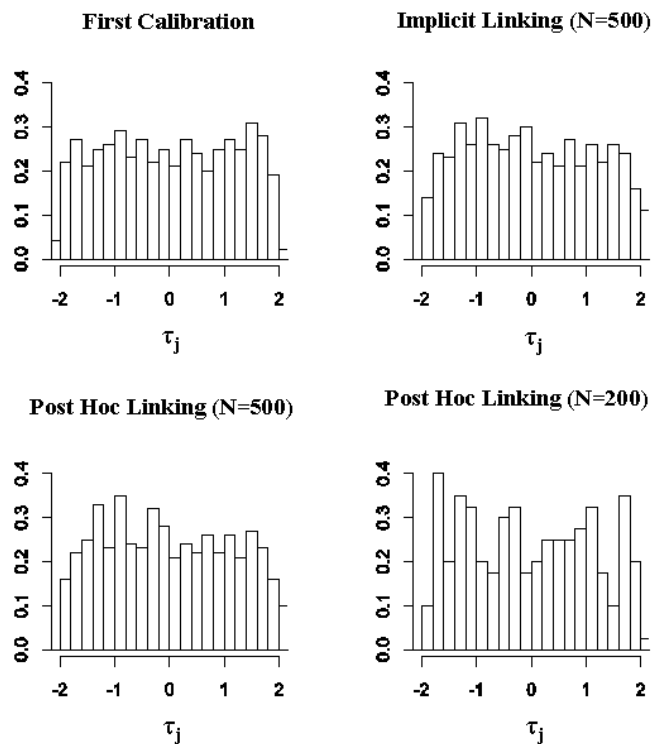


FIGURE 6. Distributions of the estimated values of τ_j for the independent samples of test takers in a linking study with a randomly equivalent groups design study. For the sample size of $N = 500$, the distributions of the estimates are from (a) the standard calibration, (b) a second calibration with 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.

Item 1 was also kept constant in this part of the study. The estimates of β_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 combination 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.

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 $\nu = 4.190 - 3.240 = .950$ for the design with $K = 5$ and $\nu = 3.880 - 2.908 = .972$ for the design with $K = 10$ anchor items. In Figure 7, the two sets of estimates of the time intensity parameters β_i of the common items in the two versions of the linking design are compared for the runs with both implicit and post hoc linking. Again, each of the comparisons shows estimates that are close to identical.

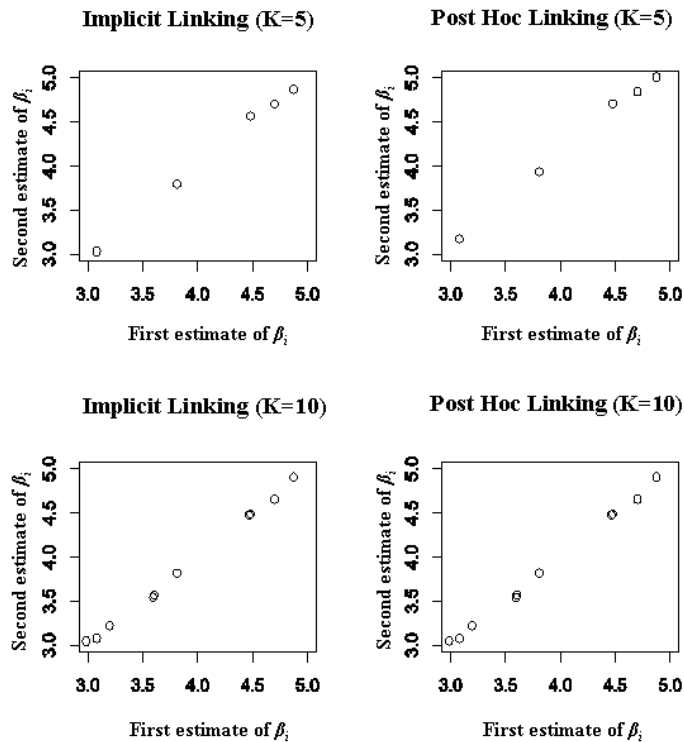


FIGURE 7. Comparison of the estimated values of β_i for $K = 5$ and $K = 10$ common items in linking study with an anchor item design with (a) implicit and (b) post hoc linking

Concluding Comments

As already noted, linking of RT parameters is liable to the same implementation problems as linking of response parameters. For instance, if test takers were 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 if the introduction of a new time limit changed the degree of speededness of the tests.

An important problem inherent in the linking of response and RT parameters is the potential role 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 not, therefore, 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 a different issue. Such changes would violate the assumption of constancy of speed that underlies the use of the RT model in (2) but would 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, in press.)

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 already 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 (7). In fact, as shown elsewhere (van der Linden, Klein Entink, & Fox, in press), when test items are calibrated jointly under the two models (i.e., 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.

References

- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.
- 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.
- Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*, 59–77.
- Johnson, N. I., & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions* (Vol. 1). New York: Wiley.
- Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to simultaneous modeling of accuracy and speed on test items. *Psychometrika*, *74*, 21–48.
- Kolen, M. K., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.
- Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200–219.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.
- van der Linden, W. J. (2006a). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.
- van der Linden, W. J. (2006b). Equating error in observed-score equating. *Applied Psychological Measurement*, *30*, 355–378.
- 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. (2007, October). *Test design and speededness*. Paper presented at the 23rd Workshop on Item Response Theory, University of Twente, Enschede, The Netherlands, October 9–11, 2007.
- van der Linden, W. J. (2008). Using response times for item selection in adaptive tests. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.
- van der Linden, W. J. (2009). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*, 25–41.

- van der Linden, W. J. (in press). Conceptual issues in response-time modeling. *Journal of Educational Measurement*.
- 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.
- van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (in press). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2005). *The kernel method of test equating*. New York: Springer.

Appendix: Least-Squares Solution

The estimates of the linking constants in (19) and (A2) are a least-squares solution to the problem of fitting a line with unit slope and unknown intercept to a plot of their two sets of 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.

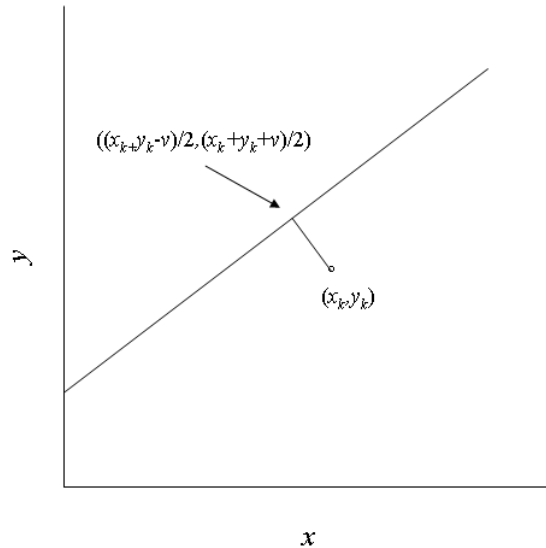


FIGURE A1. Derivation of the linking constant as a least-squares solution to the problem of fitting a line with unit slope and unknown intercept to a plot of two sets of parameter estimates

Let (x_k, y_k) denote the two estimates of a parameter in design with K common items or person. The line we want to fit is

$$y = x + v \quad (\text{A1})$$

for the best value of v .

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

$$y = -x + x_k + y_k. \quad (\text{A2})$$

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

$$((x_k + y_k - v)/2, (x_k + y_k + v)/2). \quad (\text{A3})$$

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

$$(x_k - y_k + v)^2 / 2. \tag{A4}$$

The least-squares criterion finds v as the solution of

$$\min_b \sum_{k=1}^K (x_k - y_k + v)^2. \tag{A5}$$

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

$$v = \mu_y - \mu_x.$$

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