Linking Response Model Parameters

Michelle D. Barrett

RESPONSE MODEL PARAMETER LINKING

MICHELLE D . BARRETT

Graduation Committee

Chairman	Prof. Dr. T. A. J. Toonen
Promotor	Prof. Dr. W. J. van der Linden
Members	Prof. Dr. M. P. F. Berger
	Prof. Dr. T. J. H. M. Eggen
	Prof. Dr. Ir. J.-P. Fox
	Prof. Dr. H. Holling
	Prof. Dr. Ir. B. P. Veldkamp
	Prof. Dr. J. Vermunt

Michelle D. Barrett
Response Model Parameter Linking
Ph.D. thesis, University of Twente, Enschede, the Netherlands
ISBN: 978-90-365-3911-1
DOI: 10.3990/1.9789036539111
Printed by Ipskamp Drukkers, B.V., Enschede

RESPONSE MODEL PARAMETER LINKING

DISSERTATION

to obtain
the degree of doctor at the University of Twente
on the authority of the rector magnificus,
Prof. Dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday, October 7th, 2015 at 16:45

by

Michelle Derbenwick Barrett
born on May 8th, 1974
in Summit, New Jersey, USA

This dissertation has been approved by the promotor:
Prof. Dr. W. J. van der Linden

Acknowledgments

As the past several years of research culminate in this thesis, I would like to thank the many people who have contributed to my efforts.

First and foremost, I would like to thank my supervisor Prof. Dr. Wim van der Linden. It has been a great honor and my pleasure to work with and learn from you. It has always been important to me that my work be grounded in theory yet practical in use. I couldn't have asked for a supervisor better aligned to those goals; your dedication to formal analysis of psychometric and research methods and your deep understanding of the practical issues the educational assessment field faces challenged me to attend to both theory and practice every step of the way. I know that this experience will help me continue to pursue theoretically valid solutions with practical relevance.

For the most part I completed this work while also employed at McGraw-Hill Education CTB. I am grateful for the many colleagues there who contributed to a stimulating and challenging research environment that fertilized my efforts. I could not have understood as deeply the context, the challenges, and the impact of research in psychometrics without knowing each of you. In particular, I am lucky to have known Gary Schaeffer, who recognized my potential, helped to identify this opportunity to pursue my doctorate, and mentored me through many career advancements and personal milestones. I still rely on the many lessons he taught me. I also owe thanks to Rich Patz, Seung Choi, and Craig Mills, each of whom provided professional guidance (and always deadlines!).

To my colleagues across the ocean at the University of Twente and CITO, thank you for your thoughtful questions about my research and for your hospitality. I have always felt welcome during my visits to the University. In addition, I found the RCEC IRT workshop each year to be an excellent opportunity for me to broaden my horizon beyond United States K-12 educational assessment.

Finally, I would like to thank all of my family and friends as this thesis would not have been possible without your support. Mom and Dad, thank you for the gifts of curiosity, enjoyment of a good problem, and persistence. Frankie, thank you for your love, encouragement, enthusiasm, and much-needed humor. And to my daughters Ayden and Drew, who have for most of their lives only known a mom who was writing her dissertation while also working a full time job, thank you for your patience. You are my inspiration to improve the methods by which your teachers can understand the many fantastic things you and your friends know and can do.

Michelle D. Barrett

Enschede, October 2015

Contents

List of Figures

List of Tables

Symbols

a_i	discrimination parameter of item i
b_i	difficulty parameter of item i
c_i	guessing parameter of item i
c	index used to denote a categorical attribute or constraint
d	dummy index used in polytomous model probability expressions, $d = 2, ..., K$
g	equal distance between an item's adjacent category b parameters
h	dummy index used in polytomous model probability expressions, $h = 1, ..., K$
i	item index, $i = 1, ...I$
j	index for items available as non-common items $j = 1, ..., J$
k	item response category index, $k = 1, ..., K$
k	step index, $k = 1, ..., K - 1$, where K is number of response categories
m	minimal linking element index (item, pair of items, pair of test-takers), $m = 1, ..., M$
p	test-taker index, $p = 1, ..., P$
q_m	value of attribute for linking item m for quantitative constraint q
q_j	value of attribute for non-common item j for quantitative constraint q
q	index used to denote a quantitative attribute or constraint
t	test administration index, $t = 1, 2$
U_{pik}	response by test-taker p to polytomous item i in category k
U_{pi}	response by test-taker p to dichotomous item i
u_m	u linking parameter for minimal linking element m
u	u linking parameter
v_m	v linking parameter for minimal linking element m
v	v linking parameter
V_c^{item}	set of items with categorical attribute c
V_c^{stim}	set of shared stimuli with categorical attribute c
w	w linking parameter (chap. 2, theorem 6)
x	constant (chap. 2, theorem 5)

x_m	decision variables for each of m linking items
x_j	decision variables for each of j non-common items
x_s	decision variables for each of s shared reading passages
y	common lower bound for IRT objective in optimal test design
s	index for shared stimuli, $s = 1, ..., S$
α_i	slope parameter of item i
β_i	intercept parameter of item i
γ_i	$1 - c_i$
δ_i	$c_i \exp(b_i)$, with $b_i = \ln \beta_i$ (chap. 2, theorem 2ii)
ζ	constant used in optimal linking design
η_{pik}	1 - π_{pik} (polytomous item)
η_{pi}	1 - π_{pi} (dichotomous item)
θ_p	person (test-taker) ability
θ_t	values of θ used for IRT objective function in optimal linking design
κ	constant (chap. 2, theorem 2iii)
κ	linking parameter for γ_i parameter (chap. 2, theorem 4)
λ_a	bound on number of all items in test design
λ_m	bound on number of common items in linking design
λ_q	bound on quantitative attribute q
λ_c	bound on categorical attribute c
λ_u	upper bound on σ_u
λ_v	upper bound on σ_v
ϑ_p	test-taker ability in slope-intercept parameterization of response model
π_{pik}	probability of a response by test-taker p to polytomous item i in category k
π_{pi}	probability of a correct response by test-taker p on dichotomous item i
$\sigma_{a_{ik_t}}$	SEE of a parameter for item i, response category k, administration t
σ_{u_m}	ASE of u linking parameter for minimal linking element m
σ_{v_m}	ASE of v linking parameter for minimal linking element m
σ_u	ASE of u linking parameter
σ_v	ASE of v linking parameter
ω_u	weight of u for optimal linking design
ω_v	weight of v for optimal linking design
$\mathbf{x}$	random vector
$\boldsymbol{\eta}$	vector of linking parameters, e.g., $\boldsymbol{\eta} = (u, v)$
$\boldsymbol{\xi}$	vector of response model parameters, e.g., $\boldsymbol{\xi} = (\theta, a, b, c)$

$\boldsymbol{\pi}$	vector-valued parameter
$\boldsymbol{\varphi}$	vector-valued parameter with a bijective relation to $\boldsymbol{\pi}$
$\Psi_{pik}(\boldsymbol{\xi})$	logistic function (both 2PL and step function)
$\varphi(\boldsymbol{\xi})$	linking function
$\boldsymbol{J}_{\varphi}$	Jacobian matrix associated with function $\varphi(\boldsymbol{\xi})$
$I_i(\theta)$	Fischer's information on θ in the response to item i
$U[\min, \max]$	Uniform distribution with minimum min and maximum max
$N(\mu, \sigma)$	Normal distribution with mean μ and standard deviation σ

Chapter 1

Introduction

1.1 Introduction

For measurement of ability, aptitude, or psychological traits, test-takers respond to items intended to represent the construct of interest. Statistical models are then applied to the response data to quantify the location of test-takers along a scale representing the measured construct. A simple example of a common statistic is the number of the items correct, for which test-takers with a higher number correct are seen as having "more" of the tested construct. In some instances, it is enough to understand how the test-takers who took that single test compare with one another. For those cases, using a number correct or other classical statistic along with a notion of measurement error may suffice.

However, in most applications of educational testing, users of the test results typically want to make comparisons across test forms (different items) and test administrations (different test-takers), including comparisons of test-taker ability distributions and the use of previously established cut-scores to make classification decisions. In this case, comparing the number-correct score across test administrations becomes inappropriate. As the difficulty of the items on the second test form may have increased or decreased in relation to the first test form, it is relatively more difficult or easier, respectively, for test-takers who take the second test to obtain the same number-correct score as their same-ability peers who took the first test. Indeed, it would be unfair to make the desired comparisons across test form and test administration.

There is an extensive literature on observed-score equating (OSE), which is a collection of procedures intended to allow comparison of observed test scores when different test forms are

administered. In such procedures, transformations (linear or otherwise) are claimed to make the scores on the second test form comparable with scores on the first test form. For example, equipercentile methods are intended to establish the transformation to map the number-correct score on the second form to the number-correct score on the first form for which the different test-takers have the same percentile rank. If a set of common items is administered, often referred to as an anchor set or the NEAT design, test-taker performance on the common set of items is used to try to establish the transformation. Often cited is a text by Kolen and Brennan (2004) that serves as a basic guide to these methods; von Davier (2013) provides a recent summary of newer OSE methods.

More appropriately, the response data may be modeled with an item response theory (IRT) response model. In IRT, the probability of a correct response to an item by a test-taker p on item i with ability $\theta_p \in \mathbb{R}$ is modeled using a second level of item and test-taker parameters as in

$$\Pr\{U_{pi} = 1; \theta_p\} \equiv p(\theta_p; a_i, b_i, c_i) \equiv c_i + (1 - c_i)\frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}, \tag{1.1}$$

with item difficulty $b_i \in \mathbb{R}$, item discrimination $a_i > 0$, and lower asyomptote to the probability $c_i \in (0, 1]$ for the well-known three-parameter logistic (3PL) response model. Ability and item parameters are estimated based on response data in a process called calibration. The distribution of the θ parameter yields the ability scale, and the score of test-taker p is an estimated θ_p. (Sometimes the estimated θ_p parameters are linearly transformed into scale scores, which is done simply to provide a score distribution that has a mean and standard deviation more likely to be meaningful to the test result user.) The use of IRT already makes adjustments to account for differences in difficulty among items on different test forms, as it separates the ability parameter from the parameters describing the items. Yet, statistical adjustments are still required when comparisons across test forms are desired. These statistical adjustments and the random error associated with them are the topic of this thesis.

1.2 Motivation

With a few exceptions, the problem of linking item response model parameters from different item calibrations has been conceptualized as an instance of the problem of equating observed scores on different test forms (e.g., Kolen & Brennan, 2004, p. 156). The process is described as such: (i) Estimate the item parameters in the response model for a new test form, (ii) scale the

parameters back to the base scale using a linear transformation, and (iii) if expected number-correct scores are used, convert the scores on the new form to expected number-correct scores on the old form and then to scale scores. Current response model parameter linking methods such as the mean/mean (Loyd & Hoover, 1983), mean/sigma (Marco, 1977), response-function (Haebara, 1980), and test characteristic curve (Stocking & Lord, 1983) methods assume that a simple a linear transformation of the θ scale is all that is required accordingly; this assumption was the foundation for the development of their methods. Perhaps the perception that response model linking is a special case of observed score equating has also been reinforced by the fact that the two problems require similar designs of common test items on the test forms to be compared. An analysis of the formal nature of the response model linking function for well-known response models is not present in the literature.

Attempts to establish quality of the currently used methods are made through simulation studies that vary dimensions such as number of common test items and the response generating θ distributions. Conclusions are drawn based on how closely the resulting score distribution matches the generating θ distribution subsequent to application of the methods. Along with missing a formal treatment of the linking function for each model, these studies miss a formal analysis of possible sources of error in the linking function estimates, with the exception of Ogasawara (2000, 2001, 2011) who derived the asymptotic standard error for each of the aforementioned linking methods for dichotomous response models. In fact, there appears to be confusion about sources of error in these linking transformations (e.g., Doorey, 2011) most likely due to the observed-score equating framework under which they were developed. Certainly, error in the scores from the second test administration introduced by linking with these methods is not reported in operational tests. Rather, the methods are applied operationally assuming that the item or ability parameters used to calculate the transformation are error-free.

This thesis argues that the use of item response models does not require any post hoc *observed-score equating*, but that the necessity of parameter linking is due to a fundamental problem inherent in the formal nature of these models—their general lack of *identifiability*. As illustrated in *Chapter 2*, common response models are not identified as multiple sets of item and ability parameters are available that do imply the same probability distributions of the responses. For example, multiplication of a_i in (1.1) by a constant u may be offset by multiplying both θ_p and b_i by $1/u$ to result in an identical probability π_{pi}. As a consequence, identification restrictions must be applied when estimating the parameters to select one set of parameters from the larger observationally equivalent set. It is the effect of these restrictions that prevent direct comparability across different calibrations, not the notion of an arbitrary scale for the θ parameter which ignores the linking of the other three (unidentified) parameters in a 3PL model

(and more in polytomous models). A correct theoretical framework of the linking procedure is critical if we are to understand response model linking functions, their estimators and their associated errors. Therefore, the motivation for this thesis lies in the reconceptualization of the response model linking problem to address the nature of the unidentified statistical model in use, rather than borrowing procedures from the observed-score equating framework.

1.3 Research Questions

This consideration leads to the following research questions:

(1) What is the formal nature of the linking function required to map true response model parameters from one calibration onto true response model parameters from a second calibration?

(2) How does error from estimation of the response model parameters propagate into the linking function identified in (1)? Are methods available that will reduce the propagation of error?

(3) Is it possible to identify, in advance of the second test administration, linking elements that should be common to both tests that will minimize random error in the linking function?

To lay the foundation, Chapter 2 addresses the first research question. Chapters 3 and 5 address the second research question for dichotomous and polytomous items, respectively. The final question is addressed in Chapters 4 and 6, for tests of dichotomous items and of mixed-format items, respectively. Throughout all chapters, an analytic approach is taken followed by simulation and/or empirical examples.

1.4 Overview of the Thesis

The motivation for the research and the research questions are presented in the current chapter. The following chapters follow a logical order, but have been written to be self-contained. Therefore, overlap could not be avoided.

In *Chapter 2*, formal definitions for observational equivalence and model identification are presented in accordance with past literature on model identifiability. The 3PL is used to illustrate the general lack of identifiability of the well-known dichotomous response models. The purpose of this work is not to establish sufficient identifiability criteria for the models, but to establish the need to link response model parameters when they are calibrated under different

identifiability restrictions. The chapter then presents main theorems that characterize the formal nature of linking functions for monotone, continuous response models, derive their specific shapes for different parameterizations of the 3PL model, and show how to identify them from the parameter values of the common items or persons in different linking designs. For the traditional parameterization of the 3PL, it is shown that the minimal linking elements required for identification of the linking function include a single common item, a pair of common items, or a pair of common test-takers.

The issue of estimation of the linking function parameters for dichotomous response models is explored in *Chapter 3*. The linking functions derived in *Chapter 2* are functions of the model parameters from common linking elements in each calibration. Model parameter estimation error therefore propagates into the estimated linking functions. Closed-form expressions of asymptotic standard errors of linking for the new approach are derived using a multivariate delta method for the linking elements minimally required to identify the linking function. A precision-weighted approach is then proposed, in which linking function estimates from minimal linking elements in the linking design are combined so as to reduce the relative contribution of linking elements with higher linking ASE. Results are compared to the ASE of the currently used mean/mean and mean/sigma linking methods derived by Ogasawara (2000, 2001, 2011), and empirical examples from a few recent linking studies are presented.

Because linking error is now understood at the level of the minimal linking element, optimal linking design – the selection of linking elements to minimize linking error subject to other content, psychometric, and practical constraints – is made possible. *Chapter 4* formulates the mathematical models for optimal linking design for mixed integer programming methods. Objective functions that minimize a weighted composite of the error in the slope and intercept linking parameters, minimize the error in one linking parameter while bounding error in the other linking parameter, and that maximize test information for an entire test while bounding error in the linking parameters from common items are formulated. A method is then proposed to anticipate the parameter estimation error on a second (as of yet unadministered) test form such that the optimal linking design procedure may be used to select which of the available items should be placed in the second test form to serve as linking elements *before* test administration. The models and methods are then illustrated through empirical examples. Model files and R code for these examples are available in Appendices A and B.

Chapter 5 then extends the methods in *Chapters 2 and 3* to well-known polytomous models. A formal treatment of linking of polytomous models has not been conducted to date. Previous literature relies on test-characteristic methods to compute linear transformation constants

(Baker, 1992, 1993; Muraki & Chang, 1994; Kim & Hanson, 2002; Koenig & Roberts, 2007; von Davier & von Davier, 2007), which are indirect functions of model parameters, or leave open important questions about how to combine estimates of the item difficulty parameters within an item before conducting the traditional mean/mean or mean/sigma methods, and whether the same linking function may be applied when a slope/intercept parameterization of the polytomous items is used while the $a_i(\theta_p - b_i)$ parameterization is used for dichotomous items (Kim & Hanson, 2002; Kim & Lee 2006). Neither have closed-form asymptotic standard errors for any of the previously defined linking methods been derived. *Chapter 5* presents a theorem that establishes the nature of the linking function for the nominal, generalized partial credit, sequential, and graded response models. Minimal linking elements for identification of the linking function are then presented. The closed-form asymptotic standard error of each linking element is derived, and a precision-weighted average is again proposed to combine estimates from multiple linking elements. Empirical examples are presented.

Optimal linking design of mixed-format tests (tests with both dichotomous and polytomous items) is presented in *Chapter 6*. A natural question emerged from the research conducted in *Chapter 5* – will items with more response categories be favored or discarded during optimal selection of items for linking elements? In *Chapter 6*, an analysis of the expressions for the ASE of linking parameters for polytomous items led to two hypotheses about the relative ASE of linking parameters for items with more response categories. A simulation study was conducted to test these hypotheses. Empirical examples of optimal linking design for an operational exam with mixed-format items are then presented.

The thesis concludes with a summary of main results and a discussion of their contribution to response model linking methods in *Chapter 7*. Suggestions for future research are also included.

References

Baker, F. B. (1992). Equating under the graded response model. *Applied Psychological Measurement, 16*, 87–96.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17*, 239–251.

Chang, H. & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59*, 391–404.

Cohen, A. S., & Kim, S.-H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22*, 116–130.

Doorey, N. A. (2011). *Addressing two commonly unrecognized sources of score instability in annual state assessments*. Washington, DC: Council of Chief State School Officers.

Haebara, T. (1980). Equating logistic ability scales by weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Kim, J. S., & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement, 26,* 255–270.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed-format texts. *Journal of Educational Measurement, 43*, 53–76.

Koenig, J. A., & Roberts, J. S. (2007). Linking parameters estimated with the generalized graded unfolding model: a comparison of the accuracy of characteristic curve methods. *Applied Psychological Measurement, 31*, 504–523.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices (2nd ed.).* New York: Springer.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Masters, G. M. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–173.

Masters, G. M., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York: Springer.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164) New York: Springer.

Muraki, E., & Chang, H. (1994). *Horizontal and vertical test equating methods based on the generalized partial credit model.* (ETS Internal Report). Princeton NJ: Educational Testing Service.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce), 51*, 1–23.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25,* 53–67.

Ogasawara, H. (2011). Applications of asymptotic expansion in item response theory linking. In A. A. von Davier (Ed.), *Statistical models for test equating scaling, and linking* (pp. 261–280). New York: Springer.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika, 78*, 605–623.

Chapter 2

Basic Issues in Item Response Model Parameter Linking[1]

2.1 Introduction

The literature on item-response model parameter linking tends to conceptualize the problem of linking the parameters from different calibrations as a step in the process of test score equating. For instance, for the well-known dichotomous logistic response models, Kolen and Brennan (2004, p. 156) treat the linking problem as the second step in a three-step process consisting of (i) estimating the item parameters in the response model for a new test form, (ii) scaling the parameters back to a base scale using a linear transformation, and (iii) if number-correct scoring is used, number-correct scores on the new form are converted to number-correct scores on an old form and then to scale scores. References to the problem addressed in this paper as an equating problem are found, for instance, in Dorans, Pommerich and Holland (2007), Holland and Rubin (1982), and von Davier (2011).

The main model considered in this paper is the fixed-effects three-parameter logistic (3PL) response model, which explains the probability of a correct response $U_{pi} = 1$ for a test-taker p on item i with ability $\theta_p \in \mathbb{R}$ as

$$\Pr\{U_{pi} = 1; \theta_p\} \equiv p(\theta_p; a_i, b_i, c_i) \equiv c_i + (1 - c_i)\Psi[a_i(\theta_p - b_i)], \tag{2.1}$$

[1] van der Linden, W. J., & Barrett, M. D. (in press). Linking item response model parameters. *Psychometrika.* doi: 20.1007/s11336-015-9469-6.

with

$$\Psi[a_i(\theta_p - b_i)] \equiv \frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}. \tag{2.2}$$

where $b_i \in \mathbb{R}$ and $a_i > 0$ are parameters for the difficulty and discriminating power of item i, respectively, and $c_i \in (0, 1]$ represents the height of the lower asymptote to the probability for the item. Let θ^* and θ denote the abilities of arbitrary test-takers on an old and new test form calibrated under this model. The linear transformation used in the second step above is

$$\theta^* = u\theta + v, \tag{2.3}$$

with parameters u and v to be derived from response data for the two forms.

The standard argument for the claim of linearity of the transformation in the literature relies on the notion of indeterminacy of the scale of the (estimated) θ scores (e.g., Kim, Harris & Kolen, 2010, p. 264; Lord, 1980, sect. 3.5). More precisely, it points at an arbitrary zero and unit for these scores, which manifest themselves by the fact that we can always transform θ as in (2.3), provided the two remaining parameters in (2.2) are replaced by

$$a_i^* = a_i/u \tag{2.4}$$

and

$$b_i^* = ub_i + v. \tag{2.5}$$

for all i.

Popular methods to estimate the parameters u and v are the mean/sigma method (Marco, 1977), the mean/mean method (Loyd & Hoover, 1980), and the methods based on the entire response functions for the common items in the two test forms by Haebara (1980) and Stocking and Lord (1983). The first two methods are based on a choice from the following relationships between the parameter values in the two calibrations:

$$u = \frac{\mu(a)}{\mu(a^*)},\ \mu(a^*) > 0, \tag{2.6}$$

$$= \frac{\sigma(b^*)}{\sigma(b)},\ \sigma(b) > 0, \tag{2.7}$$

$$= \frac{\sigma(\theta^*)}{\sigma(\theta)}, \sigma(\theta) > 0. \tag{2.8}$$

and

$$v = \mu(b^*) - u\mu(b) \tag{2.9}$$
$$= \mu(\theta^*) - u\mu(\theta), \tag{2.10}$$

with $\mu(\cdot)$ and $\sigma(\cdot)$ denoting means and standard deviations. These methods are applied substituting the means and/or variances of the parameter estimates for the common item and test-takers in the linking study in these expressions. The *BILOG* computer program (Mislevy & Back, 1990) included a version of these methods with the arithmetic means of the b parameters in (2.9) but the geometric instead of the arithmetic means of the a parameters in (2.6); for a generalization of this log-mean/mean procedure, see Haberman (2009). The Stocking-Lord method finds estimates of u and v minimizing the squared difference between the sums of the response functions for common items in the two test forms. For the three-parameter logistic (3PL) model in (2.1), the criterion to be minimized is

$$\left[\sum_i p(\theta; a_i^*, b_i^*, c_i^*) - \sum_i p(\theta; a_i/u, ub_i + v, c_i)\right]^2, \tag{2.11}$$

for a selection of θ values and with estimates substituted for the item parameters. For further details on these methods, see Kolen and Brennan (2004, sect. 6.2–6.3).

From a practical point of view, this treatment of the linking of response model parameters as a step in test score equating seems to make sense. IRT parameter linking does have some history in the context of IRT observed-score equating (Lord, 1980). And the fact that the necessary data for the estimation of the linear transformation in (2.3) are collected using the same type of sampling designs as used in plain observed-score equating (equivalent-groups designs; anchor-item designs; etc.) seems to lend additional support to the treatment of IRT parameter linking in this context.

From a more theoretical perspective, however, objections to this point of view are possible. First of all, a characteristic feature of all item response models is the presence of separate parameters for the effects of the properties of the items and the test-takers' abilities on the response probabilities. A naïve observer may note that the item parameters already adjust the probabilities for the differences between the items in the test forms and wonder where the necessity of the additional linking does come from.

A more fundamental puzzle is the assumed linearity of the linking transformation. The argument of the logistic function in (2.2) is definitely nonlinear (products of a_i with θ_p and b_i).

So it would be wrong to use this aspect of the parameter structure to motivate the shape of the transformation. Further, although the notion of a measurement scale for an ability with an indeterminate zero and unit has a long tradition in the behavioral and social sciences, enforced by classical publications such as Stevens (1946), its use in the current context focuses our attention exclusively on the scale of the θ parameter that is measured. But the model in (2.1)–(2.2) has four parameters for each response probability. Why should we be interested in the a_i and b_i parameters only because of features of the scale of θ? And how about the c_i parameters? If their scale is determinate, how come the Haebara and Stocking-Lord methods, which are sensitive to estimation error in these parameters, have been claimed to outperform the mean/mean and mean/sigma methods (e.g., Baker & Al-Karni, 1991)? But if their scale is indeterminate, how could we ever motivate the popularity of the last two methods in the linking literature, which ignore the c_i parameter completely?

Similar questions arise if we reparameterize the model. For example, for computational reasons (use of the Gibbs sampler), it has become convenient for Bayesian estimation to reparameterize the argument of the logistic function in (2.2) as $\alpha_i\vartheta_p + \beta_i$ (Albert, 1992). But does the scale of ϑ for this version of the model still have an indeterminate zero and unit? And is its linking transformation still linear?

The view of parameter linking in this paper is solely as a fundamental problem due to a formal feature of item response models—their general lack of identifiability. The notion of model identifiability seems akin to the one of indeterminacies of scales in Stevens' (1946) classification of measurement scales. But, unlike Stevens' classification, which just consists of a set of definitions of different levels for the scale of the parameter we try to measure and then lets us wonder how to establish the nature of the scale in a specific measurement situation, it implies a formal criterion that can be applied directly to the measurement model that is used. Loosely speaking, the criterion requires us to check if each possible distribution of the response data implies a unique set of values for *all* parameters in the model.

If a models lacks identifiability, the problem can be resolved by adjusting its numbers of equations and/or parameters, which in the current context of IRT parameter estimation with fixed numbers of item and ability parameters leads to the necessity of extra restrictions on them. The well-known practice of setting the mean and standard deviation of the ability parameters in a maximum marginal likelihood (MML) calibration of the items equal to

$$\mu_\theta = 0, \sigma_\theta = 1 \tag{2.12}$$

is an example of the use of such restrictions.

However, even if the problem of identifiability is resolved, a new problem arises. The general effect of the use of identifiability restrictions is different values for the parameters of the same items and test-takers in different calibrations. Hence, these parameters can only be compared if we know the function that maps the set of their values for one calibration onto those for the other. Once all parameters are linked, the response model automatically adjusts for any relevant differences between items or test-takers, and future estimates of any ability parameter for given item parameters (or reversely) are always directly comparable. Thus, linking functions are not necessary to correct for arbitrary units and zeroes of the θ parameters but, more generally, *to adjust for the different effects of the identifiability restrictions used in separate calibrations.*

Observe that the differential effect also arises if we use (2.12) for two separate calibrations. Using superscripts to index different groups of test-takers, we then have

$$\mu_\theta^{(1)} = 0,\ \sigma_\theta^{(1)} = 1 \tag{2.13}$$

and

$$\mu_\theta^{(2)} = 0,\ \sigma_\theta^{(2)} = 1. \tag{2.14}$$

The prevalent practice of not indexing different groups of test-takers in (2.12) may easily lead to the erroneous belief that these restrictions always have the same effect. However, as explained in more detail below, each of these two sets of restrictions yields a different intersection with the model equations and hence different identified values for all model parameters.

As just noted, linking functions are thus mathematical functions that map the set of values for the item and ability parameters in the response model for one calibration onto those for another that are necessary because of its lack of identifiability. The main theorems in this paper characterize these functions for the general class of monotone, continuous item response models, and derive their specific shapes for different parameterizations of the 3PL model with the fixed ability parameters in (2.1)–(2.2). In addition, they show how to identify the linking functions from the parameter values of common items or test-takers for different linking designs. As the current focus is only on the mathematical definition of linking functions, we treat all item and ability parameters as known and postpone the treatment of the statistical problem of estimating such functions. Before presenting the theorems, a few notions from the literature on model and parameter identifiability necessary to understand the nature of these functions are reviewed.

2.2 Observational Equivalence and Identifiability

We restrict the review of the problem of identifiability to the class of models that serve as parametric probability functions for the distribution of (discrete) random variables. Except for an occasional definition (e.g., Casella & Berger, 2002, sect. 11.2), the problem does not have much of a history in textbooks on statistics. All families of distributions typically discussed in these texts have probability functions with standard parameters that are identifiable. But the problem does have an active history of research in econometrics, mainly because of its tradition of modeling these standard parameters as functions of quantities of substantive interest, as well as in generalized latent variable modeling. Classical papers in the econometric literature discussing parameter identifiability include Koopmans (1949), Fisher (1961, 1965), Rothenberg (1971), Richmond (1974), and Gabrielsen (1978). Bekker, Merckens and Wansbeek (1994) offer an enlightening analysis of the problem of identifiability in structural equation modeling. For an introduction from the perspective of generalized latent variable modeling, see Skrondal and Rabe-Hesketh (2004, chap. 5).

The problem of identifiability does arise in IRT because of its similar attempt to explain standard parameters of response distributions as functions of item and ability parameters. Relevant papers addressing the problem for a variety of response models include Bechger, Verhelst, et al. (2001), Bechger, Verstralen, et al. (2002), Fischer (2004), Maris (2002), Maris and Bechger (2004; 2009), Reiersøl (1950), Revuelta (2009), San Martín, Gonzáles and Tuerlinckx (2009), San Martín, Jara, et al. (2011), Tsai (2000), and Volodin and Adams (2002). The problem of linking parameters estimated under different identifiability restrictions seems to be restricted mainly to item response theory, however. At least, these authors are not aware of any other field where parameters estimates are linked as frequently and routinely as in educational and psychological testing; the only exception known to them is Luijben's (1991) treatment of the equivalence of two differently restricted versions of an unidentifiable structural equations model (also addressed by Bekker et al., 1994, chap. 7). For the treatment of parameter linking for a response-time model with item and person parameters from the perspective of model identifiability, see van der Linden (2010).

The following definitions, which can be found throughout the literature just referred to, are for a model of a random vector with a multidimensional parameter space:

Definition 1. Two points in a parameter space are observationally equivalent if they imply the same joint distribution of the random vector.

Definition 2. A parameter is identifiable if for any of its points there is no other point that is observationally equivalent.

More formally, let $\mathbf{x}$ denote the random vector that is considered and $f(\mathbf{x};\boldsymbol{\pi})$ its probability function, which is assumed to have vector-valued parameter $\boldsymbol{\pi}$. Then, $\boldsymbol{\pi}$ is identifiable if for any pair $\boldsymbol{\pi}_0 \neq \boldsymbol{\pi}_1$, it holds that $f(\mathbf{x};\boldsymbol{\pi}_0) \neq f(\mathbf{x};\boldsymbol{\pi}_1)$ for all $\mathbf{x}$. Observe that lack of identifiability may be due to some of the components of $\boldsymbol{\pi}$ only. As $\mathbf{x}$ typically represents observed data, it is common to refer to a parameter as being "identifiable from the data." The use of this phrase emphasizes the practical meaning of parameter identifiability: If different values of $\boldsymbol{\pi}$ imply the same probability distribution, it becomes impossible to use observed data to distinguish between them, let alone infer a "true" value of $\boldsymbol{\pi}$. Consequently, if $\boldsymbol{\pi}$ is not identifiable, then for some of its values, the likelihood function $f(\boldsymbol{\pi};\boldsymbol{x})$ associated with the observations does not allow us to discriminate between them. Indeed, if a parameter lacks identifiability, it does not have a consistent estimator (Gabrielsen, 1978).

Definition 2 immediately suggests possible refinements of the criterion of identifiability, such as local identifiability of $\boldsymbol{\pi}$ at $\boldsymbol{\pi}_0$ (i.e., $\boldsymbol{\pi}$ is identifiable in a neighborhood of $\boldsymbol{\pi}_0$) or identifiability from a restricted set of values of x. But more important to our current goals is a discussion of a slightly generalized version of a theorem in Bartels (1985):

Theorem 1. If $\boldsymbol{\pi}$ and $\boldsymbol{\varphi}$ are the parameters in alternate versions of a model of a given random variable that have a bijective relationship**,** then $\boldsymbol{\pi}$ is identifiable if and only if $\boldsymbol{\varphi}$ is.

The theorem is immediately obvious if we realize that, in the current context, parameters serve as quantities that index individual members of families of probability distributions. As the relation between $\boldsymbol{\pi}$ and $\boldsymbol{\varphi}$ is bijective (one-to-one and onto), their role as index is entirely exchangeable.

The theorem explains why we can reparameterize a model (replace its structure with one set of parameters by a structure with another set) without losing its identifiability, provided the two sets of parameters have a bijective relationship. An example is the slope-intercept parameterization $\alpha_i\vartheta_p + \beta_i$ of the logistic model referred to earlier. In order to give the response function this parameter structure, we have to substitute

$$\begin{aligned} \theta_p &= \theta(\vartheta_p) = \vartheta_p; \\ a_i &= a(\alpha_i) = \alpha_i; \\ b_i &= b(\alpha_i, \beta_i) = -\beta_i/\alpha_i, \end{aligned} \tag{2.15}$$

$\vartheta_p, \beta_i \in \mathbb{R}$ and $\alpha_i > 0$, into (2.2). The relation between the two alternative sets of parameters is invertible, and thus bijective.

A special instance of the function $\boldsymbol{\varphi} = \varphi(\boldsymbol{\pi})$ in Theorem 1 are vector functions

$$\boldsymbol{\varphi} = (\varphi_1(\pi_1), ..., \varphi_d(\pi_d)), \tag{2.16}$$

with $\varphi_1, ..., \varphi_d$ being scalar-valued, bijective functions of the d different components of $\boldsymbol{\pi}$**.** An example of this *componentwise* type of bijective function is the well-known reparameterization of the Rasch (1960) model,

$$p(\vartheta_p; \beta_i) \equiv \frac{\vartheta_p}{\vartheta_p + \beta_i}, \tag{2.17}$$

$\vartheta_p, \beta_i > 0$, which, maintaining our current notation, follows from (2.2) with $a_i = 1$ upon substitution of

$$\begin{aligned} \theta_p &= \theta(\vartheta_p) = \ln \vartheta_p; \\ b_i &= b(\beta_i) = \ln \beta_i, \ \vartheta_p, \beta_i > 0. \end{aligned} \tag{2.18}$$

The critical difference between the two types of reparameterization resides thus in the fact that *each* component in (2.18) is a bijective function of its counterpart as well, whereas the component for b_i in (2.15) is not.

Reparameterization is a useful tool if we have to prove identifiability of a model. It allows us to replace the set of model equations with the current parameters by an equivalent set for which the proof is simpler. We will use this trick to prove some of our later results. Also, the problem of parameter linking will appear to be one in which we have to derive and estimate functions as in (2.16).

As already noted, the typical solution to an identifiability problem for an item response model is the introduction of extra restrictions on its parameters. Their effect is a reduction of the parameter space to one that uniquely represents each possible member of the family of response distributions posited by the model. However, such restrictions generally have a differential impact on the parameter space, lead to different unique representations, and consequently leave us with a linking problem. Because our focus is mainly on this problem, we only highlight the nature of the identifiability problem for the 3PL model in (2.1)–(2.2), presenting a few cases in which different sets of parameters in the 3PL model in (2.1)–(2.2) clearly show lack of identifiability (including the commonly believed to be invariant c_i parameters). It is not our intention to provide a solution for it. In fact, as follows from Theorem 3 below, in order

to derive a linking function for monotone continuous response model, it is not necessary to know the identifiability restrictions actually used in the different calibrations at all; only their differential impact on the item and ability parameters values counts.

2.3 3PL Model

The distributions addressed by the 3PL model in (2.1) are for the dichotomous responses $U_{pi} = 0, 1$ by test-takers $p = 1, ..., P$ on items $i = 1, ..., I$. The distributions are Bernoulli with probability functions

$$f(u_{pi}; \pi_{pi}) = \pi_{pi}^{u_{pi}}(1 - \pi_{pi})^{1-u_{pi}},\ p = 1, ..., P;\ i = 1, ..., I, \tag{2.19}$$

which have success parameters $\pi_{pi} \in [0, 1]$ representing the probability of a correct response by each test-taker on each item.

For different values of its parameter π_{pi}, each of the probability functions in (2.19) yields a different distribution; hence these parameters are identifiable. Observe that if these probability functions are reparameterized by substituting $\pi_{pi} = 1 - \eta_{pi}$, Theorem 1 guarantees that η_{pi} is also identifiable. We will use this feature frequently.

Making the usual assumption of independence within and between test-takers, the probability function of the joint distribution of a complete response matrix, $\mathbf{U} = (U_{pi})$, is the product of $P \times I$ of these Bernoulli distributions,

$$f(\mathbf{u}; \boldsymbol{\pi}) = \prod_p \prod_i \pi_{pi}^{u_{pi}}(1 - \pi_{pi})^{1-u_{pi}} \tag{2.20}$$

with parameter vector $\boldsymbol{\pi} = (\pi_{11}, ..., \pi_{1I}, ..., \pi_{P1}, ..., \pi_{PI})$. Clearly, as each of its components is identifiable, so is $\boldsymbol{\pi}$.

The 3PL model specifies each π_{pi} as a function of the parameters $(\theta_p, a_i, b_i, c_i)$ for the effects of the test-taker's ability and the properties of the item on it. Rather than a direct probability function for a response distribution, the model is thus a (second-level) mathematical model in the form of a system of $P \times I$ nonlinear equations

$$\pi_{pi} = c_i + (1 - c_i)\Psi[a_i(\theta_p - b_i)],\ p = 1, ..., P;\ i = 1, ..., I, \tag{2.21}$$

one for each of the success parameters.

2.3.1 Lack of Identifiability

The following three cases illustrate the lack of identifiability of the 3PL model:

Theorem 2. The system of equations for the 3PL model in (2.21) is not identifiable in the following cases: (i) c_i known for all i; (ii) $a_i = a \in \mathbb{R}^+$, for all i; and (iii) $\theta_p = \theta \in \mathbb{R}$ for all p.

Proof. (i) This case basically amounts to the 2PL model in (2.2). The fact that different values of a_i, b_i, and θ_p parameters do not need to imply different values for π_{pi} follows from (2.3)–(2.5), and is well known. (ii) Lack of identifiability of the b_i, c_i, and θ_p parameters for this case was established by Maris (2002) and later re-analyzed in Maris and Bechger (2009). Without loss of generality, his proof sets $a = 1$ and reformulates (2.1)–(2.2) as

$$\pi_{pi} = \frac{\exp(\theta_p) + c_i \exp(b_i)}{\exp(\theta_p) + \exp(b_i)}, \tag{2.22}$$

which, upon substitution of $\theta_p = \ln \vartheta_p$, $b_i = \ln \beta_i$, and $c_i = \ln \delta_i/\beta_i$, leads to

$$\pi_{pi} = \frac{\vartheta_p + \delta_i}{\vartheta_p + \beta_i}. \tag{2.23}$$

Adding a constant to ϑ_p for all p and subtracting the same constant from β_i and δ_i for all i gives different sets of values for these parameters with the same π_{pi}. (iii) Let $\pi_{pi} = 1 - \eta_{pi}$ and $c_i = 1 - \gamma_i$. Theorem 1 implies that π_{pi} and c_i are identifiable if and only if η_{pi} and γ_i are. From (2.1)–(2.2), using $\theta_p = \theta$ for all p,

$$\begin{aligned} \eta_i &= \frac{\gamma_i}{[1 + \exp(a_i(\theta - b_i))]} \\ &= \frac{\gamma_i^*}{\kappa[1 + \exp(a_i(\theta - b_i))]}, \end{aligned} \tag{2.24}$$

for all i and any $\kappa \in (0, \gamma_i]$, where $\gamma_i^* = \kappa\gamma_i$. In order to have b_i absorb κ, we need to substitute b_i^* for b_i, where b_i^* is the solution of

$$1 + \exp(a_i(\theta - b_i^*)) = \kappa[1 + \exp(a_i(\theta - b_i))], \tag{2.25}$$

which is

$$b_i^* = \theta - \ln[\kappa[1 + \exp(a_i(\theta - b_i))] - 1]/a_i. \tag{2.26}$$

For the relation in (2.26) to hold, it is necessary that $\kappa[1 + \exp(a_i(\theta - b_i))] - 1 > 0$; or,

$$\begin{aligned} \kappa &> \frac{1}{1 + \exp(a_i(\theta - b_i))} \\ &= 1 - \Psi_i, \end{aligned} \tag{2.27}$$

with Ψ the logistic function in (2.2). Alternatively, the change in the γ_i parameters can be traded off by

$$a_i^* = \ln[\kappa[1 + \exp(a_i(\theta - b_i))] - 1]/(\theta - b_i) > 0, \tag{2.28}$$

or

$$\theta^* = b_i + \ln[\kappa[1 + \exp(a_i(\theta - b_i))] - 1]/a_i \tag{2.29}$$

where the nonnegativity requirement for a_i^* implies both $\kappa > 2(1 - \Psi)$ and $\theta > b_i$. □

The status of the c_i parameters in the version of the 3PL model with all parameters free is still unknown. But the last two cases are already enough to illustrate the problematic role of these parameters, which has largely been ignored in the literature. Lord (1980, p. 36, 184–185) even claims that the c_i parameters are actually identifiable. As already noted, an exception was Maris (2002), who introduced the second case above.

For the third case, Figure 2.1 illustrates how dramatic the trade-off between the γ_i and a_i, b_i and θ parameters can be. It displays each of the trade-offs as a function of κ for an item with $a_i = 1.0$, $b_i = -.5$ and common ability parameter $\theta = 0$. For these parameter values, $1 - \Psi = .378$; hence, the range of admissible values for κ is $(.378, 1]$ for the functions in (2.26) and (2.29) but $(.755, 1]$ for the one in (2.28). Observe that $\kappa = 1$ means no change in the γ_i parameter; for smaller values of κ, γ_i decreases in size (and c_i thus becomes larger). The change in the b_i and θ parameters as a function of the decrease in γ_i is remarkable, especially closer to their vertical asymptote at $\kappa = .378$. On the other hand, the a_i parameter appears to be quite robust across its range of admissible values for κ. The fact that the trade-off between the c_i and the a_i parameters seems much less dramatic than for the other two parameters might go against our intuition, but is entirely due to the non-negativity requirement for the a_i parameter. If we did admit negative values, this parameter could go down, for example, all the way to $a_i^* = -10$ for $\kappa = .38$ (just above the vertical asymptote).

The trade-offs in (2.3)–(2.5) imply lack of identifiability of all parameters in the 2PL model and 1PL/Rasch model. Wood (1978) analyzed a special version of the Rasch model for equivalent items with success probability

$$\pi_{pi} \equiv \frac{\exp(\theta_p)}{1 + \exp(\theta_p)}, \tag{2.30}$$

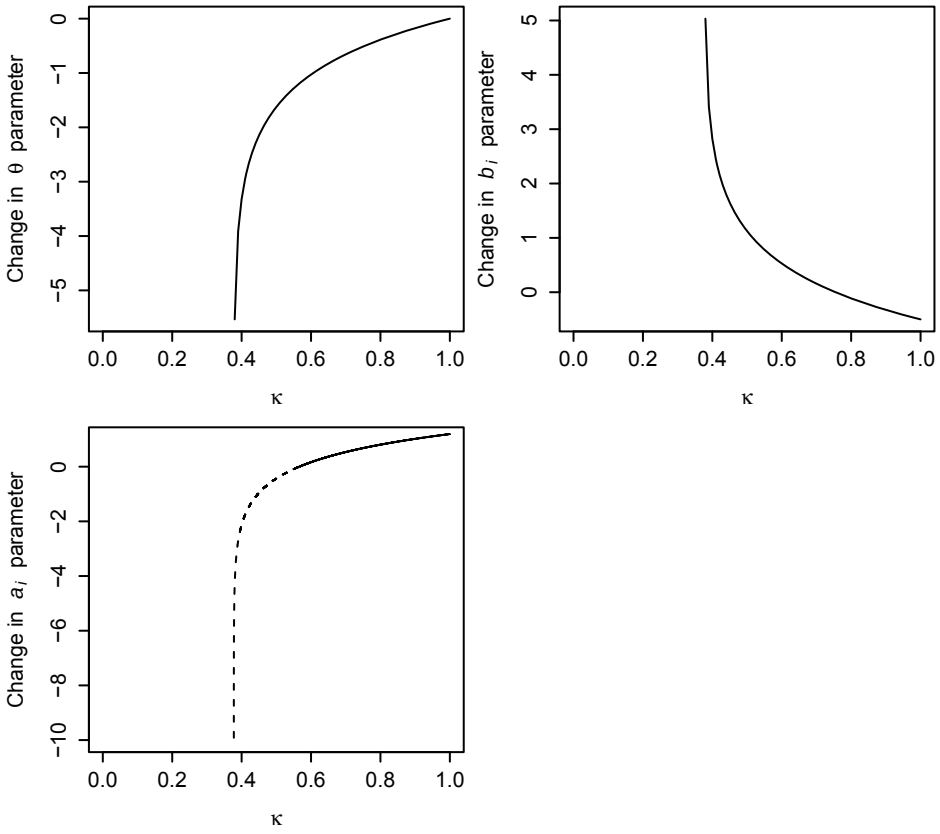

Figure 2.1. Change in θ, b_i, and a_i parameters compensating the change in $\gamma_i = 1 - c_i$ by a factor κ, for an item with $a_i = 1$ and $b_i = -0.5$ and the ability parameter fixed at $\theta = 0$. Note: Dashed line represents negative values for the a_i parameter.

for all i. This "0PL model" is just a reparameterization of the Bernoulli probability function in (2.19). It is thus fully identifiable (although unlikely to show any satisfactory fit to the items in a real-world testing program).

The literature offers only a few examples of sets of identifiability restrictions for special cases of the 3PL model in (2.1)–(2.2) for which sufficiency has formally been proven. For instance, for the 1PL/Rasch model, as is well known, it is sufficient to set the difficulty parameter of one item or the ability parameter of one test-taker equal to a known constant. Equivalently, we could impose a linear constraint on a subset of these parameters (e.g., constrain their mean to a known value). As shown by San Martín et al. (2009) for the 1PL-G model (i.e., 1PL/Rasch model extended with a guessing parameter for each item) it is sufficient to fix the difficulty and guessing parameters of one item to known constants. Recently, the same authors have shown that, contrary to what one might have expected intuitively, it is not sufficient to fix all three parameters of an arbitrary item to known constants to make the (fixed-effects) 3PL model in

(2.21) identifiable (San Martín, Gonzáles, & Tuerlinckx, 2015). In fact, we are not aware of any existing proof of a set of identifiability restrictions sufficient for it. In the absence of such proofs (but the presence of large amounts of response data), the practical solution in the testing industry has been to circumvent the problem using a two-stage calibration procedure. In its first stage, the ability parameters are temporarily treated as a random sample from a population distribution, which allows for marginalization of the likelihood for the fixed-effects version of the model with respect to an assumed ability distribution and maximum marginal likelihood (MML) estimation of all item parameters. The second stage consists of subsequent maximum likelihood or Bayesian (e.g., EAP) estimation of the individual ability parameters assuming the item parameters have been estimated from enough response data to treat them as known. Typically, the ability distribution in the first stage is taken to be the standard normal, which implies the adoption of (2.12) as *de facto* standard set of identifiability restrictions for the 3PL model in the field of educational testing. Although its effectiveness has been confirmed in the daily practice of item calibration (e.g., convergence of parameter estimates, which would have been problematic with lack of identifiability) as well as through numerous parameter recovery studies, we still await a formal proof of its sufficiency. As for the second stage, if the item parameters can be treated as known, (2.21) defines known monotonic relationships between each of the ability parameters θ_p and their success parameters π_{pi}, and the former are therefore identified as well.

For all practical purposes, the use of (2.12) in this two-stage detour thus restricts the parameters in the specification of the 3PL model in (2.21) to fixed values (to which the MML estimators of the item parameters and the subsequent estimates of the ability parameters converge with the sample size and test length, respectively). The reverse does not necessarily hold, though; identifiability of a fixed-effect specification does not automatically imply the same for its random-effect specification (San Martín, Rolin, et al., 2013).

2.4 Linking Functions

The main goal of this section is to define the problem of linking IRT parameters from different calibration studies and derive the specific linking functions necessary for the 3PL model. More specifically, we address the problem of post hoc linking; that is, mapping the parameters from one study onto the values they would have had if they had been included in another *after* both calibrations have been conducted. This type of linking is common in the testing industry. As both the item and ability parameters are to be linked simultaneously, the linking automatically

is for the fixed-effects specification of the 3PL model. In principle, it is possible to avoid the problem by concurrent recalibration of the response data collected in different studies, capitalizing on the presence of common items or test-takers in them or imposing constraints on the parameters. For this approach, it has even been proposed to tentatively impose constraints, e.g., linear constraints as in (2.3)-(2.5) and check on their appropriateness using Lagrange multiplier tests (von Davier & von Davier, 2011). But such strategies are not always practical for testing programs that have to link their parameters continuously across multiple test administrations.

Our current treatment of the linking problem only deals with its mathematical aspects at the level of the model parameters, without any bothering about the fact that these parameters are unknown. In order to actually use them in practice, linking functions need to be estimated from response data. But we are only able to find defendable estimators and evaluate their statistical quality once we have an explicit definition of their estimand.

We begin with considering the more general case of a parametric response model used to calibrate the responses for a set of P test-takers on I items with the vector of success probabilities $\boldsymbol{\pi} = (\pi_{pi})$ in (2.20). Let $f(\cdot)$ be the response function specified by the model, $\boldsymbol{\xi}_{pi}$ its vector of parameters for the combination of test-taker p and item i, and $\boldsymbol{\xi} = (\boldsymbol{\xi}_{pi})$ the vector of parameters for all test-takers and items. The choice of model amounts to the adoption of a system of $P \times I$ equations $\pi_{pi} = f(\boldsymbol{\xi}_{pi})$. As the probabilities π_{pi} are identified and thus have fixed values for all combinations of p and i, each of the equations introduces a level surface (contour) in the domain of f, which is the subset of all values of $\boldsymbol{\xi}$ for which $f(\boldsymbol{\xi}_{pi}) = \pi_{pi}$ is true. The solution set for the system of equations is the intersection of all $P \times I$ surfaces. As the system lacks identifiability of the model parameters, the set consists of more than one point. Identifiability restrictions are extra equations added to the system. The intersection of their solution sets with the set for the system reduces the latter to a unique point, whose coordinates are the true values of the item and test-taker parameters for the calibration.

Now, suppose we have conducted two separate calibration studies that had both unique and common test-takers and/or items in them. Both studies are assumed to have used appropriate sets of identifiability restrictions. Obviously, the use of different restrictions implies different intersections of their solution sets with those determined by the two systems of model equations, and hence different true values for the common parameters in the two calibrations. But different true values can also arise if formally identical sets of identifiability restrictions have been imposed on the two calibrations. The presence of unique test-takers and/or items in the calibrations implies different vectors of success probabilities $\boldsymbol{\pi}^* \neq \boldsymbol{\pi}$ for them and thus different solutions sets for their model equations. Consequently, their intersections with the solution set

for the identifiability restrictions generally differ, and the same items or test-takers assigned to the two calibrations can therefore have different true parameter values. The critical factor is the scope of the restrictions. For example, if they fix the values of some of the common parameters to the same known constants in the two calibrations, obviously their impact on them is identical. But if they include unique parameters and leave the common parameters free, they yield different true values for the latter—an observation confirmed by the educational testing industry, where invariably different parameter values are found for common parameters in separate calibrations with large samples of test-takers for the restrictions in (2.13) and (2.14), as well as by our example later in this paper. In fact, if this differential effect did not exist, we would not have to link any parameters.

Consider a hypothetical combination of a test-taker and item assigned to two of these calibration studies with identified parameters. Let $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}$ denote the vectors with the unique true values for the combination in the two studies (where the indices have been omitted for notational convenience, as well as to emphasize the hypothetical nature of the combination). For example, for the 3PL model, $\boldsymbol{\xi}^* = (\theta^*, a^*, b^*, c^*)$ and $\boldsymbol{\xi} = (\theta, a, b, c)$. The question of how to map $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$ onto one another is the topic of this section. Observe that, although different, both $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}$ are associated with the *same* success probability π for the combination of test-taker and item. This fact is key in our derivation of the mapping below.

Our first theorem is for a general response model that specifies success probability π for each combination of a test-taker and item only as a monotone continuous function of their parameters, where the monotonicity is taken to mean that π is strictly increasing or decreasing in each of the components of $\boldsymbol{\xi}$ with all other components fixed at any of their admissible values. We then present our results for the 3PL model. The version of this model for the regular parameterization in (2.1)–(2.2) is both continuous and monotone in each of its parameters, provided we exclude the case of $\theta = b$ for the a parameter; the version with the slope-intercept parameterization is monotone in each of its parameters without any further restriction.

Again, except for an illustrative example below, the current paper only deals with the mathematical aspects of linking functions; the problem of how to actually estimate them and evaluate their estimation error deserves separate treatment.

Theorem 3. Assume a response model with a fixed parameter structure that (i) specifies π as a monotone continuous function of its parameters and (ii) has been used in two separate calibration studies with identified parameters $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$. Then $\boldsymbol{\xi}^*$ is linked to $\boldsymbol{\xi}$ by a vector

function

$$\boldsymbol{\xi}^* = \varphi(\boldsymbol{\xi}) = (\varphi_1(\xi_1), ...\varphi_d(\xi_d)) \tag{2.31}$$

with components $\varphi_1, ..., \varphi_d$ that are both monotone and continuous.

Proof. Let ξ^* and ξ be an arbitrary pair of corresponding components of $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}$**.** Fixing all other components, the monotonicity of the model implies the existence of monotone functions $\pi = f(\xi^*)$ and $\pi = g(\xi)$. Hence, there also exists a function $\xi^* = f^{-1}(g(\xi)) = \varphi(\xi)$. Being a composite of continuous functions, φ is continuous. Further, as both f and g are bijective, φ is bijective as well. Suppose that φ is not monotone. It then has an interior point ξ_0 in its domain with a local optimum. But this implies the existence of points $\xi' < \xi_0 < \xi''$ with $\varphi(\xi') = \varphi(\xi'')$, which contradicts the fact that φ is bijective. Thus, φ is monotone. □

The feature of monotonicity should not come as a surprise. If it did not hold, the two sets of the identifiability restrictions would imply a different order of some of the parameters, for instance, a reversal of the difficulties of two items, which is impossible without violating the requirement of observational equivalence. Observe, however, that $\varphi(\boldsymbol{\xi})$ is only required to be *componentwise* monotone; it does not need to hold that all components be increasing or all of them be decreasing.

It is thus possible to view the impact of the use of different sets of identifiability restrictions as a componentwise reparameterization of the response model, which leaves the structure of the model intact. However, unlike the earlier case of a known function in (2.16) applied to unknown parameter values, we now have to address the reverse problem: This time the two sets of parameter values are given, and we have to find the componentwise bijective function that maps them onto one another. Observe again that the specific identifiability restrictions used in the calibrations need not be known at all; neither do we need to assume anything about the statistical estimation method through which the restrictions might have been imposed. Only their impact on the item and test-taker parameters counts.

For the 3PL model, the linking function $\boldsymbol{\varphi} = (\varphi_\theta(\theta), \varphi_a(a), \varphi_b(b), \varphi_c(c))$ has to be derived from (2.1)–(2.2) for two arbitrary sets of values $(\theta^*, a^*, b^*, c^*)$ and (θ, a, b, c). However, it is simpler to use the first equation in (2.24), and find φ_θ, φ_a, φ_b, and φ_γ as the solution of

$$\frac{\varphi_\gamma(\gamma)}{1 + \exp[\varphi_a(a)(\varphi_\theta(\theta) - \varphi_b(b))]} = \frac{\gamma}{1 + \exp[a(\theta - b)]}, \tag{2.32}$$

with additional back transformation of φ_γ to φ_c. The required linking function is thus the solution of a functional equation in four unknowns (for relevant theory of functional equations,

see, for instance, Sahoo & Kannappan, 2011, or Small, 2007). The next theorem shows the solution:

Theorem 4. Given the conditions in Theorem 3, the linking function for the 3PL model is

$$\varphi_a(a) = u^{-1}a, \tag{2.33}$$

$$\varphi_b(b) = ub + v, \tag{2.34}$$

$$\varphi_c(c) = c, \tag{2.35}$$

and

$$\varphi_\theta(\theta) = u\theta + v, \tag{2.36}$$

with

$$u \equiv \frac{\varphi_\theta(\theta) - \varphi_b(b)}{\theta - b},\ \theta \neq b, \tag{2.37}$$

and

$$v = \varphi_b(b) - ub = \varphi_\theta(\theta) - u\theta. \tag{2.38}$$

Proof. From (2.32),

$$\varphi_\gamma(\gamma) = \frac{1 + \exp[\varphi_a(a)(\varphi_\theta(\theta) - \varphi_b(b))]}{1 + \exp[a(\theta - b)]}\gamma. \tag{2.39}$$

As this function is monotone in γ, φ_γ is a monotone component of $\boldsymbol{\varphi}$, and therefore

$$\frac{1 + \exp[\varphi_a(a)(\varphi_\theta(\theta) - \varphi_b(b))]}{1 + \exp[a(\theta - b)]} = \kappa > 0, \tag{2.40}$$

is a constant independent of γ. Thus, $\varphi_\gamma(\gamma) = \kappa\gamma$. However, since φ_γ is a monotone mapping from $[0, 1]$ onto itself, $\kappa = 1$ and (2.35) follows. We now have to find φ_θ, φ_a, and φ_b as the solution of (2.40) for $\kappa = 1$; that is,

$$\varphi_a(a)[\varphi_\theta(\theta) - \varphi_b(b)] = a(\theta - b). \tag{2.41}$$

Rewriting the equation,

$$\varphi_a(a) = \frac{\theta - b}{\varphi_\theta(\theta) - \varphi_b(b)}a, \tag{2.42}$$

with $\varphi_\theta(\theta) \neq \varphi_b(b)$. But, as φ_a is a monotone component of $\boldsymbol{\varphi}$ as well ,

$$\frac{\varphi_\theta(\theta) - \varphi_b(b)}{\theta - b} = \text{const}, \tag{2.43}$$

which is our key equation. First, (2.33) follows directly from (2.43) along with the definition of its constant in (2.37). Further, (2.43) shows that $\varphi_\theta(x) - \varphi_b(x)$ is equal to a constant times $\theta - b$. Substituting $x = \theta = b$ yields $\varphi_\theta(x) - \varphi_b(x) = 0$, and it thus holds that $\varphi_\theta = \varphi_b = \varphi$. Observe also that (2.43) implies a constant difference quotient for φ. Hence, φ is linear, and (2.34) and (2.36) hold. Finally, (2.38) follows from (2.34)–(2.36). □

Although the proof did not make any assumptions as to the general shape of the functions that map the values of the a_i, b_i,and θ_p parameters in the new calibration onto those in an earlier calibration, they appear to be linear, just as currently assumed in the literature; see our review in (2.3)-(2.5). However, a new result is the definition of linking parameter u, and consequently of v. Unlike (2.6)–(2.8), (2.37) defines it as the ratio of the differences between the test taker's ability and the difficulty of the item in the two calibrations. The reason for the difference between these new and old definitions may be the failure in the current literature to distinguish between the formal definitions of u and v and their solutions from the system of linking equations implied by the choice of linking design. As demonstrated in the next section, separating the two does give us large flexibility to derive alternative solutions for u and v from alternative designs. Another new result is the derivation of the identity function for the c_i parameters. We will further reflect on its practical implications in the last section of this paper.

In addition, it is important to note the different nature of these functions for the four different types of parameters. The one for the c_i parameters is an identity function, which does not involve either of the linking parameters u and v. On the other hand, the function for the a_i parameters involves one linking parameter, u, whereas those for the b_i and θ_p parameters depend both on u and v. Thus, unlike the c_i parameters, the latter can be linked only when the numerical values of the linking parameters are known. This obvious point takes us to another identifiability requirement, namely for the system of linking equations to be derived from (2.33)–(2.38) for the specific design adopted in the linking study.

2.5 Identification of Linking Parameters

Basically, a linking design is a combination of two calibration designs with common items and/or test-takers. Once it has been selected, (2.33)–(2.38) can be used to derive a system of equations of the unknown linking parameters u and v in the parameter values for the common items or test-takers in the two calibrations. Of course, u and v have unique values only when the system is identified. The problem of linking item response model parameters thus involves

three different types of identifiability requirements, two for the systems of the model equations in (2.21) associated with the two calibrations and another for the system of linking equations that is used. At this stage, the former have already been met through the adoption of extra restrictions in the two calibrations. The latter, although sometimes critical (e.g., Theorem 7 below), involves only an appropriate choice of linking design; no additional identifiability restrictions are necessary.

We illustrate the process for three minimal designs. In doing so, $p = 1, ..., n$ and $i = 1, ..., m$ are now used as indices for the common test-takers and items in the design, respectively, while $t = 1, 2$ will be used to denote the two calibrations. Thus, (a_{i_t}, b_{i_t}) and θ_{p_t} are the true values of the pertinent parameters for item i and the parameter for test-taker p in the tth calibration, respectively. Because the linking function for the c_i parameters is already known, these parameters can further be ignored.

2.5.1 One Common Item

Linking parameters u and v are already identified if the two calibrations have one common item, $i = 1$. The system of linking equations then follows from (2.33) and (2.38) as

$$u = \frac{a_{1_1}}{a_{1_2}},\ a_{1_2} > 0, \tag{2.44}$$

$$v = b_{1_2} - ub_{1_1}. \tag{2.45}$$

2.5.2 Two Common Items

For a pair of common items $i = 1, 2$, we could use (2.44)–(2.45) for either of them. But now an alternative is available in the form of the substitution of their two sets of parameter values (b_{1_1}, b_{2_1}) and (b_{1_2}, b_{2_2}), $b_{1_1} \neq b_{2_1}$, into (2.38). Elimination of v then gives

$$u = \frac{b_{1_2} - b_{2_2}}{b_{1_1} - b_{2_1}}, \tag{2.46}$$

whereupon v is equal to

$$v = b_{i_2} - ub_{i_1},\ \ i = 1, 2. \tag{2.47}$$

This simple system of equations gives unique values for the linking parameters once their item parameters are identified. A similar property does not hold for the next type of design.

2.5.3 One Common Test-Taker

It appears to be impossible to derive a system of equations from (2.33) and (2.38) for a common test-taker with $(\theta_{1_1}, \theta_{1_2})$ from which u and v are identifiable.

2.5.4 Two Common Test-Takers

For a pair of test-takers with $\theta_{1_1} \neq \theta_{2_1}$, u and v can be obtained as

$$u = \frac{\theta_{1_2} - \theta_{2_2}}{\theta_{1_1} - \theta_{2_1}}, \tag{2.48}$$

and

$$v = \theta_{p_2} - u\theta_{p_1}, \quad p = 1, 2. \tag{2.49}$$

Because of their practical importance, we document these results as a theorem:

Theorem 5. For the 3PL model with standard parameterization, linking parameters u and v are already identifiable for a common-item design with at least one common item and a common-test-taker design with at least two common test-takers.

As a typical linking study has more than these minimal numbers of items or test-takers, we easily have multiple systems of linking equations, each returning the same unique values for the linking parameters u and v. At the current level of true parameter values, any choice from them would thus suffice. However, in practice, except when some of item or ability parameters were fixed at known constants, in which case we can just substitute these constants into (2.44)–(2.49), all parameters are estimated. A first suggestion of how to combine estimates of u and v from multiple systems of linking equations as in (2.44)–(2.49) is offered in the empirical example below.

2.5.5 Slope-Intercept Parameterization

Earlier we wondered what the impact of a change of parameter structure on the linking function would be. Theorem 6 highlights the impact for the slope-intercept parameterization for the 3PL model.

Theorem 6. Given the conditions in Theorem 3, the linking function for the slope-intercept parameterization $\alpha\vartheta + \beta$ of the 3PL model is

$$\varphi_\alpha(\alpha) = (\alpha - u)/v,\ v > 0, \tag{2.50}$$

$$\varphi_\beta(\beta) = \beta + u, \tag{2.51}$$

$$\varphi_c(c) = c \tag{2.52}$$

and

$$\varphi_\vartheta(\vartheta) = (\vartheta - u)/w,\ w \neq 0, \tag{2.53}$$

with

$$u = \varphi_\beta(0), v = \varphi_\vartheta(1), \text{ and } w = \varphi_\alpha(1). \tag{2.54}$$

Proof. The component for the linking of c does not change because these parameters are left untouched by the reparameterization. But we now have to find φ_ϑ, φ_α, and φ_β as the solution of

$$\varphi_\alpha(\alpha)\varphi_\vartheta(\vartheta) + \varphi_\beta(\beta) = \alpha\vartheta + \beta, \tag{2.55}$$

or, equivalently,

$$\varphi_\beta(\beta) = \beta + \alpha\vartheta - \varphi_\alpha(\alpha)\varphi_\vartheta(\vartheta). \tag{2.56}$$

Substituting $\beta = 0$ and using the definition of u in (2.54) yields

$$u = \alpha\vartheta - \varphi_\alpha(\alpha)\varphi_\vartheta(\vartheta). \tag{2.57}$$

Hence, (2.51) follows from (2.56)–(2.57). Likewise, substituting $\vartheta = 1$ into (2.57), we obtain (2.50) with v given by (2.54), while substitution of $\alpha = 1$ leads to (2.53). □

Although still linear, the linking functions in (2.50)–(2.54) differ considerably from those for the regular parameterization of the 3PL model; in fact, they now appear to have three rather than two unknown parameters. More importantly, an attempt to derive an identified system of linking equations from them appears to run into practical problems. Still assuming all model parameter to be known, (2.54) suggests selecting a common item with $\beta_{i_1} = 0$ and $\alpha_{i_1} = 1$ and common test-taker with $\vartheta_{p_1} = 1$ for the first calibration and equating the three linking parameters to their values in the second calibration, that is, setting $u = \beta_{i_2}$, $w = \alpha_{i_2}$, and $v = \vartheta_{p_2}$. However, the presence of items and test-takers with such exact parameter values in a linking study is highly unlikely.

An alternative seems to solve (2.50)–(2.53) for u, v, and w, obtaining them as

$$u = \beta_{i_2} - \beta_{i_1} \tag{2.58}$$

$$v = \frac{\alpha_{i_1} - u}{\alpha_{i_2}},\ \alpha_{i_2} \neq 0, \tag{2.59}$$

and

$$w = \frac{\vartheta_{p_1} - u}{\vartheta_{p_2}},\ \vartheta_{p_2} \neq 0. \tag{2.60}$$

for an arbitrary p and i. However, we now need a linking design with both at least one common item for (2.58)–(2.59) and one common test-taker for (2.60), a new requirement entirely due to the change in parameter structure created by (2.15). Hence the following theorem:

Theorem 7. For the 3PL model with slope-intercept parameterization, linking parameters u, v, and w are only identifiable for designs with both common test-takers and common items.

The result in this theorem has a major practical implication. If one believes that IRT models always have a scale for the person parameter with an arbitrary unit and zero, as the literature referenced in our introductory section appears to do, it may seem natural to adopt the same linking functions as for the model with the standard parameterization in (2.33)–(2.38). This choice is incorrect; the proper functions are those in (2.58)–(2.60). However, we do not expect the latter to be practised regularly as they require a linking design with test-takers responding twice but independently to the same items—an assumption unlikely ever to be met because of memory effects. An alternative would be to transform the estimated slope and intercept parameters back to the standard parameters and link the latter, but then we miss the covariance matrices for the estimators of the model parameters necessary to evaluate the standard errors of linking as in the example in the next section.

2.6 Illustrative Example

Although the focus of this paper was not yet on a statistical treatment of the linking problem, a small example might already illustrate some of the practical consequences of our theoretical results. The example is for a common-item design for the 3PL model in its standard parameterization. The design allows us to use (2.44)–(2.45) to estimate linking parameters u and v in (2.33)–(2.38) for the two calibrations.

Table 2.1. *Generating and estimated parameter values for the common items*

Common	Generating Values			Calibration 1			Calibration 2		
Item	a_i	b_i	c_i	$\widehat{a}_i$	$\widehat{b}_i$	$\widehat{c}_i$	$\widehat{a}_i$	$\widehat{b}_i$	$\widehat{c}_i$
1	1.500	-2.000	0.250	2.612	-0.843	0.213	2.162	-1.725	0.191
2	1.500	-1.500	0.250	2.707	-0.558	0.234	2.334	-1.333	0.212
3	1.500	-1.000	0.250	2.612	-0.297	0.232	2.358	-0.959	0.290
4	1.500	-0.500	0.250	2.751	-0.008	0.241	2.125	-0.728	0.228
5	1.500	0.000	0.250	2.830	0.283	0.258	2.140	-0.364	0.238
6	1.500	0.000	0.250	2.722	0.296	0.264	2.283	-0.338	0.248
7	1.500	0.500	0.250	2.596	0.536	0.241	2.013	-0.046	0.233
8	1.500	1.000	0.250	2.506	0.773	0.231	2.183	0.330	0.253
9	1.500	1.500	0.250	3.150	1.090	0.249	2.478	0.663	0.257
10	1.500	2.000	0.250	2.605	1.331	0.246	2.176	1.004	0.249
11	0.500	-2.000	0.250	1.022	-0.652	0.295	0.745	-1.518	0.284
12	0.500	-1.500	0.250	0.897	-0.523	0.245	0.685	-1.486	0.256
13	0.500	-1.000	0.250	0.860	-0.510	0.171	0.698	-1.071	0.234
14	0.500	-0.500	0.250	0.854	-0.209	0.181	0.715	-0.857	0.187
15	0.500	0.000	0.250	0.938	0.311	0.246	0.770	-0.447	0.225
16	0.500	0.000	0.250	0.904	0.205	0.233	0.715	-0.455	0.228
17	0.500	0.500	0.250	0.971	0.497	0.244	0.647	-0.276	0.173
18	0.500	1.000	0.250	0.924	0.841	0.248	0.762	0.248	0.227
19	0.500	1.500	0.250	0.946	1.087	0.245	0.655	0.562	0.229
20	0.500	2.000	0.250	0.826	1.352	0.225	0.647	0.860	0.216

Response data were generated for two test forms each existing of 20 unique items and 20 common items. All common items had $c_i = .25$, while their a_i and b_i parameters were chosen to represent one of the possible combinations of $b_i = -2(.5)2$ with $a_i = .5,\ 1.5$, using $b_i = 0$ twice to get a total of 20 items. All unique items had $c_i = .25$ as well, but their a_i and b_i parameters were randomly sampled from $U(.5, 2)$ and $N(0, 1)$, respectively. The first calibration had ability parameters for 10,000 test takers sampled from $N(-.5, 2)$; for the second calibration, the parameters for 10,000 test takers were sampled from $N(.5, 1.5)$. The item parameters for the two test forms were estimated separately using the *MIRT Scaling Program*, version 1.0 (Glas, 2010), with MML estimation with $\theta \sim N(0, 1)$ in both runs.

Table 2.1 contains the generating and estimated values of the common item parameters. Observe that even though the response data were generated for exactly the same sets of parameters for the common items, the parameters of the common items in the two calibrations were estimated to be quite different. As argued in our earlier discussion of (2.13)-(2.14), the reason is the different effect of the $\theta \sim N(0, 1)$ restriction on all parameter values in the presence of test takers with different abilities in the two calibrations.

Table 2.2. *Estimated (co)variances for the estimators of the common item parameters*

Common	Calibration 1						Calibration 2					
Item	$\widehat{\sigma}^2_{a_i}$	$\widehat{\sigma}^2_{b_i}$	$\widehat{\sigma}^2_{c_i}$	$\widehat{\sigma}_{a_ib_i}$	$\widehat{\sigma}_{a_ic_i}$	$\widehat{\sigma}_{b_ic_i}$	$\widehat{\sigma}^2_{a_i}$	$\widehat{\sigma}^2_{b_i}$	$\widehat{\sigma}^2_{c_i}$	$\widehat{\sigma}_{a_ib_i}$	$\widehat{\sigma}_{a_ic_i}$	$\widehat{\sigma}_{b_ic_i}$
1	0.023	0.001	0.002	0.001	0.006	0.001	0.019	0.004	0.012	0.002	0.012	0.005
2	0.020	0.001	0.001	0.001	0.003	0.001	0.017	0.002	0.003	0.000	0.006	0.001
3	0.015	0.001	0.000	0.001	0.002	0.000	0.015	0.001	0.001	0.000	0.003	0.001
4	0.015	0.001	0.000	0.002	0.001	0.000	0.009	0.001	0.001	0.001	0.002	0.001
5	0.018	0.001	0.000	0.004	0.001	0.000	0.010	0.001	0.000	0.002	0.002	0.001
6	0.017	0.001	0.000	0.004	0.001	0.000	0.011	0.001	0.000	0.001	0.001	0.000
7	0.015	0.002	0.000	0.004	0.001	0.000	0.009	0.001	0.000	0.002	0.001	0.001
8	0.016	0.003	0.000	0.006	0.001	0.000	0.014	0.002	0.000	0.004	0.001	0.000
9	0.044	0.006	0.000	0.016	0.001	0.000	0.020	0.003	0.000	0.007	0.001	0.000
10	0.033	0.009	0.000	0.016	0.001	0.000	0.022	0.006	0.000	0.011	0.001	0.001
11	0.008	0.036	0.006	0.015	0.006	0.014	0.009	0.215	0.031	0.039	0.015	0.081
12	0.007	0.053	0.007	0.018	0.006	0.019	0.009	0.292	0.036	0.047	0.017	0.103
13	0.007	0.055	0.008	0.017	0.007	0.021	0.008	0.180	0.021	0.034	0.012	0.061
14	0.006	0.048	0.005	0.016	0.005	0.016	0.007	0.125	0.015	0.027	0.009	0.043
15	0.008	0.032	0.002	0.014	0.003	0.008	0.007	0.073	0.007	0.020	0.006	0.022
16	0.007	0.036	0.002	0.014	0.004	0.009	0.007	0.109	0.009	0.026	0.008	0.031
17	0.007	0.027	0.001	0.013	0.003	0.006	0.007	0.152	0.011	0.031	0.008	0.040
18	0.008	0.036	0.001	0.016	0.003	0.006	0.008	0.066	0.004	0.021	0.005	0.015
19	0.009	0.035	0.001	0.017	0.002	0.005	0.009	0.136	0.005	0.033	0.006	0.025
20	0.010	0.062	0.001	0.023	0.003	0.008	0.010	0.139	0.004	0.034	0.006	0.023

Table 2.2 summarizes the covariance matrices for the MML estimators for each of the common items produced by the scaling program. The data in this table are needed to evaluate the estimates of the parameters for the linking function between the two calibrations. Observe that the variances for the item parameter estimates are as expected for data sets of the current size and generating parameter values.

We know that the c_i parameters in the two calibrations are linked by the identity transformation. Thus, for example, if we needed to know the value of the c_i parameter that an arbitrary item in the first calibration would have had in the second calibration, we could just use $\widehat{c}_i$ obtained in the first calibration as its estimate. (For the common items, it makes more sense to pool their two estimates, though.) For the other parameters, we need to know the linking parameters u and v, which can be estimated simply by plugging $\widehat{a}_i$ and $\widehat{b}_i$ for each common item into (2.44)–(2.45). The results are shown in Table 2.3. Although each of these 20 estimates reveals the same trend, they show considerable random variation. In order to evaluate the variation, Table 2.3 also gives the estimated standard errors for each $\widehat{u}_i$ and $\widehat{v}_i$, which were derived from the (co)variances for a_i and b_i in Table 2.2 using the (first-order) multivariate delta method (e.g.,

Table 2.3. *Linking parameters and their standard errors estimated for each common item*

Common Item	$\widehat{u}_i$	$\widehat{\sigma}_{u_i}$	$\widehat{v}_i$	$\widehat{\sigma}_{v_i}$
1	1.208	0.105	-0.707	0.104
2	1.160	0.088	-0.686	0.069
3	1.108	0.077	-0.630	0.045
4	1.295	0.082	-0.718	0.047
5	1.322	0.088	-0.738	0.074
6	1.192	0.079	-0.691	0.069
7	1.290	0.086	-0.737	0.106
8	1.148	0.084	-0.557	0.134
9	1.271	0.111	-0.723	0.228
10	1.197	0.117	-0.589	0.287
11	1.372	0.210	-0.624	0.413
12	1.309	0.218	-0.801	0.516
13	1.232	0.194	-0.443	0.425
14	1.194	0.178	-0.607	0.406
15	1.218	0.171	-0.826	0.396
16	1.264	0.191	-0.714	0.444
17	1.501	0.238	-1.022	0.573
18	1.213	0.184	-0.772	0.490
19	1.444	0.256	-1.008	0.724
20	1.277	0.247	-0.866	0.811

Casella & Berger, 2002, sect. 5.5.4). As the size of these standard errors indicate, we should have expected a considerable amount of variation indeed.

Obviously, as each $\widehat{u}_i$ and $\widehat{v}_i$ is an estimate of the same u and v, respectively, rather than using them individually, they should be combined into overall estimates. A natural suggestion is to use their precision-weighted average, with the inverse of their squared standard errors, $\sigma_{u_i}^{-2}$ and $\sigma_{v_i}^{-2}$, as measure of precision. The estimator of u is then

$$\widehat{u} = \left(\sum_{i=1}^{20} \widehat{\sigma}_{u_i}^{-2} \widehat{u}_i \right) \Bigg/ \left(\sum_{i=1}^{20} \widehat{\sigma}_{u_i}^{-2} \right) , \tag{2.61}$$

with estimated standard error

$$\widehat{\sigma}_u = \left(\sum_{i=1}^{20} \widehat{\sigma}_{u_i}^{-2} \right)^{-1/2} , \tag{2.62}$$

with similar expressions for the estimator of v.

Table 2.4 shows these overall estimates along with those for the mean/mean and mean/sigma methods. The former were obtained by plugging the estimates of the a_i and b_i parameters into (2.6) and (2.9); the latter by plugging the estimates of the b_i parameters into (2.7) and (2.9).

Table 2.4. *Overall estimates of linking parameters and their standard errors*

Method	$\widehat{u}$	$\widehat{\sigma}_u$	$\widehat{v}$	$\widehat{\sigma}_v$
Precision-Weighted Average	1.226	0.026	-0.684	0.023
Mean/Mean	1.237	0.027	-0.706	0.078
Mean/Sigma	1.197	0.118	-0.696	0.084

The standard errors for these two methods were calculated from the (co)variances for the item parameter estimates in Table 2.4 using the same the multivariate delta method. The differences between the results for all three methods were generally substantial, with the precision-weighted method being uniformly best. Especially the results for the v parameter are revealing. Whereas the precision-weighted method produced an acceptable low standard error for it, the other two methods lagged behind considerably. In more practical terms these results suggest that, even with 20 common items, these two methods are likely to seriously misspecify the location of the parameters mapped from one calibration onto the values they would have obtained in another. The extremely large errors for the mean/sigma method are assumed to be due to its ignoring of the unique information in the estimates of the a_i parameters.

It is also interesting to inspect how these overall estimates of the standard errors behave as a function of the number of common items. The curves in Figure 2.2 were obtained by adding the common items to the linking design, one at a time beginning with the first item in Table 2.3. The precision-weighted method produced results substantially better and never worse than those for the mean/mean and mean/sigma method. In fact, it already reached stability for both estimates after some five common items, whereas there still was considerable room for the other two methods to converge. Also, observe the lack of monotonicity in the curves for the standard errors of u and v for the mean/sigma method and in the one for the standard error of v for the mean/mean method. Whereas one would expect a decrease of them with the extra information in each common item added to the linking design, these methods actually showed a considerable increase for the eleventh and twelfth item. Finally, use of the precision-weighted method with the first ten items in Figure 2.2 would yield linking estimates already superior to those for all 20 items for the mean/mean and mean/sigma methods.

The results in Figure 2.2 highlight the importance of the relative precision of the linking parameter estimates contributed by each individual item to the linking design. Surprisingly, if we had added the items in a different order, different results would have been found. Figure 2.3 illustrates the linking errors for the same total set of common items as in Figure 2.2, but added by increasing item difficulty rather than increasing item discrimination as in Table 2.2. Note that, of course, the overall error associated with all 20 common items remains the same. But

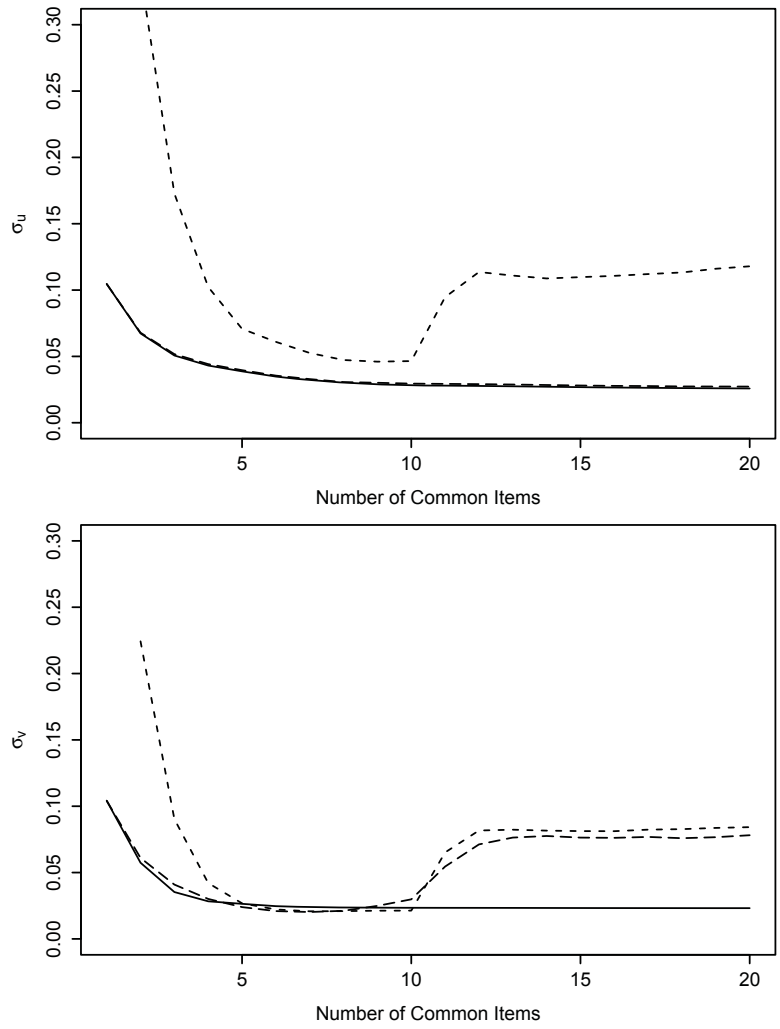

Figure 2.2. Estimated standard errors for linking parameters u and v for the precision-weighted (solid), mean/mean (longdash), and mean/sigma (shortdash) methods as a function of the number of common items in the linking design.

the final result is now reached along different trajectories for all three methods. The difference between the results in Figure 2.2 and Figure 2.3 suggests further research on the use of optimal design principles to find the best possible subset of linking items for the linking design from the larger set of candidate items typically available in practical situations.

This example was only to explore what might be possible once the problem of parameter linking in IRT has been provided with a solid formal foundation. Research on the new estimation method, including comparative studies with other linking methods using empirical data, attempts to explain the aberrances produced by the traditional linking methods, and an analysis

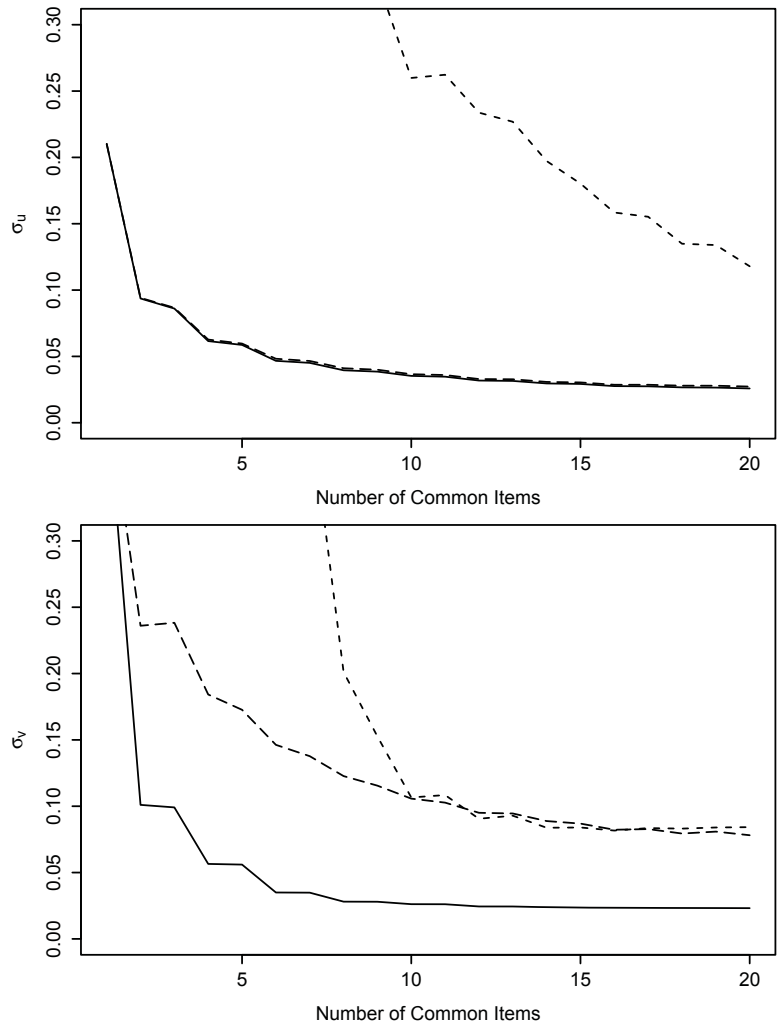

Figure 2.3. Estimated standard errors for linking parameters u and v for the precision-weighted (solid), mean/mean (longdash), and mean/sigma (shortdash) methods as a function of the number of common items in the linking design for a different order of the items than in Table 2.2.

of the consequences of the confounding of linking error with estimation error in the c_i parameters by the response-function methods, is currently conducted. Also, as just noted, further research is needed on how to find the optimal linking design given a set of candidate common items.

2.7 Concluding Remarks

We began this article with a review of the traditional conception of parameter linking in IRT, which appears to have been motivated largely by the notion of equating θ scores on scales with an indeterminate zero and unit, that is, interval scales in the tradition of S. S. Stevens' (1946) classification. An obvious way to remove the indeterminacy of a θ scale, according to the tradition, would be to set its zero and unit equal to the mean and standard deviation of the abilities for a population of test takers, and the only reason why we need to link parameters from different calibrations would be to correct for differences between population distributions. Besides, because they are not assumed to be affected by this choice of zero and unit, the c_i parameters could be treated as invariant. Lord (1980, sect. 3.5) is quite direct in his claim as to this last point.

The fundamental notion underlying the necessity of parameter linking in IRT, however, is not that of an "interval scale" for the θ parameters in the response model, but possible lack of identifiability of any of its parameters. We have been able to present several results due to this reconceptualization. First, although never formally derived before, the linking functions for the standard reparameterization of the 3PL model (Theorem 4) appear to have the general linear form for the a_i, b_i and θ_p parameters assumed in the current literature, but with the different slope and intercept parameters u and v in (2.37) and (2.38), respectively, and the identity function $\varphi_c(c) = c$ added for the c_i parameters. The definitions of u and v allowed us to derive the different solutions for a few minimal linking designs in (2.44)–(2.48). Second, the alternative slope-intercept reparameterization of the 3PL model appears to have a serious impact on the linking problem. Not only have the linking functions for its α_i, β_i, and ϑ_p parameters a different form with one more unknown linking parameter w (Theorem 6), its parameters are identifiable only for the unpractical type of design that has both common items and common test-takers (Theorem 7). Third, as a more general result, we now know from Theorem 3 that for any other monotone, continuous response model the linking functions take the general form of componentwise monotone vector function—a fact that will simplify our explorations of the linking functions required for nearly every other IRT model currently used in educational and psychological testing. Fourth, although all linking functions were derived for true model parameters and their subsequent estimation was not the focus of this paper, it is already clear that we will have to deviate from the estimation methods currently practised. For instance, as illustrated by our example, rather than estimating linking parameter u as the ratio of the mean of estimates of the a_i parameters in (2.6) for the common items, it is much more efficient to use an estimate based on the (precision-weighted) mean of their ratios. Fifth, the

derivation of the linking function for the c_i parameters in (2.35) helps us to evaluate their role in the currently used estimation methods discussed in the introductory section. The Stocking-Lord and Haebara methods admit estimation error in the c_i parameters into their estimates of the linking parameters, but the mean/mean and mean/sigma methods ignore these parameters entirely. At first sight, the lack of identifiability of the c_i parameters seems to suggest the choice of a method from the former rather than the latter category. But the fact that their linking function is the identity function $\varphi_c(c) = c$ implies that, once they have been made identifiable, no further linking is necessary. Consequently, unlike the mean/mean and mean/sigma methods, the Haebara and Stocking-Lord methods confound linking error in the a_i, b_i, and θ_p parameters with estimation error in the c_i parameters.

Any choice of identifiability restrictions has an element of arbitrariness to it, and the practice of making the 3PL model identifiable using restrictions that include the mean ($\mu_\theta = 0$) or standard deviation ($\sigma_\theta = 1$) of the ability parameters for the test takers in the calibration study therefore cannot be wrong. Nevertheless, the reliance on the notion of randomly sampling from some population of test-takers sometimes automatically associated with it is potentially dangerous. For instance, it easily leads to the idea that we now estimate a population mean and standard deviation and therefore have to account for their sampling error. Indeed, a recent study advocated this idea, along with the claim that large-scale educational assessments tend to overlook the design effects on linking error due to the typical clustering of test-takers during sampling (Doorey, 2011, p. 6). However, as demonstrated by Theorem 4, the shape of the true linking functions for the 3PL model does not depend on the actual identifiability restrictions imposed on the calibration studies, let alone on any population parameters adopted for them or even a specific choice of sampling design used to estimate such parameters.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics, 17*, 261–269.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147–172.

Bartels, R. (1985). Identification in econometrics. *The American Statistician, 39*, 102–104.

Bechger, T. M., Verhelst, N. D., & Verstralen, H. H. F. M. (2001). Identifiability of nonlinear logistic models. *Psychometrika, 66*, 357–372.

Bechger, T. M., Verstralen, H. H. F. M., & Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika, 67*, 123–136.

Bekker, P. A., Merckens, A., & Wansbeek, T. J. (1994). *Identification, equivalent models, and computer algebra*. Boston: Academic Press.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Doorey, N. A. (2011). *Addressing two commonly unrecognized sources of score instability in annual state assessments*. Washington, DC: Council of Chief State School Officers.

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.) (2007). *Linking and aligning scores and scales*. New York: Springer.

Fischer, G. H. (2004). Remarks on "Equivalent linear logistic test models" by Bechger, Verstralen, and Verhelst (2002). *Psychometrika, 69*, 305–315.

Fisher, F. M. (1961). Identifiability criteria in nonlinear systems. *Econometrica, 29*, 574–590.

Fisher, F. M. (1965). Identifiability criteria in nonlinear systems: A further note. *Econometrica, 33*, 197–205.

Gabrielsen, A. (1978). Consistency and identifiability. *Journal of Econometrics, 8*, 261–263.

Glas, C. A. W. (2010). *MIRT: Multidimensional item response theory, version 1.01* [Computer software and manual]. Enschede, The Netherlands: University of Twente. Retrieved from `http://www.utwente.nl/gw/omd/en/employees/employees/glas.doc/`.

Haebara, T. (1980). Equating logistic ability scales by weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Haberman, S. (2009). *Linking parameter estimates derived from item response model through separate calibration* (Research Report 09-40). Princeton , NJ: Educational Testing Service.

Holland, P. W., & Rubin, D. B. (Eds.) (1982). *Test equating*. New York: Academic Press.

Kim, S., Harris, D. J., & Kolen, M. J. (2010). Equating with polytomous item response models. In Nering, M. L., & Ostini, R. (Eds), *Handbook of polytomous items response theory models* (pp. 257–291). New York: Taylor & Francis.

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica, 17*, 125–144.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

Luijben, T. C. W. (1991). Equivalent models in covariance structure analysis. *Psychometrika, 56*, 653-665.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Maris, G. (2002). *Concerning the identification of the 3PL model* (Measurement and Research Department Reports 2002-3). Arnhem, The Netherlands: Cito.

Maris, G., & Bechger, T. (2004). Equivalent MIRD models. *Psychometrika, 69*, 627–639.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three-parameter logistic model. *Measurement: Interdisciplinary Research and Perspectives, 7*, 75–88.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reiersøl, O. (1950). On the identifiability of parameters in Thurstone's multiple factor analysis. *Psychometrika, 15*,121–149.

Revuelta, J. (2009). Identifiability and equivalence of GLLIRM models. *Psychometrika, 74*, 257–272.

Richmond, J. (1974). Identifiability in linear models. *Econometrica, 42*, 731–736.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica, 39*, 577–591.

Sahoo, P. K., & Kannappan, P. (2011). *Introduction to functional equations*. Boca Raton, FL: Chapman & Hall/CRC.

San Martín, E. Gonzáles, J., & Tuerlinkckx, F. (2009). Identified parameters, parameters of interest and their relationships. *Measurement: Interdisciplinary Research and Perspective, 7*, 97–105.

San Martín, E. Gonzáles, J., & Tuerlinkckx, F. (2015). On the identifiability of the fixed-effects 3PL model. *Psychometrika, 80*, 450–467.

San Martín, E., Jara, A., Rolin, J.-M., & Mouchart, M. (2011). On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika, 76*, 385–409.

San Martín, E., Rolin, J.-M., & Castro, L. M. (2013). Identification of the 1PL model with guessing parameter: Parametric and semi-parametric results. *Psychometrika, 78*, 341–379.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Baco Raton, FL: Chapman & Hall/CRC.

Small, C. G. (2007). *Functional equations and how to solve them*. New York: Springer.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Tsai, R.-C. (2000). Remarks on the identifiability of Thurstonian ranking models: Case V, Case III, or neither? *Psychometrika, 65*, 233–240.

van der Linden, W. J. (2010). Linking response-time parameters onto a common scale. *Journal of Educational Measurement, 47*, 92–114.

Volodin, N., & Adams, R. J. (2002). *The estimation of polytomous item response models with many dimensions* (Internal Report). Parkville, Victoria, Australia: University of Melbourne, Faculty of Education.

von Davier, A. A. (Ed.) (2011). *Statistical models for test equating scaling, and linking*. New York: Springer.

von Davier, M., & von Davier, A. A. (2011). A general model for IRT scale linking and scale transformations. In A. A. von Davier (Ed.), *Statistical models for test equating scaling, and linking* (pp. 225–242). New York: Springer.

Wood, R. (1978). Fitting the Rasch model—A heady tale. *British Journal of Mathematical and Statistical Psychology, 31*, 27–32.

Chapter 3

Estimating Linking Functions for Response Model Parameters[1]

3.1 Introduction

A model lacks identifiability if any of its parameters does. In the case of the well-known item response theory (IRT) models for dichotomous responses, while their probabilities of a correct response are identifiable, the parameters used to model them typically are not. For each of these models, multiple parameter vectors can yield an identical likelihood function. They therefore have no unique true values for their parameters, and additional restrictions have to be imposed to identify one of the parameter vectors. Consequently, if parameter estimates from two or more different calibrations are to be compared, linking must be conducted to correct for the differences between their restrictions. As linking functions have to be inferred from estimated model parameters, parameter estimation error automatically propagates into linking error.

The lack of identifiability described above specifically holds for the three-parameter logistic (3PL) model, which explains the probabilities of a correct response $U_{pi} = 1$ on items $i = 1, ..., I$ for test-taker p with ability $\theta_p \in \mathbb{R}$ as

$$\Pr\{U_{pi} = 1; \theta_p\} \equiv p(\theta_p; a_i, b_i, c_i) \equiv c_i + (1 - c_i) \{1 + \exp[-a_i(\theta_p - b_i)]\}^{-1}, \tag{3.1}$$

[1]Barrett, M. D., & van der Linden, W. J. (2015). *Estimating linking functions for response model parameters.* Manuscript submitted for publication.

where $b_i \in \mathbb{R}$ and $a_i > 0$ are parameters for the difficulty and discriminating power of item i, respectively, and $c_i \in (0, 1]$ represents the height of the lower asymptote to the probability for the item. In order to make the model identifiable, many common commercial and open-source calibration tools assume a normal true theta distribution with $\mu_\theta = 0$ and $\sigma_\theta = 1$ during maximum marginal likelihood (MML) estimation. Note that although these restrictions look formally identical, in different calibrations they restrict the mean and variance of the θ parameters for different sets of examinees and are therefore empirically different. As a result, estimates from parameter estimation with different instances of these restrictions are not directly comparable, and linking is required. The model in (3.1) will be central in the rest of this paper, but our results hold equally well for the 2PL and Rasch model upon setting of the appropriate item parameters to a common constant.

The problem of estimating linking functions has been addressed in earlier literature, mainly for the mean/mean (Loyd & Hoover, 1980) and mean/sigma (Marco, 1977) methods, as well as the response-function methods by Haebara (1980) and Stocking and Lord (1980). Standard errors for each of these linking methods were obtained by Ogasawara (2000, 2001, 2011). The goal of the current paper is to report research on a new linking method for common-item and common-test-taker designs based on the statistical plausible idea of calculating precision-weighted averages of estimates of the linking function parameters from the minimal linking elements in these designs (single items; pairs of items; pairs of test-takers) from which these parameters are already exactly identified. We also derive the asymptotic standard errors of estimation (ASEs) for these linking parameters for different linking designs, discuss their nature, present results for several empirical examples of recent linking problems, and show their improvements on those for the current mean/mean and mean/sigma methods.

Before doing so, we will review the linking functions for the 3PL model recently derived directly from the presence of different identifiability restrictions in different calibrations (van der Linden and Barrett, in press). The derivation did prove the generally assumed linearity of the linking functions currently in use for the θ_p, a_i, and b_i parameters but provided definitions of their parameters that deviated from those used in the current methods. It also added a explicit linking function for the c_i parameters. As explained later, the new definitions of the linking parameters give us the possibility to derive their true values from different item response model parameters for any feasible linking design. The result for the c_i parameters implies that their estimation error is correctly ignored in the mean/mean and mean/sigma methods but leads to confounding of linking error for the response-function methods. For this reason we have omitted a comparison between the ASEs for our new method and the latter in our examples. Instead, more general conclusions as to the role of the guessing parameters in linking are presented.

3.2 Linking Functions, Designs, and Equations

Linking requires selection of a linking design and inference of the function that maps the true values of the parameters in one calibration onto the values that would have been obtained for the identifiability restrictions used in the other calibration. As shown by van der Linden and Barrett (in press, Theorem 3), linking functions for monotone response models are always component-wise monotone. Consequently, for the 3PL model they can be contained as the solution to the functional equation

$$\frac{\varphi_\gamma(\gamma)}{1+\exp[\varphi_a(a)(\varphi_\theta(\theta)-\varphi_b(b))]} = \frac{\gamma}{1+\exp[a(\theta-b)]}, \tag{3.2}$$

with $c = 1-\gamma^{-1}$, where each component $\varphi_a(a), \varphi_b(b), \varphi_c(c)$ and $\varphi_\theta(\theta)$ can be assumed to be monotone. Solving (3.2) for each of them gives

$$\varphi_a(a) = u^{-1}a, \tag{3.3}$$

$$\varphi_b(b) = ub + v, \tag{3.4}$$

$$\varphi_c(c) = c, \tag{3.5}$$

$$\varphi_\theta(\theta) = u\theta + v, \tag{3.6}$$

with

$$u = \frac{\varphi_\theta(\theta)-\varphi_b(b)}{\theta-b},\ \theta \neq b, \tag{3.7}$$

and

$$v = \varphi_b(b) - ub = \varphi_\theta(\theta) - u\theta. \tag{3.8}$$

Fortunately, the functions for the a_i, b_i, and θ_p parameters are linear, as assumed for those currently in use. However, the solution reveals two new results. First, the formal definition of u in (3.7) as the ratio of the difference between the ability and difficulty parameter for the same *arbitrary* test-taker and item in the two calibrations differs from the one for the current mean/mean and mean/sigma methods, which define u directly as the ratio of the means of the discrimination parameters or the variances of the difficulty or ability parameters *actually* present in the linking design. Consequently, the definition of v, which depends on u, is different as well. As shown below, one of the benefits of the definitions in (3.7)–(3.8) is the flexibility it gives us in deriving solutions for u and v for different linking designs. Second, the identity function for the c_i parameters is a new result. Basically, it tells us that once the model is made identifiable, no matter the restrictions used to do so, the c_i parameters are always directly comparable. As

already indicated, the result will be used below to clarify the hitherto ambiguous role of the c_i parameters in the current linking methods.

Notice the different nature of these functions for the different types of parameters. The identity function for the c_i parameters in (3.5) does not involve any unknown quantities, and is always ready for use. The function for the a_i parameter in (3.3) involves one unknown linking parameter, u,which has to be estimated before it can be used. Likewise, the functions for the b_i and θ_j parameters depend on two unknown linking parameters, u and v, and their use requires the estimation of two unknown parameters.

Several steps are required to estimate u and v. First, a linking design has to be adopted, which can be defined as a combination of two calibration designs with common items and/or test-takers. From now on, we use sub-index $t = 1, 2$ to denote the two different calibrations. Thus, $(a_{i_t}, b_{i_t}, c_{i_t})$ and θ_{p_t} denote the true values of the parameters for item i and test-taker p in the tth calibration, respectively.

Second, substituting the values for the common item and/or ability parameters into the left-hand and right-hand sides of (3.3)–(3.8), we create a system of equations in the unknown linking parameters u and v for the given linking design. This step thus involves a change from the system of equations of $\varphi_\xi(\xi)$ in the variables $\xi = (a, b, c, \theta)$ in (3.3)–(3.8) into a different system of equations of u and v in the true values of $(a_{i_t}, b_{i_t}, c_{i_t})$ and/or θ_{p_t}. It is precisely at this second step that the formal definitions of u and v in (3.7)–(3.8) become convenient.

Third, the system has to be solved for u and v. However, in order to be able to do so, a necessary condition for the system is to be identified. It is important to observe that we now have met two different types of identifiability requirements: (i) the earlier identifiability of the system of the equations for the 3PL model (3.1) in all item and ability parameters in the two calibrations and (ii) identifiability of the system of equations for linking parameters u and v in the parameters $(a_{i_t}, b_{i_t}, c_{i_t})$ and/or θ_{p_t} for the common items and test-takers in the linking design only. The former led to restrictions to be imposed in the two different calibration studies; the latter involves a restriction on our choice of linking design—a condition that appears to be mild, easy to check, but nevertheless critical. From now on, we use $m = 1, ..., M$ to denote the minimal linking element (common items or test-takers) that results in identifiability of u and v for the chosen linking design.

Fourth, to become practical, we need to derive estimates of u and v from the estimates of $(a_{i_t}, b_{i_t}, c_{i_t})$ and/or θ_{p_t}. At this point, estimation error in the parameters for the common items

and/or test-takers propagates into the estimates of linking parameters u and v and it becomes important to know the standard errors of the latter.

We first discuss the choice of linking design and the derivation of the system of equations for the linking parameters u and v, postponing the statistical problem of their estimation. As an introductory example, consider the simple case of one common test-taker p responding to the same common item i in two calibrations. Substitution of $(\theta_{p_1},\theta_{p_2})$ and (b_{i_1},b_{i1}), with $\theta_{p_1} \neq b_{i_1}$ into (3.7)–(3.8) would give us

$$u = \frac{\theta_{p_2} - b_{i_2}}{\theta_{p_1} - b_{i_1}} \tag{3.9}$$

and

$$v = b_{i_2} - ub_{i_1} = \theta_{p_2} - u\theta_{p_1}, \tag{3.10}$$

as a solution. But, although this system is obviously identified, due to memory and/or learning effects, designs with the same test-takers responding to the same items twice are seldom practically feasible. Our main focus is therefore on linking design with common items or common test-takers only.

3.2.1 Common-Item Designs

Suppose we have observations of the pairs of parameter values (a_{1_1}, a_{1_2}) and (b_{1_1}, b_{1_2}) for a design with as linking element m a single common item which, for convenience, we denote as $i = 1$. The two linking parameters follow immediately upon substitution of these observations into (3.3)–(3.4) as

$$u = \frac{a_{1_1}}{a_{1_2}} \tag{3.11}$$

and

$$v = b_{1_2} - ub_{1_1}. \tag{3.12}$$

Thus, more technically, the observation of one common item is already sufficient to generate a system of equations in u and v from (3.3)–(3.8) that has a unique solution.

Alternatively, for a design with as linking element m a pair of two common items, $i = 1$ and $i = 2$, we will have pairs of observations on the difficulty parameters of the two items, (b_{1_1}, b_{1_2}) and (b_{2_1}, b_{2_2}). Assuming $b_{1_1} \neq b_{2_1}$, substitution into (3.4) and elimination of v gives us

$$u = \frac{b_{1_2} - b_{2_2}}{b_{1_1} - b_{2_1}} \tag{3.13}$$

along with a choice from

$$v = b_{i_2} - ub_{i_1},\ i = 1, 2. \tag{3.14}$$

The different solutions for u and v provided by (3.11)–(3.14) are mathematically equivalent. But our choice from them does have immediate statistical relevance. For instance, estimation of the ratio of two discrimination parameters in (3.11) implies a different error for $\widehat{u}$, and consequently for $\widehat{v}$, than estimation of the ratio of the differences between the two difficulty parameters in (3.13). Such differences are the main subject of the remainder of this paper.

3.2.2 Common-Test-Taker Designs

A case comparable to (3.11)–(3.12) does not hold for designs with common test-takers only. If a single test-taker $p = 1$ is attempted as the only linking element, one pair of values $(\theta_{1_1}, \theta_{1_2})$ is observed, and it appears impossible to derive a system of equations from (3.3)–(3.8) for it that allows us to identify u and v.

For a linking element m comprised of a pair of test-takers, $p = 1$ and $p = 2$, we have two common test-takers with pairs of ability values $(\theta_{1_1}, \theta_{1_2})$ and $(\theta_{2_1}, \theta_{2_2})$, and the system does have a unique solution. Assuming $\theta_{1_1} \neq \theta_{2_1}$, substitution of the values into (3.8) and elimination of v gives

$$u = \frac{\theta_{1_2} - \theta_{2_2}}{\theta_{1_1} - \theta_{2_1}}, \tag{3.15}$$

while substitution into (3.8) results in

$$v = \theta_{p_2} - u\theta_{p_1},\ p = 1, 2. \tag{3.16}$$

The linking designs chosen in practice typically have $M > 2$ linking elements. But before dealing with these cases, we address the consequences of the possible different choices from (3.11)–(3.16) for the standard errors of estimation for u and v based on these simple designs.

3.3 Standard Errors of Estimated Linking Parameters

Item parameters are typically estimated using MML estimation, that is, with the ability parameters as nuisance parameters integrated out of the likelihood function assuming a common marginal distribution for each of them. As we will demonstrate later, in order to derive the standard errors for the linking parameters for a common-item design, the only input we need from

the computer program used to estimate the item parameters are the estimates of their covariance matrices typically produced along with them. On the other hand, individual ability parameters are typically estimated in a next step using simple maximum likelihood estimation, this time with the item parameters assumed to be known. Consequently, for a common-test-taker design, we only need estimates of the variances of the ability parameter estimates, which likewise can be taken from the output by the computer program used to produce them. For Bayesian estimation, we would need to derive the posterior distributions of the linking parameters—a topic reserved for subsequent research.

It is important to note that error in the estimates of the linking parameters is a function of estimation error in the common item or ability parameters only. More specifically, although sometimes claimed otherwise (e.g., Doorey, 2011), linking error does not depend on any parameters describing some population from which the test-takers are supposed to be sampled, even if restrictions on such parameters (e.g., $\mu_\theta = 0$ and $\sigma_\theta = 1$) were imposed on the item calibration to identify the response model parameters. In order to derive the impact of random error in the values of the item or ability parameters on estimates of the linking parameters, we only need to account for our choice from (3.11)–(3.16). It is not necessary to know how the common items or test-takers that figure in them were actually obtained.

3.3.1 Common-Item Designs

For the case of a linking element consisting of one common item $i = 1$, let $\boldsymbol{\eta} = (\eta) = (u, v)$ denote the vector of the linking parameters, $\boldsymbol{\xi}_{1_t} = (a_{1_t}, b_{1_t})$ the values of the a_i and b_i parameters of the common item in calibration $t = 1, 2$. In addition, we will use $\boldsymbol{\xi} = (\boldsymbol{\xi}_{1_t})$. The system of linking equations in (3.11)–(3.12) defines a vector function $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$. We will use ξ to denote an arbitrary element of $\boldsymbol{\xi}$.

The (first-order) multivariate delta method (e.g., Casella & Berger, 2002, sect. 5.5.4) enables us to approximate the 2×2 covariance matrix of $\widehat{\boldsymbol{\eta}}$ as

$$\mathrm{Cov}(\widehat{\boldsymbol{\eta}} \mid \boldsymbol{\eta}) = \boldsymbol{J}'_{\varphi} \mathrm{Cov}(\widehat{\boldsymbol{\xi}} \mid \boldsymbol{\xi}) \boldsymbol{J}_{\varphi}, \tag{3.17}$$

where

$$\boldsymbol{J}_{\varphi} = \left(\frac{\partial \eta}{\partial \xi} \right) \tag{3.18}$$

is the 4×2 Jacobian matrix associated with $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$ and $\mathrm{Cov}(\widehat{\boldsymbol{\xi}} \mid \boldsymbol{\xi})$ the 4×4 covariance matrix for the estimators of the elements of $\boldsymbol{\xi}$.

Because of independence between the two administrations of any common item (different test-takers), $\mathrm{Cov}(\widehat{\boldsymbol{\xi}} \mid \boldsymbol{\xi})$ is block diagonal with the two 2×2 covariance matrices

$$\begin{bmatrix} \sigma^2_{a_{1_t}} & \sigma_{a_{1_t}b_{1_t}} \\ \sigma_{a_{1_t}b_{1_t}} & \sigma^2_{b_{1t}} \end{bmatrix}, \; t = 1, 2, \tag{3.19}$$

of the estimators of the two item parameters in each calibration as blocks.

As we are interested in the standard errors of the estimators of $\boldsymbol{\eta} = (u_m, v_m)$, the only elements required are those in the diagonal of the left-hand side of (3.17), which follow from (3.19) using the corresponding elements of (3.18) as

$$\mathrm{Var}(\widehat{u}_m \mid \boldsymbol{\xi}) = \sum_{t=1}^{2} \left[\left(\frac{\partial u}{\partial a_{1t}} \right)^2 \sigma^2_{a_{1t}} + 2 \frac{\partial u}{\partial a_{1t}} \frac{\partial u}{\partial b_{1t}} \sigma_{a_{1_t}b_{1_t}} + \left(\frac{\partial u}{\partial b_{1t}} \right)^2 \sigma^2_{b_{1_t}} \right] \tag{3.20}$$

and

$$\mathrm{Var}(\widehat{v}_m \mid \boldsymbol{\xi}) = \sum_{t=1}^{2} \left[\left(\frac{\partial v}{\partial a_{1t}} \right)^2 \sigma^2_{a_{1_t}} + 2 \frac{\partial v}{\partial a_{1t}} \frac{\partial v}{\partial b_{1t}} \sigma_{a_{1_t}b_{1_t}} + \left(\frac{\partial v}{\partial b_{1t}} \right)^2 \sigma^2_{b_{1_t}} \right]. \tag{3.21}$$

With the function $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$ defined by the equations in (3.11)–(3.12), the partial derivatives in (3.18) are

$$\frac{\partial u}{\partial a_{1_1}} = \frac{1}{a_{1_2}} \tag{3.22}$$

$$\frac{\partial u}{\partial b_{1_t}} = 0, \; t = 1, 2 \tag{3.23}$$

$$\frac{\partial v}{\partial a_{1_1}} = -\frac{b_{1_1}}{a_{1_2}} \tag{3.24}$$

$$\frac{\partial v}{\partial b_{1_1}} = -\frac{a_{1_1}}{a_{2_2}} = -u \tag{3.25}$$

$$\frac{\partial u}{\partial a_{1_2}} = \frac{-a_{1_1}}{a^2_{1_2}} = -\frac{u}{a_{1_2}} \tag{3.26}$$

$$\frac{\partial v}{\partial a_{1_2}} = \frac{a_{1_1}b_{1_1}}{a^2_{1_2}} = \frac{ub_{1_1}}{a_{1_2}} \tag{3.27}$$

$$\frac{\partial v}{\partial b_{1_2}} = 1 \tag{3.28}$$

To obtain the standard errors of the linking parameters, we only need to substitute the two covariance matrices for the common item in (3.19) and the partial derivatives in (3.22)–(3.28) into (3.20) and (3.21) and take their square roots. Simplifying the resulting expressions, we

have

$$\sigma_{u_m} = \left(\frac{\sigma^2_{a_{1_1}} + u^2 \sigma^2_{a_{1_2}}}{a^2_{1_2}} \right)^{1/2}, \tag{3.29}$$

with $a_{1_2} \neq 0$ and

$$\sigma_{v_m} = \left\{ \frac{b^2_{1_1} \left(\sigma^2_{a_{1_1}} + u^2 \sigma^2_{a_{1_2}} \right) + 2 a_{1_1} b_{1_1} \left(\sigma_{a_{1_1} b_{1_1}} + \sigma_{a_{1_2} b_{1_2}} \right) + a^2_{1_2} \left(u^2 \sigma^2_{b_{1_1}} + \sigma^2_{b_{1_2}} \right)}{a^2_{1_2}} \right\}^{1/2}, \tag{3.30}$$

with $a_{1_2} \neq 0$. For this case of exactly one common item as linking element, the estimates of u and v from the mean/mean moment method specialize to the one of (3.11)–(3.12), whereas the mean/sigma methods fails to produce either. The same holds for their standard errors. However, as the number of items increases, the method by which these estimates are combined directly influences the estimates and their standard errors. Statistical intuition suggests a precision-weighted average of estimates across all linking elements rather than any of the existing methods. We will further pursue this idea when we discuss the more general case of multiple common items or test-takers below.

Likewise, for the case of a pair of common items as linking element, we can use the system of linking equations in (3.13)–(3.14). However, there are now two equivalent definitions of v and we have information about this parameter from both items. It makes therefore sense to combine them and estimate

$$v = \frac{1}{2} \sum_{i=1}^{2} \left(b_{i_2} - u b_{i_1} \right), \tag{3.31}$$

allowing the delta method to account for the estimate of u shared by both items.

The system of linking equations in this case thus defines a vector function $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$ with $\boldsymbol{\xi} = (\boldsymbol{\xi}_t)$, where $\boldsymbol{\xi}_t = (b_{1_t}, b_{2_t})$. For this case, $\boldsymbol{J}_\varphi$ is still of dimension 4×2 and, because of local independence, $\mathrm{Cov}(\widehat{\boldsymbol{\xi}} \mid \boldsymbol{\xi})$ becomes a 4×4 diagonal matrix with elements $\sigma^2_{b_{i_t}}$, $i = 1, 2$, $t = 1, 2$. Consequently, $\mathrm{Cov}(\widehat{\boldsymbol{\eta}} \mid \boldsymbol{\eta})$ is now diagonal as well.

The partial derivatives of $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$ are now

$$\frac{\partial u}{\partial b_{1_1}} = \frac{b_{2_2} - b_{1_2}}{(b_{1_1} - b_{2_1})^2} = -\frac{u}{b_{1_1} - b_{2_1}} \tag{3.32}$$

$$\frac{\partial u}{\partial b_{1_2}} = \frac{1}{b_{1_1} - b_{2_1}} = \frac{u}{b_{1_2} - b_{2_2}} \tag{3.33}$$

$$\frac{\partial u}{\partial b_{2_1}} = \frac{b_{1_2} - b_{2_2}}{(b_{1_1} - b_{2_1})^2} = \frac{u}{b_{1_1} - b_{2_1}} \tag{3.34}$$

$$\frac{\partial u}{\partial b_{2_2}} = \frac{-1}{b_{1_1} - b_{2_1}} = -\frac{u}{b_{1_2} - b_{2_2}} \tag{3.35}$$

$$\frac{\partial v}{\partial b_{1_1}} = \frac{-b_{2_1}(b_{1_2} + b_{2_2})}{(b_{1_1} - b_{2_1})^2} \tag{3.36}$$

$$\frac{\partial v}{\partial b_{1_2}} = \frac{-b_{2_1}}{b_{1_1} - b_{2_1}} \tag{3.37}$$

$$\frac{\partial v}{\partial b_{2_1}} = \frac{-b_{1_1}(b_{1_2} + b_{2_2})}{(b_{1_1} - b_{2_1})^2} \tag{3.38}$$

$$\frac{\partial v}{\partial b_{2_2}} = \frac{b_{1_1}}{b_{1_1} - b_{2_1}} \tag{3.39}$$

The standard errors follow upon substitution of (3.32)–(3.39) into

$$\text{Var}(\widehat{u}_m \mid \boldsymbol{\xi}) = \sum_{i=1}^{2}\sum_{t=1}^{2}\left[\left(\frac{\partial u}{\partial b_{i_t}}\right)^2 \sigma^2_{b_{i_t}}\right] \tag{3.40}$$

and

$$\text{Var}(\widehat{v}_m \mid \boldsymbol{\xi}) = \sum_{i=1}^{2}\sum_{t=1}^{2}\left[\left(\frac{\partial v}{\partial b_{it}}\right)^2 \sigma^2_{b_{i_t}}\right] \tag{3.41}$$

as

$$\sigma_{u_m} = \left[u^2\left(\frac{\sigma^2_{b_{1_1}} + \sigma^2_{b_{2_1}}}{(b_{1_1} - b_{2_1})^2} + \frac{\sigma^2_{b_{1_2}} + \sigma^2_{b_{2_2}}}{(b_{1_2} - b_{2_2})^2}\right)\right]^{1/2}, \tag{3.42}$$

with $b_{1_1} - b_{2_1} \neq 0$ and $b_{1_2} - b_{2_2} \neq 0$ and

$$\sigma_{v_m} = \left[\frac{\left(b^2_{2_1}\sigma^2_{b_{1_1}} + b^2_{1_1}\sigma^2_{b_{2_1}}\right)(b_{1_2} + b_{2_2})^2 + \left(b^2_{2_1}\sigma^2_{b_{1_2}} + b^2_{1_1}\sigma^2_{b_{2_2}}\right)(b_{1_1} - b_{2_1})^2}{(b_{1_1} - b_{2_1})^4}\right]^{1/2}, \tag{3.43}$$

with $b_{1_1} - b_{2_1} \neq 0$.

In this case, the estimate of u from the mean/sigma method now specializes to the one of (3.13), but the mean/mean method produces a different estimate. The same relationships holds for their standard errors. As for the estimate of v, our treatment of the two different choices from (3.14) below will yield results different from those for both current methods.

3.3.2 Common-Test-Taker Designs

For a design with a pair of common test-takers, $p = 1$ and $p = 2$, as linking element, the system of linking equations in (3.15)–(3.16) is used. Again, as there are two equivalent definitions of v, with a common parameter u shared by the two test-takers, analogous to (3.31) we estimate

$$v = \frac{1}{2}\sum_{p=1}^{2}\left(\theta_{p_2} - u\theta_{p_1}\right). \tag{3.44}$$

The system of linking equations defines a vector function $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$, with $\boldsymbol{\xi}_t = (\theta_{1_t}, \theta_{2_t})$ and $\boldsymbol{J}_\varphi$ of dimension 4×2 again. Because of independence between the test-takers, $\mathrm{Cov}(\widehat{\boldsymbol{\xi}} \mid \boldsymbol{\xi})$ is a 4×4 diagonal matrix with elements $\sigma^2_{\theta_{p_t}}$. Hence, $\mathrm{Cov}(\widehat{\boldsymbol{\eta}} \mid \boldsymbol{\eta})$ will be diagonal again. Consequently, the standard errors for the linking parameters we seek are the square roots of

$$\mathrm{Var}(\widehat{u}_m \mid \boldsymbol{\xi}) = \sum_{p=1}^{2}\sum_{t=1}^{2}\left[\left(\frac{\partial u}{\partial \theta_{p_t}}\right)^2 \sigma^2_{\theta_{p_t}}\right] \tag{3.45}$$

and

$$\mathrm{Var}(\widehat{v}_m \mid \boldsymbol{\xi}) = \sum_{p=1}^{2}\sum_{t=1}^{2}\left[\left(\frac{\partial v}{\partial \theta_{p_t}}\right)^2 \sigma^2_{\theta_{p_t}}\right] \tag{3.46}$$

Partial derivatives $\partial\eta/\partial\xi$ are entirely analogous to (3.32)-(3.37). Substitution of them into (3.45)–(3.46) gives

$$\sigma_{u_m} = \left[u^2\left(\frac{\sigma^2_{\theta_{1_1}} + \sigma^2_{\theta_{2_1}}}{\left(\theta_{1_1} - \theta_{2_1}\right)^2} + \frac{\sigma^2_{\theta_{1_2}} + \sigma^2_{\theta_{2_2}}}{\left(\theta_{1_2} - \theta_{2_2}\right)^2}\right)\right]^{1/2}, \tag{3.47}$$

with $\theta_{1_1} - \theta_{2_1} \neq 0$ and $\theta_{1_2} - \theta_{2_2} \neq 0$ and

$$\sigma_{v_m} = \left[\frac{\left(\theta^2_{2_1}\sigma^2_{\theta_{1_1}} + \theta^2_{1_1}\sigma^2_{\theta_{2_1}}\right)\left(\theta_{1_2} + \theta_{2_2}\right)^2 + \left(\theta^2_{2_1}\sigma^2_{\theta_{1_2}} + \theta^2_{1_1}\sigma^2_{\theta_{2_2}}\right)\left(\theta_{1_1} - \theta_{2_1}\right)^2}{\left(\theta_{1_1} - \theta_{2_1}\right)^4}\right]^{1/2}, \tag{3.48}$$

with $\theta_{1_1} - \theta_{2_1} \neq 0$. Estimates of the standard errors $\sigma_{\theta_{j_t}}$ can be taken from the output of the computer program used to estimate the ability parameters in the two calibrations. However, for ability estimation with item parameters that can be assumed to be known, we could also use the well-known (asymptotic) relationship

$$\sigma_{\theta_{p_t}} = \left[\sum_{i;p_t} I_i(\theta_{p_t})\right]^{-1/2}, \tag{3.49}$$

where $I_i(\theta)$ is Fisher's information on θ in the response to item i,

$$I_i(\theta) = a_i^2 \frac{1 - p_i(\theta)}{p_i(\theta)} \left(\frac{p_i(\theta) - c_i}{1 - c_i} \right)^2 , \tag{3.50}$$

and the sum in (3.49) is across all items taken by test-taker p according to the design of calibration t.

3.4 Multiple Common Items or Test-Takers

As already noted several times, typical linking designs have more than one linking element. In such cases, we have multiple independent estimates of the same linking parameter u, one for each single common item from (3.11), pair of common items from (3.13), or pair of common test-takers from (3.15). Likewise, we have independent estimates of v for each common item in (3.12) and pair of common items or test-takers in (3.31) and (3.44), respectively. Each of these estimates has its own standard error. Thus, we can combine estimates $\widehat{u}_m$ and $\widehat{v}_m$ for $m = 1, ..., M$ linking elements in the design. Standard errors for each linking element in (3.29)–(3.30), (3.42)–(3.43), or (3.47)–(3.48) are denoted as σ_{u_m} and σ_{v_m}.

Statistical intuition suggest pooling the estimates using precision-weighted averaging. The pooled estimator of u then becomes

$$\widehat{u} = \left(\sum_{m=1}^{M} \frac{\widehat{u}_m}{\widehat{\sigma}^2_{u_m}} \right) \Bigg/ \left(\sum_{m=1}^{M} \frac{1}{\widehat{\sigma}^2_{u_m}} \right) , \tag{3.51}$$

with estimated standard error

$$\widehat{\sigma}_u = 1 \Bigg/ \left(\sum_{m=1}^{M} \frac{1}{\widehat{\sigma}^2_{u_m}} \right)^{1/2} , \tag{3.52}$$

where the latter is obtained by substituting estimates for the parameters in the right-hand side of (3.29), (3.42), or (3.47). The estimator and standard error for v are defined entirely analogously.

Observe that this type of pooling is possible only because of the derivation of the standard errors in (3.29)–(3.30), (3.42)–(3.43), and (3.47)–(3.48) for the estimates of u and v at the level of the minimal elements of the linking design for which these parameters are identifiable, a feature missed by the current mean/mean and mean/sigma methods of parameter linking.

3.5 Guessing Parameters

Guessing parameters c_i lack general identifiability; for a few examples illustrating this point, see van der Linden and Barrett (in press, Theorem 2). But, surprisingly, as documented in (3.5), once the 3PL model is identified, the linking function for these parameters is an identity function. This simple fact has two important consequences. First, if the same item is used in two different calibrations, we have two independent estimates of the same true value for it. Consequently, without any linking, we are always able to pool the two estimates to obtain one more accurate estimate. Second, if an item is used in one calibration only, say $t = 1$, and we want to estimate the true value of the guessing parameter it would have obtained in $t = 2$, we should just use its estimate from the former; that is, $\widehat{c}_{i_1}$. No linking parameters have to be estimated, and the only "linking error" involved in the use of $\widehat{c}_{i_1}$ as an estimate of c_{i_2} is its current estimation error.

Further, although for items calibrated under the 3PL model the estimates of c_i do covary with those of a_i and b_i, the standard errors of u and v for a common-item design are independent of it. This fact follows immediately from our previous derivation of them: Extending $\text{Cov}(\widehat{\boldsymbol{\xi}} \mid \boldsymbol{\xi})$ in (3.17) with all covariances between the c_i parameters and the other two parameters of the common items would have led to the requirement of a similar extension of the Jacobean matrix in (3.18) with the partial derivatives $\partial\eta/\partial c_i$, $\eta = a_i, b_i$. However, as $\varphi(c) = c$ is independent of η, these derivatives are equal to zero. Consequently, all terms with the covariances for the c_i parameters in $\text{Cov}(\widehat{\boldsymbol{\eta}} \mid \boldsymbol{\eta})$ in (3.17) would vanish and we would just have retained (3.20)–(3.21).

Linking methods based on the full response functions, such as those by Haebara (1980) and Stocking and Lord (1983), require substitution of estimates of the c_i parameters for the common items. In doing so, they actually confound estimation error for the a_i and b_i parameters, and thus for linking parameters u and v, with that for the c_i parameters. As a result, the standard errors generally ascribed to them can be expected to be inflated.

3.6 Empirical Examples

The goal of our empirical examples is twofold. First, we want to study the behavior of the new precision-weighted estimators of u and v more closely. In particular, we want to know how their ASEs behave as a function of the number of linking elements in the linking design. It seems reasonable to expect the ASE to decrease monotonically in them, but it is also important to

know how quickly it decreases. For instance, common items in a linking design do take away time from the examinees, so we may want to minimize or constrain their number. The two questions will be investigated both for the precision-weighted estimators based on elements of single common items and elements of common pairs of items. Second, we want to compare the behavior of the new estimators with those for the mean/mean and mean/sigma methods, which estimate u as

$$\widehat{u} = \frac{\mu(\widehat{a})}{\mu(\widehat{a}^*)} \tag{3.53}$$

and

$$\widehat{u} = \frac{\sigma(\widehat{b}^*)}{\sigma(\widehat{b})}, \tag{3.54}$$

respectively, while estimating v as

$$\widehat{v} = \mu(\widehat{b}^*) - \widehat{u}\mu(\widehat{b}). \tag{3.55}$$

The comparison will be for empirical data from two real-world testing programs. As it is quite uncommon for testing organizations to conduct linking studies with common test-takers, we are unable to present any empirical examples from studies with this type of design.

Our response data were from two state-wide testing programs with test forms calibrated using the 3PL. Both programs administer a different exam form once a year, with common items in subsequent forms for linking purposes. The data set was for a high-school mathematics exam that included 13 common items between the two forms used in this study, with 36,939 and 37,740 test-takers who had taken the first and second form, respectively. The other data set was for two forms for a high-school reading exam that included 23 common items, with 32,931 and 36,157 test-takers who had taken the first and second form, respectively.

We first estimated the response model parameters in independent calibrations for each different test form from their response data, using the *MIRT Scaling Program,* version 1.01 (Glas, 2010) for MML estimation. A normal theta distribution with identifiability restrictions $\mu_\theta = 0$ and $\sigma_\theta = 1$ was used in each calibration. The software provided the complete information matrix for each item, the inverse of which was used as its asymptotic covariance matrix. Estimated item parameters and covariances for each test administration are included in Tables 3.1–3.4.

To explore the behavior of the ASEs of the precision-weighted average methods as a function of the number of the linking elements in the design, we considered the case of adding identical elements to the design; that is, elements comprised of items with identical parameter values and covariance matrices. Empirical item parameters and covariance matrices from the mathematics

Table 3.1. *High-school mathematics exam, item parameter estimates and covariance for administration* $t = 1$

Item	$\widehat{a}_{i_1}$	$\widehat{b}_{i_1}$	$\widehat{c}_{i_1}$	$\widehat{\sigma}_{a_{i_1}}$	$\widehat{\sigma}_{b_{i_1}}$	$\widehat{\sigma}_{c_{i_1}}$	$\widehat{\sigma}_{a_{i_1}b_{i_1}}$	$\widehat{\sigma}_{a_{i_1}c_{i_1}}$	$\widehat{\sigma}_{b_{i_1}c_{i_1}}$
1	1.8440	-0.0510	0.1500	0.0118	0.0198	0.0098	0.0001	0.0000	0.0002
2	1.3020	-0.1470	0.0540	0.0513	0.0574	0.0292	0.0026	0.0013	0.0016
3	1.2880	0.6620	0.0600	0.0494	0.0518	0.0133	0.0023	0.0006	0.0007
4	1.8370	0.5050	0.1680	0.0667	0.0382	0.0101	0.0023	0.0006	0.0004
5	2.8680	1.1300	0.2110	0.1395	0.0570	0.0044	0.0077	0.0004	0.0002
6	2.2290	1.3970	0.3270	0.1368	0.0854	0.0052	0.0113	0.0005	0.0004
7	0.4500	-0.0970	0.2000	0.0785	0.7915	0.1391	0.0607	0.0107	0.1100
8	1.2630	-0.0670	0.1740	0.0561	0.0663	0.0274	0.0033	0.0014	0.0018
9	3.0870	1.0110	0.0810	0.1247	0.0415	0.0034	0.0050	0.0003	0.0001
10	1.1940	-0.3360	0.2390	0.0621	0.0869	0.0379	0.0048	0.0021	0.0033
11	1.5070	0.3100	0.4110	0.0820	0.0674	0.0161	0.0050	0.0011	0.0010
12	1.1060	-0.9710	0.1290	0.0611	0.1260	0.0786	0.0069	0.0044	0.0098
13	1.0730	2.0530	0.2950	0.1023	0.1723	0.0109	0.0169	0.0010	0.0018

Table 3.2. *High-school mathematics exam, item parameter estimates and covariance for administration* $t = 2$

Item	$\widehat{a}_{i_2}$	$\widehat{b}_{i_2}$	$\widehat{c}_{i_2}$	$\widehat{\sigma}_{a_{i_2}}$	$\widehat{\sigma}_{b_{i_2}}$	$\widehat{\sigma}_{c_{i_2}}$	$\widehat{\sigma}_{a_{i_2}b_{i_2}}$	$\widehat{\sigma}_{a_{i_2}c_{i_2}}$	$\widehat{\sigma}_{b_{i_2}c_{i_2}}$
1	1.6320	-0.1410	0.1270	0.0627	0.0394	0.0214	0.0021	0.0012	0.0008
2	1.2570	-0.1970	0.0520	0.0498	0.0582	0.0299	0.0026	0.0013	0.0017
3	1.3150	0.6850	0.0630	0.0500	0.0498	0.0124	0.0023	0.0005	0.0006
4	1.6640	0.4480	0.1100	0.0589	0.0375	0.0112	0.0020	0.0006	0.0004
5	2.7810	1.2100	0.2100	0.1369	0.0600	0.0042	0.0080	0.0004	0.0002
6	2.6350	1.4930	0.3800	0.1902	0.1047	0.0044	0.0195	0.0005	0.0003
7	0.4500	-0.2030	0.1990	0.0836	0.8542	0.1539	0.0700	0.0126	0.1314
8	1.4590	-0.2960	0.2500	0.0671	0.0568	0.0281	0.0033	0.0017	0.0016
9	2.5230	1.1290	0.0830	0.0954	0.0421	0.0037	0.0039	0.0002	0.0001
10	1.2250	-0.4350	0.2080	0.0604	0.0790	0.0389	0.0042	0.0021	0.0030
11	1.3820	0.0520	0.3780	0.0719	0.0689	0.0209	0.0044	0.0013	0.0014
12	1.1700	-0.8020	0.1170	0.0575	0.0947	0.0597	0.0048	0.0031	0.0056
13	1.9300	1.9540	0.3170	0.1634	0.1420	0.0048	0.0226	0.0006	0.0005

exam as given in Table 3.1 were used such that the behavior could be examined for item characteristics.

Our first example focuses on the precision-weighted average of linking parameter estimates with linking elements comprised of single common items in (3.11)–(3.12). Figure 3.1 presents plots of the ASEs of $\widehat{u}$ and $\widehat{v}$ calculated from (3.29)–(3.30) as identical items were added to the design. The identical items represented in each of its panels had the item parameter and covariance matrices of each of the items in Table 3.1. Note that the ASEs of both the $\widehat{u}$ and $\widehat{v}$ parameters for this method demonstrate the desirable characteristic of being monotonically

Table 3.3. *High-school reading exam, item parameter estimates and covariance for administration* $t = 1$

Item	$\widehat{a}_{i_1}$	$\widehat{b}_{i_1}$	$\widehat{c}_{i_1}$	$\widehat{\sigma}_{a_{i_1}}$	$\widehat{\sigma}_{b_{i_1}}$	$\widehat{\sigma}_{c_{i_1}}$	$\widehat{\sigma}_{a_{i_1}b_{i_1}}$	$\widehat{\sigma}_{a_{i_1}c_{i_1}}$	$\widehat{\sigma}_{b_{i_1}c_{i_1}}$
1	1.3250	-1.9470	0.0540	0.0513	0.0833	0.0960	0.0032	0.0043	0.0077
2	2.1590	-1.0330	0.1470	0.0620	0.0171	0.0206	0.0000	0.0010	0.0002
3	1.1500	0.0920	0.0500	0.0395	0.0387	0.0147	0.0013	0.0005	0.0005
4	1.0310	-1.9410	0.0530	0.0478	0.1357	0.1120	0.0055	0.0048	0.0150
5	1.2930	-1.4600	0.0520	0.0444	0.0538	0.0535	0.0017	0.0020	0.0027
6	0.6600	-0.6100	0.0520	0.0463	0.1980	0.0684	0.0086	0.0030	0.0135
7	2.6710	-1.6440	0.1760	0.0979	0.0286	0.0431	-0.0014	0.0033	0.0001
8	0.8400	-1.4830	0.0500	0.0449	0.1523	0.0884	0.0061	0.0036	0.0134
9	1.3560	-1.0990	0.0500	0.0435	0.0373	0.0348	0.0011	0.0013	0.0012
10	0.3240	0.6230	0.1030	0.0862	1.3976	0.1777	0.1186	0.0151	0.2482
11	0.5070	-0.0480	0.0670	0.0553	0.3736	0.0842	0.0199	0.0045	0.0314
12	1.2590	-1.1270	0.0640	0.0418	0.0445	0.0375	0.0014	0.0013	0.0016
13	1.3690	-1.4280	0.0800	0.0472	0.0479	0.0497	0.0016	0.0020	0.0022
14	1.0340	-0.5100	0.0510	0.0392	0.0592	0.0306	0.0020	0.0011	0.0018
15	1.9170	-1.0530	0.1760	0.0577	0.0211	0.0240	0.0003	0.0011	0.0004
16	3.3370	-1.4520	0.1660	0.1212	0.0297	0.0273	-0.0029	0.0024	-0.0002
17	0.8760	0.4040	0.1570	0.0501	0.0951	0.0239	0.0043	0.0011	0.0022
18	1.0860	-0.8240	0.0510	0.0397	0.0558	0.0351	0.0018	0.0012	0.0019
19	1.1480	-0.9210	0.0540	0.0401	0.0519	0.0361	0.0016	0.0013	0.0018
20	1.6130	-0.0240	0.2010	0.0568	0.0292	0.0127	0.0013	0.0006	0.0003
21	0.6680	-0.1250	0.1770	0.0546	0.2056	0.0535	0.0105	0.0027	0.0109
22	0.6250	1.5410	0.1810	0.0695	0.2457	0.0307	0.0163	0.0020	0.0075
23	0.8230	-0.6330	0.0670	0.0432	0.1116	0.0480	0.0043	0.0019	0.0053

decreasing with additional items. Also, note that the initial impact of the addition of items for each of the cases is relatively great, but then quickly tapers off. After four or five items, both ASEs are already half the size of the ones for a design with one item, while after some ten items or so the marginal benefit of each extra item is hardly noticeable. Note that the ASE of $\widehat{v}$ for Item 7 was too large to appear on the plot due to low a_i parameter estimates and large standard errors for its b_i parameter estimates.

Our second example is for the ASEs of the precision-weighted average of linking parameter estimates using linking elements comprised of pairs of common items in (3.13)–(3.14). For this example, again we used the items in Table 3.1, paired in the order they appear in the table (e.g., Pair 1 included Item 1 and 2). Figure 3.2 presents plots of the ASEs of $\widehat{u}$ and $\widehat{v}$ calculated from (3.42)–(3.43) as identical pairs were added to the design. Note that the ASEs for this type of precision-weighted estimate again display the desirable feature of being monotonically decreasing with increase in number of item included in the design. In addition, we can note

Table 3.4. *High-school reading exam, item parameter estimates and covariance for administration* $t = 2$

Item	$\widehat{a}_{i_2}$	$\widehat{b}_{i_2}$	$\widehat{c}_{i_2}$	$\widehat{\sigma}_{a_{i_2}}$	$\widehat{\sigma}_{b_{i_2}}$	$\widehat{\sigma}_{c_{i_2}}$	$\widehat{\sigma}_{a_{i_2}b_{i_2}}$	$\widehat{\sigma}_{a_{i_2}c_{i_2}}$	$\widehat{\sigma}_{b_{i_2}c_{i_2}}$
1	1.2270	-1.9450	0.0520	0.0470	0.0901	0.0938	0.0033	0.0038	0.0083
2	2.0440	-1.0770	0.0980	0.0514	0.0164	0.0200	0.0001	0.0008	0.0002
3	1.2660	0.0920	0.0800	0.0411	0.0335	0.0131	0.0011	0.0004	0.0004
4	0.9240	-2.0110	0.0530	0.0447	0.1667	0.1207	0.0066	0.0049	0.0200
5	1.4600	-1.3780	0.0550	0.0425	0.0365	0.0407	0.0010	0.0014	0.0014
6	0.6480	-0.4770	0.0500	0.0447	0.1948	0.0635	0.0082	0.0027	0.0123
7	2.3580	-1.6770	0.1800	0.0781	0.0272	0.0443	-0.0004	0.0027	0.0005
8	0.8090	-1.4220	0.0510	0.0424	0.1531	0.0834	0.0059	0.0033	0.0127
9	1.3220	-1.1380	0.0510	0.0407	0.0388	0.0354	0.0011	0.0012	0.0013
10	0.3010	0.7920	0.1220	0.0843	1.5622	0.1777	0.1297	0.0147	0.2775
11	0.5020	0.0540	0.0820	0.0544	0.3687	0.0788	0.0193	0.0041	0.0290
12	1.1820	-1.1020	0.0970	0.0409	0.0526	0.0392	0.0017	0.0014	0.0020
13	1.1230	-1.4330	0.0640	0.0418	0.0703	0.0572	0.0024	0.0021	0.0039
14	1.0300	-0.4370	0.0580	0.0377	0.0560	0.0276	0.0018	0.0009	0.0015
15	1.9120	-0.8810	0.1720	0.0534	0.0186	0.0190	0.0003	0.0008	0.0003
16	2.9950	-1.5580	0.1160	0.0952	0.0264	0.0318	-0.0018	0.0022	-0.0001
17	0.8510	0.2690	0.1320	0.0454	0.0916	0.0251	0.0038	0.0010	0.0023
18	1.3480	-1.0420	0.0860	0.0401	0.0358	0.0305	0.0010	0.0010	0.0010
19	1.5100	-1.0620	0.0560	0.0409	0.0279	0.0280	0.0007	0.0009	0.0007
20	1.6190	-0.1190	0.1940	0.0522	0.0263	0.0126	0.0010	0.0005	0.0003
21	0.6370	-0.1650	0.1660	0.0525	0.2209	0.0566	0.0109	0.0028	0.0124
22	0.6430	1.0450	0.1540	0.0576	0.1984	0.0322	0.0108	0.0017	0.0063
23	0.8360	-0.6320	0.0780	0.0419	0.1051	0.0454	0.0040	0.0017	0.0047

another rapid decrease in ASE for the first pairs of common items added to the design, with lower impact as additional pairs are included. Note that the ASE for Pair 4, which included Item 7, was too large to appear on the plot due to the same low a_i parameter estimates and large standard errors for the b_i parameters.

Also, note that the magnitude of the ASEs was generally larger for the estimators when linking elements consisted of common pairs of items than when elements were comprised of single common items (the vertical scales in Figures 3.1 and 3.2 differ by a factor of ten), and that in some cases the ASEs of $\widehat{u}$ were higher than those of $\widehat{v}$ for the pair-based method. The latter was counter-intuitive and yet we found this to be the case when $\widehat{b}_{1_1} - \widehat{b}_{2_1}$, one of the two denominators of the expressions for the ASE of $\widehat{u}_m$ in (3.42), was small. Although $\widehat{v}_m$ is a function of $\widehat{u}_m$ for the pair-based method, its ASE is a non-linear function of the ASE of latter. It is important to note that for the precision-weighted estimate, this dependence of the ASE of $\widehat{u}_m$ on $\widehat{b}_{1_1} - \widehat{b}_{2_1}$ is inconsequential though; the closer $\widehat{b}_{1_1}$ to $\widehat{b}_{2_1}$, the larger the standard error

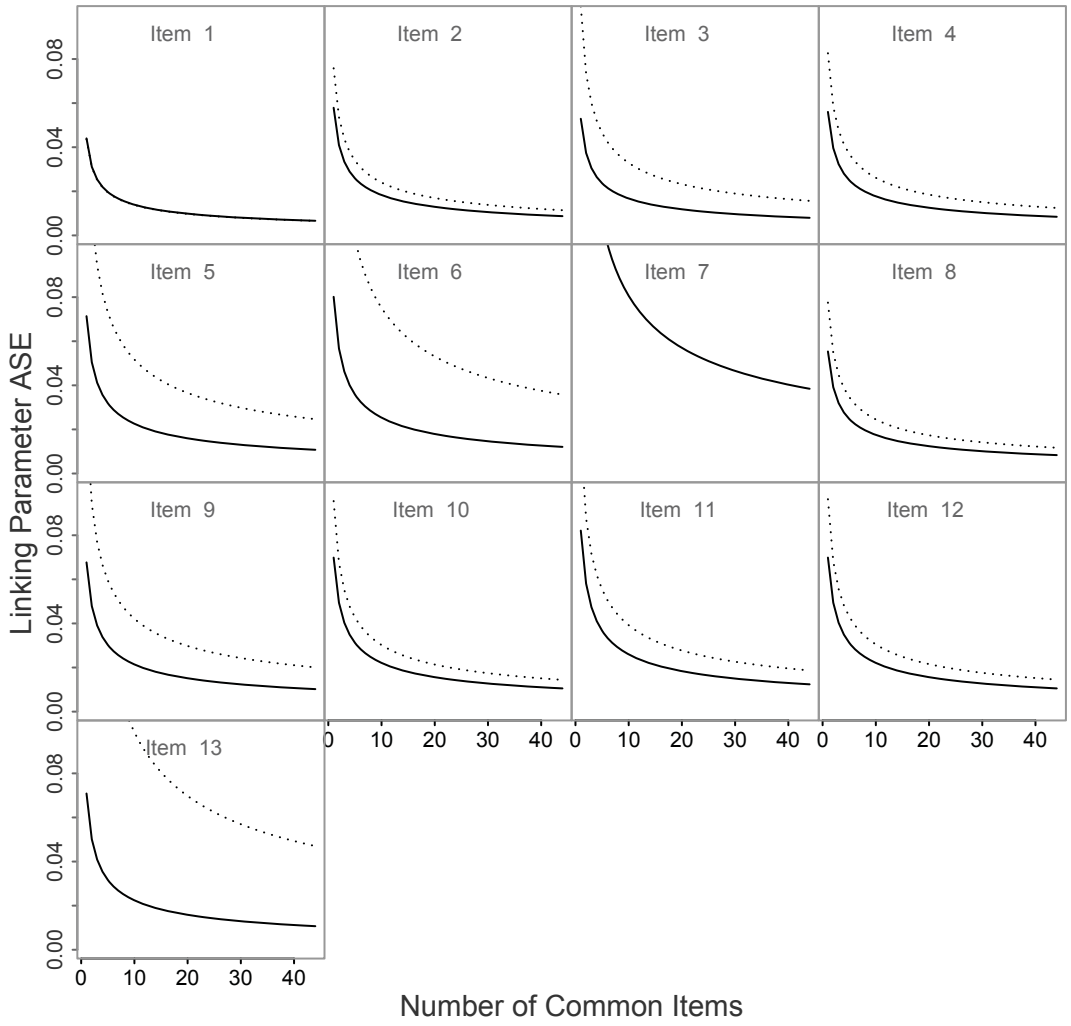

Figure 3.1. ASEs of precision-weighted linking parameter estimates u (solid) and v (dashed) as a function of the number of identical common items in single-item linking elements. Note. The ASE of $\widehat{v}$ for Item 7 is too large to appear on the plot.

for this pair, and the smaller its weight in the overall estimate from the design. In case the two $\widehat{b}_i$ parameters in a pair happen to be equal, its estimate will be automatically suppressed. However, the dependence does have important implications for the optimal selection of common item pairs in a linking design, a topic reserved for future research.

The results in Figures 3.1 and 3.2 suggest better linking for the precision-weighted estimates of the u and v parameters based on linking elements comprised of single items than pairs of items. Apparently, for the same numbers of items, the use of ratios of a_i parameter estimates leads to more stable estimates than the use of ratios of differences between b_i parameter estimates—an observation that makes sense as the latter have both denominators and numerators that are linear combinations of two independent sources of estimation error. More importantly maybe,

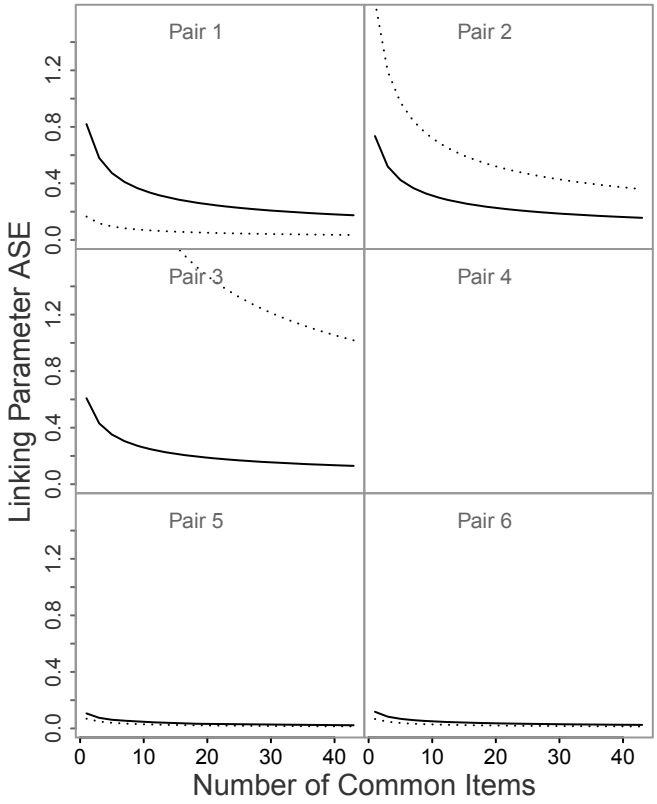

Figure 3.2. ASEs of linking parameter estimates u (solid) and v (dashed) as a function of the number of identical common items in paired-item linking elements. Note. The ASEs for Item Pair 4 are too large to appear on the plot.

just as the mean/sigma method, which performed worse than the mean/mean method in our next examples, the method based on pairs calculates both the estimates of u and v from the difficulty parameters estimates only, ignoring the unique information available in the estimates of the discrimination parameters.

In order to compare the differences between the new and existing linking methods we calculated the linking parameter estimates and ASEs from the item (co)variances in Tables 3.1–3.4 for: (i) the precision-weighted average method based on linking elements of single common items; (ii) the precision-weighted average method based on linking elements of pairs of items; (iii) the mean/mean method in Eqs. 3.53 and 3.55; and (iv) the mean/sigma method in Eqs. 3.54 and 3.55. The ASEs for the last two methods were calculated using the expressions derived by Ogasawara (2000, 2001). The raw results for the items in each design are provided in Tables 3.5–3.8.

Figure 3.3 illustrates the error in the linking parameters for the mathematics exam as different common items were added to their estimates, in the order in which they appeared in the test

Table 3.5. *High-school mathematics exam, $\widehat{u}$ and $\widehat{\sigma}_u$ for different linking parameter estimation methods*

	$\widehat{u}$				$\widehat{\sigma}_u$			
N items	PW-Single	PW-Pair	M/M	M/S	PW-Single	PW-Pair	M/M	M/S
1	1.1299		1.1299		0.0440		0.0440	
2	1.0954	0.5833	1.0890	1.4460	0.0350	0.8198	0.0353	0.3072
3	1.0601		1.0547	1.0832	0.0292		0.0293	0.1078
4	1.0695	1.0967	1.0687	1.0228	0.0259	0.5473	0.0263	0.0756
5	1.0650		1.0567	1.0572	0.0243		0.0292	0.0463
6	1.0465	1.0803	1.0074	1.0639	0.0233	0.4069	0.0302	0.0398
7	1.0461		1.0072	1.0678	0.0232		0.0307	0.2006
8	1.0192	1.0802	0.9915	1.0792	0.0214	0.4069	0.0278	0.1734
9	1.0377		1.0288	1.0873	0.0204		0.0259	0.1558
10	1.0328	1.1559	1.0249	1.0914	0.0196	0.1028	0.0245	0.1321
11	1.0359		1.0298	1.0824	0.0190		0.0235	0.1256
12	1.0296	0.9444	1.0247	1.0440	0.0184	0.0775	0.0225	0.0920
13	0.9998		0.9825	1.0079	0.0178		0.0219	0.0775

Table 3.6. *High-school mathematics exam, $\widehat{v}$ and $\widehat{\sigma}_v$ for different linking parameter estimation methods*

	$\widehat{v}$				$\widehat{\sigma}_v$			
N items	PW-Single	PW-Pair	M/M	M/S	PW-Single	PW-Pair	M/M	M/S
1	-0.0834		-0.0834		0.0436		0.0436	
2	-0.0738	-0.1112	-0.0612	-0.0258	0.0378	0.1657	0.0455	0.0739
3	-0.0609		-0.0475	-0.0519	0.0355		0.0440	0.0485
4	-0.0685	-0.1132	-0.0601	-0.0490	0.0326	0.1649	0.0388	0.0380
5	-0.0641		-0.0426	-0.0428	0.0320		0.0424	0.0328
6	-0.0574	-0.1130	-0.0040	-0.0369	0.0317	0.1648	0.0506	0.0323
7	-0.0575		-0.0183	-0.0478	0.0317		0.1768	0.2351
8	-0.0833	-0.1131	-0.0381	-0.0746	0.0293	0.1648	0.1533	0.1990
9	-0.0845		-0.0378	-0.0660	0.0287		0.1396	0.1831
10	-0.0864	-0.0551	-0.0414	-0.0680	0.0275	0.0638	0.1253	0.1563
11	-0.0957		-0.0637	-0.0843	0.0268		0.1147	0.1423
12	-0.0805	-0.1025	-0.0405	-0.0458	0.0258	0.0461	0.1048	0.1169
13	-0.0744		-0.0313	-0.0419	0.0257		0.0983	0.1087

Table 3.7. *High-school reading exam, $\widehat{u}$ and $\widehat{\sigma}_u$ for different linking parameter estimation methods*

	$\widehat{u}$				$\widehat{\sigma}_u$			
N items	PW-Single	PW-Pair	M/M	M/S	PW-Single	PW-Pair	M/M	M/S
1	1.0799		1.0799		0.0588		0.0588	
2	1.0638	0.9497	1.0651	1.0042	0.0332	0.1336	0.0335	0.0211
3	1.0055		1.0214	1.0073	0.0263		0.0269	0.0334
4	1.0176	1.0001	1.0374	1.0202	0.0248	0.0850	0.0256	0.0275
5	0.9808		1.0053	1.0062	0.0211		0.0218	0.0207
6	0.9824	1.0035	1.0065	1.0012	0.0206	0.0826	0.0217	0.0279
7	1.0004		1.0365	1.0048	0.0193		0.0211	0.0222
8	1.0026	1.0047	1.0366	0.9985	0.0188	0.0825	0.0204	0.0193
9	1.0059		1.0354	1.0009	0.0174		0.0188	0.0178
10	1.0060	1.0052	1.0364	1.0090	0.0173	0.0823	0.0209	0.2227
11	1.0061		1.0354	1.0091	0.0172		0.0210	0.2106
12	1.0121	1.0069	1.0379	1.0069	0.0163	0.0813	0.0197	0.1978
13	1.0256		1.0513	1.0065	0.0158		0.0188	0.1822
14	1.0238	1.0281	1.0483	1.0048	0.0151	0.0693	0.0179	0.1753
15	1.0213		1.0435	0.9972	0.0142		0.0166	0.1666
16	1.0274	1.1193	1.0535	1.0039	0.0137	0.0644	0.0161	0.1550
17	1.0274		1.0526	1.0015	0.0135		0.0158	0.1462
18	1.0024	1.1084	1.0383	1.0092	0.0127	0.0573	0.0150	0.1420
19	0.9719		1.0214	1.0141	0.0119		0.0142	0.1375
20	0.9734	1.0896	1.0198	1.0142	0.0116	0.0469	0.0136	0.1336
21	0.9740		1.0205	1.0143	0.0115		0.0136	0.1305
22	0.9740	1.0747	1.0194	0.9895	0.0115	0.0459	0.0137	0.1145
23	0.9743		1.0184	0.9897	0.0113		0.0135	0.1125

forms. Several things should be noticed. First, for the u parameter, the precision-weighted estimates with single item minimal linking elements outperformed the estimates for those with paired-item linking elements, but those for the v parameter were close. Second, the precision-weighted estimates based on linking elements of single common items performed actually best overall. Third, the mean/sigma method yielded standard errors that strongly violated the intuitive requirement of monotonicity in the number of items for the u parameter, while both the mean/mean and mean/sigma method showed such violations for the v parameter. At a first glance, it was surprising to see such violations. However, upon examination of the standard errors, it became evident that a few of the items, particularly those in the sixth and seventh positions, had much higher standard errors for their item parameter estimates than the other items. The standard errors for the linking parameters of the mean/mean and mean/sigma methods do not account for this, and therefore they rise, rather than fall, when those items are added to the linking design. In contrast, the precision-weighted methods do account for

Table 3.8. *High-school reading exam, $\widehat{v}$ and $\widehat{\sigma}_v$ for different linking parameter estimation methods*

	$\widehat{v}$				$\widehat{\sigma}_v$			
N items	PW-Single	PW-Pair	M/M	M/S	PW-Single	PW-Pair	M/M	M/S
1	0.1575		0.1575		0.0831		0.0831	
2	0.0497	-0.0960	0.0760	-0.0148	0.0414	0.3506	0.0586	0.0581
3	0.0336		0.0066	-0.0070	0.0324		0.0380	0.0460
4	0.0415	-0.0049	0.0171	-0.0036	0.0313	0.0478	0.0519	0.0599
5	0.0026		0.0007	0.0018	0.0260		0.0421	0.0489
6	0.0045	-0.0043	0.0246	0.0185	0.0259	0.0477	0.0513	0.0678
7	0.0132		0.0545	0.0158	0.0252		0.0487	0.0593
8	0.0172	-0.0043	0.0623	0.0145	0.0247	0.0477	0.0475	0.0579
9	0.0098		0.0540	0.0114	0.0212		0.0425	0.0518
10	0.0098	-0.0041	0.0643	0.0355	0.0212	0.0477	0.2064	0.3255
11	0.0100		0.0669	0.0417	0.0212		0.1927	0.2956
12	0.0250	-0.0032	0.0692	0.0391	0.0193	0.0475	0.1768	0.2751
13	0.0490		0.0812	0.0360	0.0185		0.1639	0.2578
14	0.0512	0.0022	0.0795	0.0373	0.0177	0.0464	0.1522	0.2392
15	0.0682		0.0844	0.0391	0.0164		0.1421	0.2243
16	0.0680	0.0064	0.0865	0.0366	0.0162	0.0463	0.1346	0.2136
17	0.0660		0.0714	0.0241	0.0162		0.1271	0.1938
18	0.0014	-0.0399	0.0446	0.0179	0.0149	0.0394	0.1194	0.1856
19	-0.0583		0.0211	0.0144	0.0137		0.1123	0.1780
20	-0.0624	-0.0617	0.0140	0.0090	0.0129	0.0304	0.1069	0.1671
21	-0.0624		0.0121	0.0069	0.0128		0.1026	0.1589
22	-0.0626	-0.0619	-0.0133	-0.0351	0.0128	0.0302	0.0991	0.1387
23	-0.0619		-0.0128	-0.0337	0.0128		0.0949	0.1337

it. For these methods, item parameter estimates that contribute relatively less to the overall estimate than more precisely estimated parameters are automatically downgraded. But, albeit now slightly, the ASEs of the linking parameters still decrease with the addition of them.

For the full linking design with 13 common items used in the testing program, the precision-weighted method based on linking elements of single common items would generally have been the best choice.

Likewise, Figure 3.4 illustrates the standard errors of the linking parameter estimates for the reading exam as different items were added to the linking design, again in the order in which they were included in the test form. The results for this exam were similar to those for the previous mathematics exam, with the exception of generally better results for the precision-weighted estimates based on elements of common item pairs. Again, the mean/mean and mean/sigma methods showed serious violations of the monotonicity requirement. In this case,

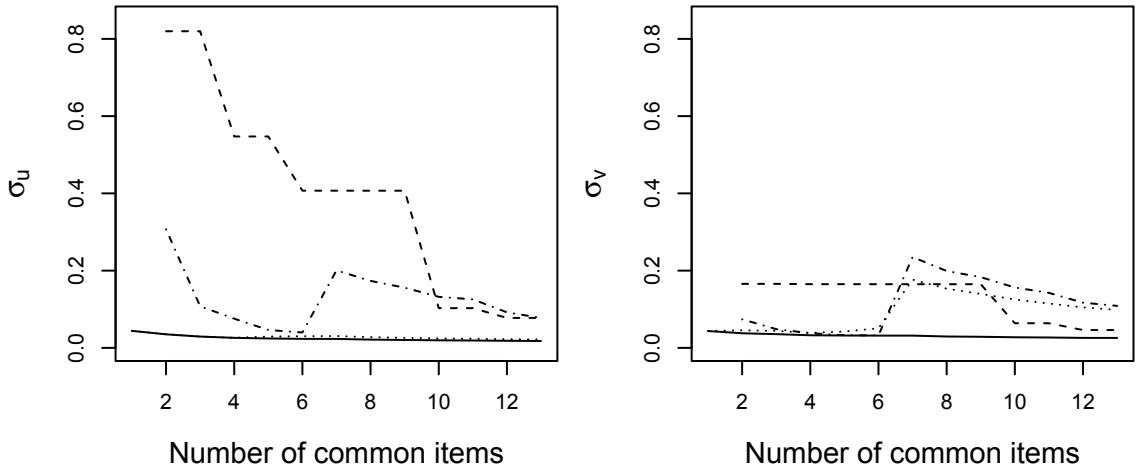

Figure 3.3. Comparison of the ASEs of linking parameter estimates for precision-weighted single item (solid), precision-weighted paired item (longdash), mean/mean (shortdash) and mean/sigma (dashdot) methods on high-school math exam as common items are added to the design.

it was the item in the tenth position that caused a remarkable increase in the ASE for both methods, especially for the v parameter.

For the design with all common items, the precision-weighted method based on linking elements of single common items would have been the best choice for this testing program again.

3.7 Concluding Remarks

In order to compare response model parameter estimates from one independent calibration to another, they need to be adjusted for the impact of different identifiability restrictions imposed on each of them. The linking functions necessary for the adjustment can be derived from such functional equations as in (3.2). Linking function parameters need to be calculated from estimates of the common item or ability parameters in the design chosen for the linking study.

In this research, we used a multivariate delta method to derive closed-form expressions for the asymptotic standard errors for the linking parameters for the 3PL model from the (co)variances for the parameter estimates for linking elements comprised of single common items, individual pairs of common items, and individual pairs of common test-taker in the linking design. To arrive at an overall estimate for the entire linking design, precision-weighted averages were

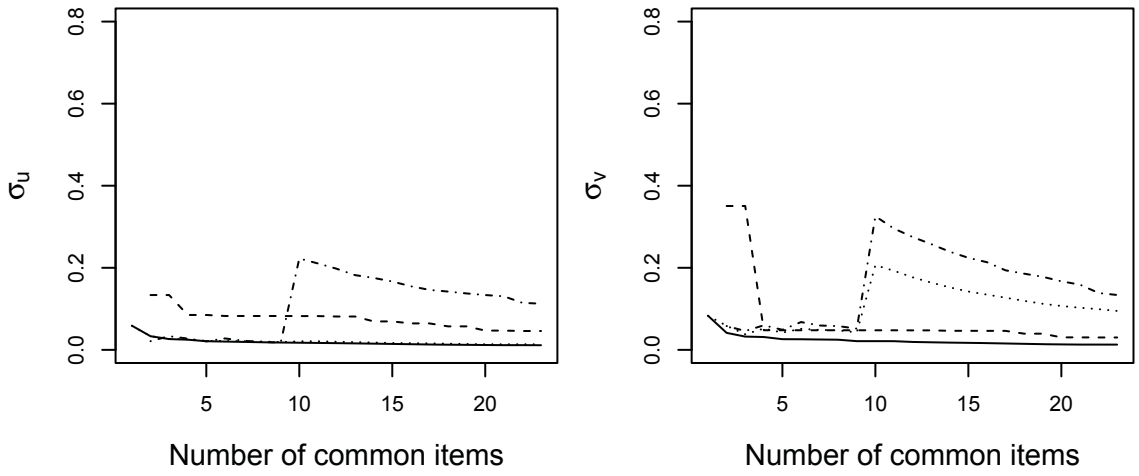

Figure 3.4. Comparison of the ASEs of linking parameter estimates for precision-weighted single item (solid), precision-weighted paired item (longdash), mean/mean (shortdash) and mean/sigma (dashdot) methods on high-school reading exam as common items are added to the design.

calculated from these individual linking elements, using the inverse of their standard errors as weights. The approach not only makes the contributions from each of these linking elements in the linking design transparent, but also favors those with the least amount of error in their parameter estimates by weighing them more heavily. Our empirical results demonstrated desirable behavior of the overall standard errors for the new precision-weighted average method, especially for the version in which the linking elements were single common items. Generally, its performance was better than for the mean/mean and mean/sigma methods, which failed to show the required monotone decrease in linking error when the number of common items was increased when some of them had less accurate estimates of their parameters.

Our approach also provided clarity to the role of the c_i parameters in linking, and brought into question the use of such response-function linking methods as by Haebara (1980) and Stocking and Lord (1983), which add unnecessary error in the c_i parameters to the standard errors of linking for the 3PL model.

The research presented in this paper assumed only random linking error due to response model parameter estimation. In practice, it is also possible for systematic error to be present. For example, there may be flaws in the implementation of the linking design, or a lack of model fit for some of the common items or test-takers. Therefore, it is important that linking designs are robust, which is another reason why it is important to include more than one linking element

in them. We also recognize that a linking design we did not address is sometimes used in practice, that of randomly equivalent groups. This type of design does entail sampling error in addition to model parameter estimation error, an issue not addressed in our current research. Besides, as its implementation is typically difficult, this type of design easily incurs systematic error.

Future research is also required to deal with other response models (e.g., polytomous models), the statistical aspects of optimal design of linking studies, the consequences of linking error for the ability estimates of individual test-takers as well as estimates of functionals defined on ability distributions (e.g., certain percentiles), and the consideration of single-step versus chain linking.

Acknowledgment

The authors are indebted to Cees A. W. Glas for his support during their use of the *MIRT Scaling Program* (version 1.01) software.

References

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Doorey, N. A. (2011). *Addressing two commonly unrecognized sources of score instability in annual state assessments*. Washington, DC: Council of Chief State School Officers.

Glas, C.A.W. (2010). *MIRT: Multidimensional item response theory, version 1.01* [Computer software and manual]. Enschede, The Netherlands: University of Twente. Retrieved from `http://www.utwente.nl/gw/omd/en/employees/employees/glas.doc/`.

Haebara, T. (1980). Equating logistic ability scales by weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce), 51*, 1–23.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25,* 53–67.

Ogasawara, H. (2011). Applications of asymptotic expansion in item response theory linking. In A. A. von Davier (Ed.), *Statistical models for test equating scaling, and linking* (pp. 261–280). New York: Springer.

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., & Barrett, M. D. (in press). Linking item response model parameters. *Psychometrika*. doi: 20.1007/s11336-015-9469-6.

Chapter 4

Optimal Linking Design for Response Model Parameters[1]

4.1 Introduction

When estimating the item and ability parameters in one of the popular item response theory (IRT) models, additional restrictions are required to identify the model. As a result, unless all parameters are estimated simultaneously, estimates for a test form used in one administration are not directly comparable with those for another form used in a different administration. Linking is the act of adjusting one set of parameter estimates to make them comparable with another set. A linking function is the mathematical function necessary to make the adjustment.

The above holds true for the well-known three-parameter logistic (3PL) model, which explains the probabilities of a correct response $U_{pi} = 1$ for test-taker $p = 1, ..., P$ with ability $\theta_p \in \mathbb{R}$ on items $i = 1, ..., I$ as

$$\Pr\{U_{pi} = 1; \theta_p\} \equiv p(\theta_p; a_i, b_i, c_i) \equiv c_i + (1 - c_i) \left\{1 + \exp[-a_i(\theta_p - b_i)]\right\}^{-1}, \tag{4.1}$$

where $b_i \in \mathbb{R}$ and $a_i > 0$ are parameters for the difficulty and discriminating power of item i, respectively, and $c_i \in (0, 1]$ represents the height of the lower asymptote to the probability for the item. During item calibration, many of the common commercial and open-source calibration tools assume a normal true theta distribution with $\mu_\theta = 0$ and $\sigma_\theta = 1$ in maximum marginal

[1]Barrett, M. D., & van der Linden, W. J. (2015). *Optimal linking design for response model parameters.* Manuscript submitted for publication.

likelihood (MML) estimation for identification. While these restrictions look formally identical, when used in different calibrations they restrict the mean and variance of the θ parameters for different sets of test-takers and are therefore empirically different. The 3PL model will be used in the examples in this paper, but the results hold equally well for the 2PL and Rasch model upon fixing the appropriate item parameters to a common constant.

Although the problem of estimating linking functions for the dichotomous item response theory models has been addressed in earlier literature, mainly for the mean/mean (Loyd & Hoover, 1980), mean/sigma (Marco, 1977) methods, and the response-function methods by Haebara (1980) and Stocking and Lord (1980), these methods were motivated by an intuitive notion of the need to adjust a scale for θ with arbitrary zero and unit. Rather, van der Linden and Barrett (in press) derived the linking function for the 3PL model directly from the presence of different identifiability restrictions in different calibrations, as the solution to a functional equation in four unknowns, which was shown to be a set of component-wise monotone linking functions $\varphi_a(a), \varphi_b(b), \varphi_c(c)$ and $\varphi_\theta(\theta)$ that map each of the a_i, b_i, c_i, and θ_p parameters in one calibration onto those in another. The resulting functions were

$$\varphi_a(a) = u^{-1}a, \tag{4.2}$$

$$\varphi_b(b) = ub + v, \tag{4.3}$$

$$\varphi_c(c) = c, \tag{4.4}$$

$$\varphi_\theta(\theta) = u\theta + v, \tag{4.5}$$

with

$$u = \frac{\varphi_\theta(\theta) - \varphi_b(b)}{\theta - b},\ \theta \neq b, \tag{4.6}$$

and

$$v = \varphi_b(b) - ub = \varphi_\theta(\theta) - u\theta. \tag{4.7}$$

Observe that the result confirms the assumed linearity of the functions in the earlier methods but with different definitions for slope and intercept parameters u and v, respectively. The result also contains a formal linking function for the c_i parameters.

Estimation of u and v requires several steps. A full discussion of these steps is provided in Barrett and van der Linden (2015a) but will be briefly summarized here. The first step is to establish the linking design, which basically is the choice of common elements (item and/or test-takers) in the two calibrations that support the estimation of the linking function. We use sub-index $t = 1, 2$ to denote two different calibrations. Thus, $(a_{i_t}, b_{i_t}, c_{i_t})$ and θ_{p_t} are the

true values of the parameters for item i and test-taker p in the tth calibration, respectively. As shown by these authors, linking could be conducted for any design with at least one common item or pair of common test-takers, but not for a design with one common test-taker, for which the linking parameters appear to be unidentified. The decision of which of the many possible linking designs to use is typically constrained by the context in which the linking is to be conducted; for example, it may not be appropriate to include common test-takers when the test items are liable to memory or learning effects, or just because some test agency does not allow its test-takers to take a test twice. The second step requires the substitution of the values of the common item parameters $(a_{i_t}, b_{i_t}, c_{i_t})$ and/or common test-taker parameters θ_{p_t} into (4.2)–(4.7), thereby creating a system of equations in the unknown linking parameters u and v for the chosen linking design. Third, the system has to be solved for u and v. Fourth, substitution of the estimates of the common parameters gives us the required estimates of u and v.

At this point, we should be interested in the propagation of estimation error in the item and test-taker parameters into the linking function estimates. As shown in Barrett and van der Linden (2015a), the actual choice of the common items and/or test-takers has substantial implications for the linking error. Hence, the problem of minimizing linking error through the application of optimal design principles addressed in this paper. More specifically, we will review the impact of the choice of common items or test-takers on the asymptotic standard error (ASE) of linking for the design, provide relevant background information on the methodology of optimal test design, and demonstrate how to select the linking design to minimize or bound overall linking error.

4.2 Review of Standard Errors of Estimation for Linking Parameters

Asymptotic standard errors for the linking parameters were derived by Barrett and van der Linden (2015a) for the linking elements comprised of one common item, two common items, or two common test-takers in the linking design, using a (first-order) delta method. While linking could be conducted with just one of these elements, multiple elements $m = 1, .., M$ are necessary to reduce linking error to an acceptable level. The overall estimate used in this paper is based on the idea of precision-weighted averaging. As will be shown below, understanding

the precise individual contribution of each single element to the overall error in the linking parameter estimates provides a benefit when selecting them for inclusion in the linking design.

4.2.1 One Common Item

From (4.2)–(4.3), when the linking design includes a linking element m with one common item $i = 1$, the linking parameters are defined as

$$u = \frac{a_{1_1}}{a_{1_2}},\ a_{1_2} > 0, \tag{4.8}$$

$$v = b_{1_2} - ub_{1_1}. \tag{4.9}$$

The expressions for the asymptotic standard error (ASE) for the estimate of u and v associated with this design are

$$\sigma_{u_m} = \left(\frac{\sigma^2_{a_{1_1}} + u^2\sigma^2_{a_{1_2}}}{a^2_{1_2}} \right)^{1/2}, \tag{4.10}$$

with $a_{1_2} > 0$ and

$$\sigma_{v_m} = \left\{ \frac{b^2_{1_1}\left(\sigma^2_{a_{1_1}} + u^2\sigma^2_{a_{1_2}}\right) + 2a_{1_1}b_{1_1}\left(\sigma_{a_{1_1}b_{1_1}} + \sigma_{a_{1_2}b_{1_2}}\right) + a^2_{1_2}\left(u^2\sigma^2_{b_{1_1}} + \sigma^2_{b_{1_2}}\right)}{a^2_{1_2}} \right\}^{1/2}, \tag{4.11}$$

with $a_{1_2} > 0$, where the σs are the elements of the covariance matrices of the item parameter estimators for the common item in the two calibrations.

4.2.2 One Pair of Common Items

Alternatively, when the design includes a linking element m of one pair of common items, $i = 1$ and $i = 2$, substitution of the difficulty parameters into (4.7) and elimination of v gives us the linking parameters as functions of the difficulty parameters only:

$$u = \frac{b_{1_2} - b_{2_2}}{b_{1_1} - b_{2_1}},\ b_{1_1} \neq b_{2_1} \tag{4.12}$$

$$v = \frac{1}{2}\sum_{i=1}^{2}(b_{i_2} - ub_{i_1}). \tag{4.13}$$

The expressions for the ASEs are now

$$\sigma_{u_m} = \left[u^2 \left(\frac{\sigma^2_{b_{1_1}} + \sigma^2_{b_{2_1}}}{(b_{1_1} - b_{2_1})^2} + \frac{\sigma^2_{b_{1_2}} + \sigma^2_{b_{2_2}}}{(b_{1_2} - b_{2_2})^2} \right) \right]^{1/2}, \tag{4.14}$$

with $b_{1_1} \neq b_{2_1}, b_{1_2} \neq b_{2_2}$ and

$$\sigma_{v_m} = \left[\frac{\left(b^2_{2_1}\sigma^2_{b_{1_1}} + b^2_{1_1}\sigma^2_{b_{2_1}} \right)(b_{1_2} + b_{2_2})^2 + \left(b^2_{2_1}\sigma^2_{b_{1_2}} + b^2_{1_1}\sigma^2_{b_{2_2}} \right)(b_{1_1} - b_{2_1})^2}{(b_{1_1} - b_{2_1})^4} \right]^{1/2}, \tag{4.15}$$

with $b_{1_1} \neq b_{2_1}$. Note that we could have estimated v separately for each of the two items using (4.9). However, as both items are informative it makes sense to combine them through estimation of the average in (4.13). The standard error of v in (4.15) allows for the fact that the contributions by the two items share the common estimate of u.

4.2.3 One Pair of Common Test-Takers

Analogous to (4.12) and (4.13), when the linking element m includes common test-takers $p = 1$ and $p = 2$ the linking parameters may be calculated as

$$u = \frac{\theta_{1_2} - \theta_{2_2}}{\theta_{1_1} - \theta_{2_1}}, \ \theta_{1_1} \neq \theta_{2_1}, \tag{4.16}$$

$$v = \frac{1}{2}\sum_{p=1}^{2}(\theta_{p_2} - u\theta_{p_1}). \tag{4.17}$$

The standard errors are analogous as well:

$$\sigma_{u_m} = \left[u^2 \left(\frac{\sigma^2_{\theta_{1_1}} + \sigma^2_{\theta_{2_1}}}{(\theta_{1_1} - \theta_{2_1})^2} + \frac{\sigma^2_{\theta_{1_2}} + \sigma^2_{\theta_{2_2}}}{(\theta_{1_2} - \theta_{2_2})^2} \right) \right]^{1/2}, \tag{4.18}$$

with $\theta_{1_1} - \theta_{2_1} \neq 0$ and $\theta_{1_2} - \theta_{2_2} \neq 0$ and

$$\sigma_{v_m} = \left[\frac{\left(\theta^2_{2_1}\sigma^2_{\theta_{1_1}} + \theta^2_{1_1}\sigma^2_{\theta_{2_1}} \right)(\theta_{1_2} + \theta_{2_2})^2 + \left(\theta^2_{2_1}\sigma^2_{\theta_{1_2}} + \theta^2_{1_1}\sigma^2_{\theta_{2_2}} \right)(\theta_{1_1} - \theta_{2_1})^2}{(\theta_{1_1} - \theta_{2_1})^4} \right]^{1/2}, \tag{4.19}$$

with $\theta_{1_1} - \theta_{2_1} \neq 0$.

4.2.4 Multiple Common Items or Test-Takers

Let $m = 1, ..., M$ denote the linking elements (common items; common pairs of items/test-takers) added to a linking design. The linking parameters may now be estimated pooling the estimators for each of the $m = 1, ..., M$ linking elements as a precision-weighted average. For the estimation of u, the precision-weighted average is

$$\widehat{u} = \left(\sum_{m=1}^{M} \frac{\widehat{u}_m}{\widehat{\sigma}^2_{u_m}} \right) \Bigg/ \left(\sum_{m=1}^{M} \frac{1}{\widehat{\sigma}^2_{u_m}} \right), \tag{4.20}$$

with estimated standard error

$$\widehat{\sigma}_u = 1 \Bigg/ \left(\sum_{m=1}^{M} \frac{1}{\widehat{\sigma}^2_{u_m}} \right)^{1/2}. \tag{4.21}$$

The estimator and standard error for v are analogous.

As is immediately clear from their expressions, these standard errors of u and v decrease monotonically as linking elements are added to the design. At the same time, thanks to (4.20), contributions to the overall estimates by elements with higher linking error are down-weighted. We also expect designs with multiple elements to be more robust to systematic errors.

As systematically found in our earlier research, the design in (4.8)–(4.9) outperforms the one in (4.12)–(4.13). That is, for equal numbers of common items, it appears difficult to attain the same linking precision for the latter as for the former (Barrett & van der Linden, 2015a). This result is not surprising for two different reasons. First, estimates of (4.12)–(4.13) are based on the item difficulty parameters only; unlike (4.8)–(4.9), they ignore the unique information about the linking parameters in the discrimination parameters of the common items. Second, the ASEs in (4.14)–(4.15) have the expressions $b_{1_1} - b_{2_1}$ and $b_{1_2} - b_{2_2}$ in their denominators, which become unstable when the item difficulties approach each other. Although (4.20) effectively reduces the impact of pairs with minimal differences on the overall estimate, their standard error in (4.21) will become large due to the presence of these differences in the denominator.

4.3 Optimal Design

When planning for linking operationally, a linking design must be chosen. As noted earlier, it is more common to link through sets of items than test-takers. Therefore, the focus of the

remainder of this paper will be on common-item designs. In particular, for the purpose of our examples, we will use the case of linking based on the precision-weighting of estimates from elements of single common items in (4.8)–(4.9) according to (4.20)–(4.21). In practice, the set of linking items is often selected from the items present in the first administration, but other items may be available in the item pool that were already linked to those in the first calibration through a different chain of linking as well. The choice of linking items is non-trivial, involving many test requirements that must be met, both for the entire new test form and its subset of linking items. The complete set of requirements is typically multivariate, complicated, and nuanced. Often, the subset of linking items must proportionately represent the content requirements of the entire form to minimize the chance of systematic errors in the linking. Also, the subset is usually included when the information function of the new form has to be optimized with respect to a target.

Optimal test design is now widely used to solve severely constrained operational test assembly problems for both fixed forms (same selection of items administered to all test-takers) and adaptive tests (linear on the fly, multi-stage on the fly, or item-level adaptive). The optimal test design framework (van der Linden, 2005) provides a method in which a set of items from an item pool is selected to optimize a statistical objective subject to all test requirements. In this approach, both the objective and constraints on the test form are formulated as a mathematical mixed integer programming (MIP) model which is then solved by a powerful optimization solver. Each item in the item pool is represented with a binary decision variable with values 1 (include) and 0 (do not include) and both the constraints and the objective are built with these variables. For example, the requirement to include between three and five items associated with the algebra sub-content area on a mathematics test would be represented with a summation constraint, in which the sum of the decision variables of all items in the pool designated as algebra items would be bounded to be greater than or equal to three and less than or equal to five. Many types of categorical, quantitative, and logical constraints are possible under this framework; see van der Linden (2005) for a more complete description of them. Objective functions are represented as the sum across all items in the pool of their contribution to the statistical objective (e.g., test information function) multiplied by their decision variable. The impact of estimation error in the item parameters to the objective function was studied by Veldkamp (2013) and Veldkamp, Matteucci, and de Jong (2013). However, as the items in our examples were calibrated from large samples of test-takers, it was not necessary us to account for parameter estimation error in the objective functions. Multiple optimization solvers are available, including *CPLEX* (International Business Machine Corporation, 2009), *Xpress Optimizer* (FICO, 2009), and the open source *lp_solve* (Berkelaar, Eikland & Notebaert, 2004).

Diao & van der Linden (2011) present some examples representing common test assembly applications using *lp_solve* with *lpSolveAPI* in *R* (Konis, 2013).

Our decision variables will be represented as $x_m = 0, 1$, where $m = 1, ...M$ are the candidate common items for the new test form. Use of the optimal test design framework for optimal linking design is greatly facilitated if the overall linking error is additive in the contribution of each linking item. Fortunately, this feature does hold for the ASE of the precision-weighted average; its expression in (4.21) is minimal when

$$\sum_{m=1}^{M} \frac{1}{\sigma^2_{u_m}} \tag{4.22}$$

is maximal, so we are able to use $\sigma^{-2}_{u_m}$ as weights for the decision variables in our model.

From this key observation, we are provided with options for objective functions that minimize linking error or constraints that bound it at reasonable values. For example, we may wish to select the set of a fixed number of items which will minimize the error in $\widehat{u}$, minimize the error in $\widehat{v}$, or minimize a weighted composite of the error in $\widehat{u}$ and $\widehat{v}$. Alternatively, we may wish to minimize the number of linking items subject to a constraint that sets an upper bound on the linking error of $\widehat{u}$ and/or $\widehat{v}$. Finally, we may wish to select the entire second form, optimized to a target test information function, while simultaneously selecting the linking items subject to an upper bound on the linking error of $\widehat{u}$ and/or $\widehat{v}$.

4.3.1 Optimal Linking Design Models

Our first model is for the selection of m common items:

$$\max \sum_{m=1}^{M} \left(\omega_u \sigma^{-2}_{u_m} + \omega_v \sigma^{-2}_{v_m}\right) x_m, \quad \text{(objective function)} \tag{4.23}$$

subject to

$$\sum_{m=1}^{M} x_m = \lambda_m \quad \text{(number of common elements)} \tag{4.24}$$

$$\sum_{m=1}^{M} q_m x_m \lessgtr \lambda_q, \quad \text{all } q, \quad \text{(quantitative constraints)} \tag{4.25}$$

$$\sum_{m \in V_c^{\text{item}}} x_m \lessgtr \lambda_c, \quad \text{all } c, \quad \text{(categorical constraints)} \tag{4.26}$$

$$x_m = 0, 1, \quad \text{all } m, \quad \text{(decision variables)} \tag{4.27}$$

where ω_u and ω_v are weights for the relative importance of the two parameters, all items have quantitative attributes q_m and the items in set V_c^{item} share categorical attribute c, while λ_m, λ_q and λ_c are bounds set by the test assembler.

Alternatively, we can constrain one of the standard errors (σ_u, say) and minimize the other. We then replace (4.23) by

$$\max \sum_{m=1}^{M} \sigma_{u_m}^{-2} x_m, \quad \text{(objective function)} \tag{4.28}$$

and add

$$\sum_{m=1}^{M} \sigma_{v_m}^{-2} x_m \geq \lambda_v^{-2}, \quad \text{(upper bound on } \sigma_v\text{)} \tag{4.29}$$

to the constraint set. The addition guarantees a standard error $\sigma_v \leq \lambda_v$, with λ_v a well-chosen constant; see (4.21).

Finally, if the entire test form for the second test administration is assembled using the current type of optimization modeling, it may make sense relax the content constraints in (4.25)–(4.26) by having their sums run over the entire pool of common and non-common items available for the test form (rather than having separate sets of constraints for both of them). Let $j = 1, ..., J$, denote the available non-common items. The key part of the constraint set now becomes

$$\sum_{m=1}^{M} \sigma_{u_m}^{-2} x_m \geq \lambda_u^{-2}, \quad \text{(upper bound on } \sigma_u\text{)} \ , \tag{4.30}$$

$$\sum_{m=1}^{M} \sigma_{u_m}^{-2} x_m \geq \lambda_v^{-2}, \quad \text{(upper bound on } \sigma_v\text{)} \ , \tag{4.31}$$

$$\sum_{j=1}^{J} x_i + \sum_{m=1}^{M} x_m = \lambda_a \quad \text{(number of total elements)} \ , \tag{4.32}$$

$$\sum_{m=1}^{M} x_m = \lambda_m \quad \text{(number of common elements)} \ , \tag{4.33}$$

$$\sum_{j=1}^{J} q_j x_j + \sum_{m=1}^{M} q_m x_m \lessgtr \lambda_q, \quad \text{all } q, \quad \text{(quantitative constraints)} \ , \tag{4.34}$$

$$\sum_{j \in V_c^{\text{item}}} x_j + \sum_{m \in V_c^{\text{item}}} x_m \lessgtr \lambda_c, \quad \text{all } c, \quad \text{(categorical constraints)} \ , \tag{4.35}$$

$$x_j = 0, 1, \quad \text{all } j, \quad \text{(decision variables)} \ , \tag{4.36}$$

$$x_m = 0, 1, \quad \text{all } m, \quad \text{(decision variables)} \ . \tag{4.37}$$

The objective function can now be chosen to maximize the test information, for instance, one that maximizes a common lower bound, y, at three well-chosen θ values:

$$\text{maximize } y \tag{4.38}$$

subject to

$$\sum_{j=1}^{J} I_j(\theta_r)x_j + \sum_{m=1}^{M} I_m(\theta_r)x_m \geq y, \text{ for } r = 1, 2, 3. \tag{4.39}$$

4.3.2 Missing Values

The only missing values in the objective functions and/or constraints are those for σ_{u_m} and σ_{v_m}. Although the expressions for the ASEs in (4.10)–(4.11) now have known item parameter values for the first calibration, those for the second test administration have not yet been obtained. Separating the two types of values and using (4.8), we can write their expressions as

$$\begin{aligned} \sigma_{u_m} &= \left(\frac{\sigma^2_{a_{i_1}} + u^2\sigma^2_{a_{i_2}}}{a^2_{i_2}} \right)^{1/2} \\ &= \left(\sigma^2_{a_{i_1}}\zeta^{(1)}_m + a^2_{i_1}\zeta^{(2)}_m \right)^{1/2} \end{aligned} \tag{4.40}$$

and

$$\begin{aligned} \sigma_{v_m} &= \left\{ \frac{b^2_{i_1}\left(\sigma^2_{a_{i_1}} + u^2\sigma^2_{a_{i_2}}\right) + 2a_{i_1}b_{i_1}\left(\sigma_{a_{i_1}b_{i_1}} + \sigma_{a_{i_2}b_{i_2}}\right) + a^2_{i_2}\left(u^2\sigma^2_{b_{i_1}} + \sigma^2_{b_{i_2}}\right)}{a^2_{i_2}} \right\}^{1/2} \\ &= \left\{ [(b_{i_1}\sigma_{a_{i_1}})^2 + (a_{i_1}\sigma_{b_{i_1}})^2 + 2a_{i_1}b_{i_1}\sigma_{a_{i_1}b_{i_1}}]\zeta^{(1)}_m + (a_{i_1}b_{i_1})^2\zeta^{(2)}_m + 2a_{i_1}b_{i_1}\zeta^{(3)}_m + \zeta^{(4)}_m \right\}^{1/2} \end{aligned} \tag{4.41}$$

where

$$\zeta^{(1)}_m = \frac{1}{a^2_{i_2}} \tag{4.42}$$

$$\zeta^{(2)}_m = \left(\frac{\sigma_{a_{i_2}}}{a^2_{i_2}} \right)^2 \tag{4.43}$$

$$\zeta^{(3)}_m = \frac{\sigma_{a_{i_2}b_{i_2}}}{a^2_{i_2}} \tag{4.44}$$

$$\zeta^{(4)}_m = \sigma^2_{b_{i_2}}. \tag{4.45}$$

Because we selected the linking design with a minimal linking element of a single common item, $m = i$ in (4.40)–(4.45).

The question of how best to establish reasonable choices of values for the ζs arises. Perhaps the most simple and flexible approach is to conduct a simulation study, in which simulated test-takers are pulled from a distribution similar to that anticipated in the second administration. For example, if the test-takers are expected to be more able, they should be sampled from an ability distribution shifted to the right relative to the first administration. As the probabilities of success on the items are invariant under linking, responses can be generated on the candidate common items for the ability levels of the sampled test-takers using the parameters from their first administration. The responses can then be used to estimate the item parameters and their standard errors of estimation for the anticipated new distribution of test-takers, which allow us to calculate the ζs in (4.42)–(4.45).

Simulation is not the only way to choose values for the ζs. Alternatively, if the samples have comparable sizes and shapes of distributions, it may be possible to simply use the same estimates and standard errors twice. Simulation, however, provides a great deal of flexibility if the second set of test-takers is expected to be different from the first.

Finally, it is important to observe that the purpose of the ζs is to support an optimal design study only. Once administered to the second set of test-takers, item parameter estimates calculated from the actual responses will be used to estimates linking parameters u and v. Misspecification of the ζs during the optimal linking design study only leads to less than optimal standard errors for these estimates, *not* to any bias in them.

4.4 Empirical Examples

We first start with a set of examples that illustrate the use of the model described in (4.23)–(4.27) for a variety of $\zeta_m^{(1)} - \zeta_m^{(4)}$ values and weights ω_u and ω_v. This model minimizes linking error given the selection of a fixed number of linking items. We then continue with examples for the model in (4.28)–(4.39) constraining linking error by carefully chosen upper bounds. All models were solved using *lp_solve* 5.5.2.0 (Berkelaar, Eikland & Notebaert, 2004) with *lpSolveAPI* for *R* (Konis, 2013). All item calibrations were performed using *MIRT v1.0* (Glas, 2010).

4.4.1 Example 1. Minimizing Linking Error Given a Fixed Number of Linking Items for a High-School Mathematics Exam

The first example uses data from a large-scale high-school mathematics exam. In the first administration, responses to 44 selected-response items were gathered from 36,939 test-takers. The data were used to calibrate the items to the 3PL model. The exam also included constructed-response items and field-test items, which were excluded from this study.

Our linking problem asked us to identify which of the 44 items to include as linking items in the new test form to minimize linking error, while meeting a variety of test requirements established by the subject matter experts and psychometricians who worked on the operational program. To limit item exposure, no more than 15 items could be selected to serve as linking items. To limit the chance of systematic error or bias, the linking items also had to represent proportionately the full test blueprint, have similar numbers of items with each of the answer-key choices, and be distributed similarly across the positions they had in the test form during their first administration. In addition, they had to perform well psychometrically, with reasonable discriminating power and difficulty and low rates of omission ($< 5\%$) across test-takers (by policy an item left unanswered by a test-taker is treated as incorrect in this dataset; however, items frequently left unanswered are not desired in the linking set). Each of these requirements could be represented using fairly simple expressions in our decision variables $x_m = 0, 1$, $m = 1, ..., M$. All constraints were categorical, using the item attributes in Table 4.1 to establish membership of the sets V_c^{item} in (4.26). The same table shows the bounds used on each of these sets.

In order to establish our objective function, we needed reasonable values for the missing parameters $\zeta_m^{(1)}$–$\zeta_m^{(4)}$, $m = 1, ..., M$. We used the approach described earlier and simulated test-taker responses on each of the 44 items, and then estimated their parameters and covariance matrices from the simulated responses to establish $\zeta_m^{(1)}$–$\zeta_m^{(4)}$. To illustrate the impact of different choices of the test-taker ability distribution for the simulation on the selection of the linking items, we implemented the procedure for seven different distributions, including those that were generally higher, lower, more skewed, or had variances different from the first administration. Table 4.2 lists all distributions used for the simulations. Responses were simulated for the same number of test-takers as tested in the first administration, as this number was known to hold roughly throughout this testing program. The pattern of item calibration error across the seven simulations followed our expectations. For example, estimation error in the parameters was generally

Table 4.1. *High-school mathematics exam, linking item constraints (Examples 1, 3, and 4)*

V_c^{item}	Lower Bound λ_c	Upper Bound λ_c
All items	15	15
Items with an item type other than selected response	0	0
Items with a use status other than 'Yes' (e.g., do not use)	0	0
Items in positions 1 through 20 in the first administration	5	10
Items in positions 21 through 40 in the first administration	5	7
Items in positions 41 through 60 in the first administration	5	7
Items with answer key = 1	3	6
Items with answer key = 2	3	12
Items with answer key = 3	3	6
Items with answer key = 4	3	6
Items flagged due to poor item fit	0	0
Items flagged due to high omit rates	0	0
Items with p-values under 0.25	0	2
Items with p-values over 0.9	0	0
Items with point biserial correlation under 0.2	0	3
Items with a point biserial over 0.05 on a distractor	0	0
Items that cannot be Brailled	0	2
Items from sub-content area Number Sense and Computation	5	8
Items from sub-content area Algebra	4	5
Items from sub-content area Statistics and Probability	1	5
Items from sub-content area Geometry	4	5

Table 4.2. *Simulated test-taker ability distributions for $t = 2$ (Examples 1 and 2)*

Case	Simulated Test-Taker Ability Distribution
1	N(0,1)
2	N(1,1)
3	N(-1,1)
4	N(0,2)
5	N(0,0.5)
6	Generalized Beta(3,5), linearly transformed $x = 8z - 4$
7	Generalized Beta(5,3), linearly transformed $x = 8z - 4$

higher in Case 5, where ability variance was low and consequently the responses were less informative about the item parameters for most of the items, with the exception of those whose difficulty was close to the mean of the ability distribution.

We then modeled the assembly problem for each of the seven ability distributions exploring the effects of three different sets of weights ($\omega_u = 1, \omega_v = 1$; $\omega_u = 2, \omega_v = 1$; $\omega_u = 1, \omega_v = 2$). The constraints were the same in all models, but the objective function in (4.23) changed as a function of the ζs calculated from the item parameter estimates and covariance matrices for

the different simulations and weighting scenarios. An example of the model file (.lp file) for *lp_solve* 5.5.2.0 for the first case is included in Appendix A for reference.

Figure 4.1 shows the ASEs anticipated for the linking parameter estimates associated with the selected linking items for each of the 21 scenarios (seven simulated ability distributions, each with three weighting scenarios). Differences were minimal across the three weighting scenarios, but the linking error followed our expectations as to the relatively higher and lower variances in parameter estimates based on the choice of expected ability distribution. Figure 4.2 illustrates the percentages with which each of the 44 available items appeared in the 21 different solutions for the given item pool and constraint set. Nine of the 44 items appeared in 100% of the solutions, while over 20 of them were never selected. The remaining items figured in different numbers of selections based on their contribution to linking error in the objective function.

Again, note that using the item selection for one of our scenarios over the others will not lead to any bias in the linking parameter estimates to be derived from the actual responses by the test-takers in the second test administration. Rather, the solutions help us to identify the set of common items most likely to have minimal linking error based on our best guess of the ability distribution prior to the second administration.

4.4.2 Example 2. Minimizing Linking Error Given a Fixed Number of Linking Items for a High-School Reading Exam

Our second example provides a similar illustration, this time with a large-scale high-school reading exam. Reading tests tend to be more tightly constrained than mathematics tests due to extra requirements related to the reading passages shared by their items. This exam included 56 items spread across 17 passages administered to 32,931 test-takers, all calibrated with the 3PL model. Fifty-three of the items became eligible for selection as linking items, as not all meta data fields for the item pool were sufficiently populated for three of the items. The optimal linking design was to contain 20-25 linking items. The constraints that had to be used are provided in Table 4.3. Note that this time, in addition to decision variables $x_m = 0, 1$, $m = 1, ..., M$, we needed decision variables $x_s = 0, 1$, $s = 1, ..., S$ for the inclusion of the shared reading passages in the test, and V_c^{stim} to represent the sets of passages with the attributes present in the test requirements.

Simulated responses were again generated with test-takers sampled from the ability distributions in Table 4.2, and models were built to represent the objective functions based on the parameter

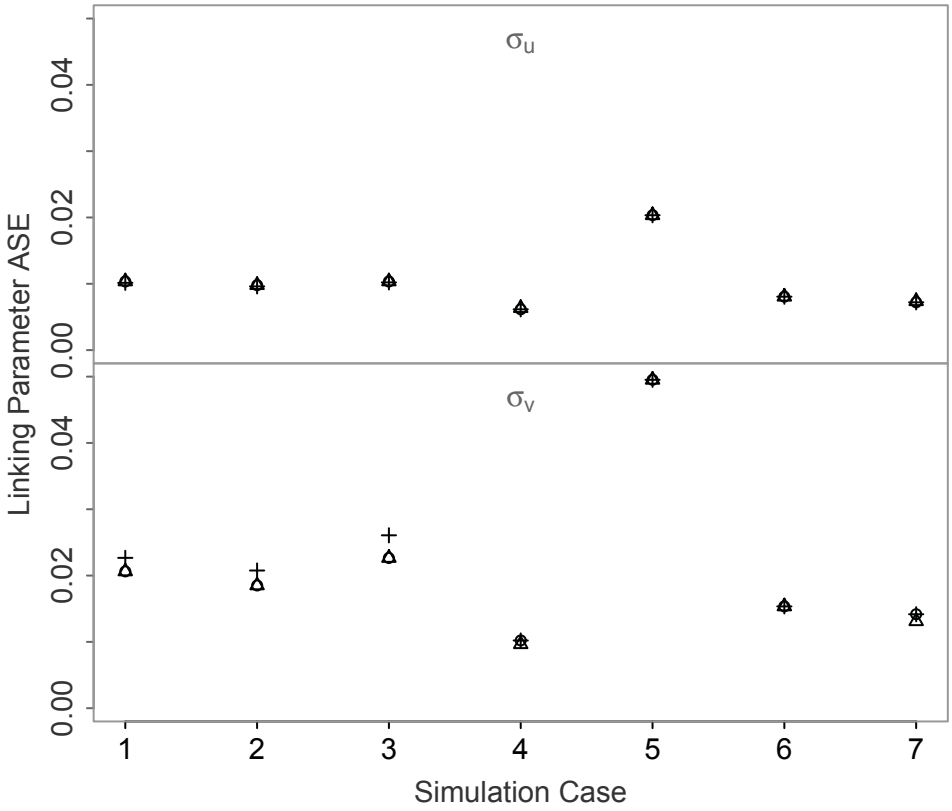

Figure 4.1. ASE of optimal linking design solutions on high-school mathematics exam for each of the simulated ability distributions for second administration (Cases 1-7) and weighting scenarios ($\omega_u = 1, \omega_v = 1$ (circle); $\omega_u = 1, \omega_v = 2$ (triangle); $\omega_u = 2, \omega_v = 1$ (cross)) in Example 1. Note. The distinction among the different weighting scenarios is hardly discernible in all cases for σ_u and in cases 4-7 for σ_v.

estimates and covariance matrices obtained from the responses and the three different weighting schemes. Figure 3 shows the standard error of the linking parameters calculated for each of the resulting selections of linking items. Again, little difference is seen among the three weighting scenarios and the standard errors follow expectation based on the simulated responses. However, there is less variation between these standard errors among the different scenarios than was observed in the mathematics example. Figure 4 presents the percentages of the 21 different solutions in which each of the 53 items and 17 reading passage appeared. Note that all of the 21 solutions included the same set of five passages. The same held for 15 items, with the additional six items varying by solution. Although the maximum number allowed by bound $\lambda_m = 25$ would have resulted in the least amount of linking error, the combination of other constraints and the attributes of the available items prevented the total number of items selected from exceeding 21.

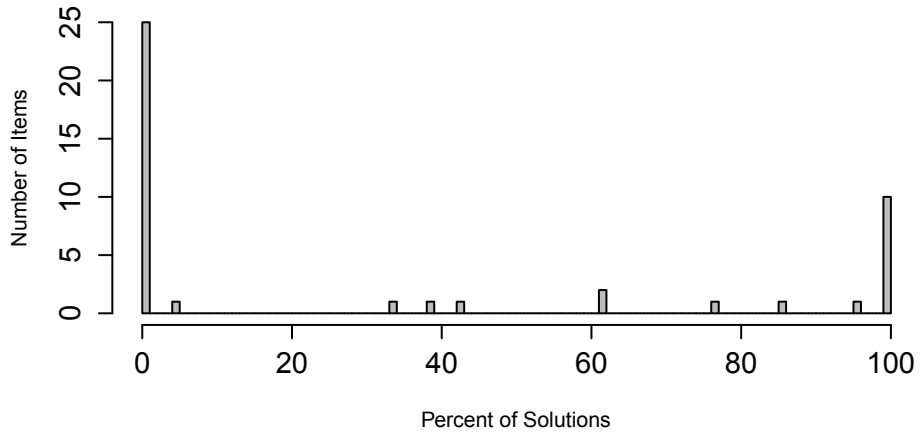

Figure 4.2. Percent of the high-school mathematics exam optimal linking design solutions in which each of the 44 available items appears in Example 1.

Table 4.3. *High-school reading exam, linking item constraints (Example 2)*

V_c^{item}	Lower Bound λ_c	Upper Bound λ_c
All items	20	25
Items with an item type other than selected response	0	0
Items with a use status other than 'Yes' (e.g., do not use)	0	0
Items flagged due to poor item fit	0	0
Items flagged due to high omit rates	0	5
Items with p-values under 0.25	0	6
Items with p-values over 0.9	0	6
Items with point biserial correlation under 0.2	0	7
Items with a point biserial over 0.05 on a distractor	0	0
Items that cannot be brailled	0	2
Items not to be used due to previous use history	0	0
Items from sub-content area Reading Comprehension	6	8
Items from sub-content area Use of Critical Thinking Skills	4	5
Items from sub-content area Use of Evidence	4	5
Items from sub-content area Literature	4	5
Items associated with any one reading passage	3	8
V_c^{stim}	Lower Bound λ_c	Upper Bound λ_c
All reading passages	3	9
Poem reading passage	1	3
Fiction reading passage	1	3
Non-fiction reading passage	1	3
No passage	0	0

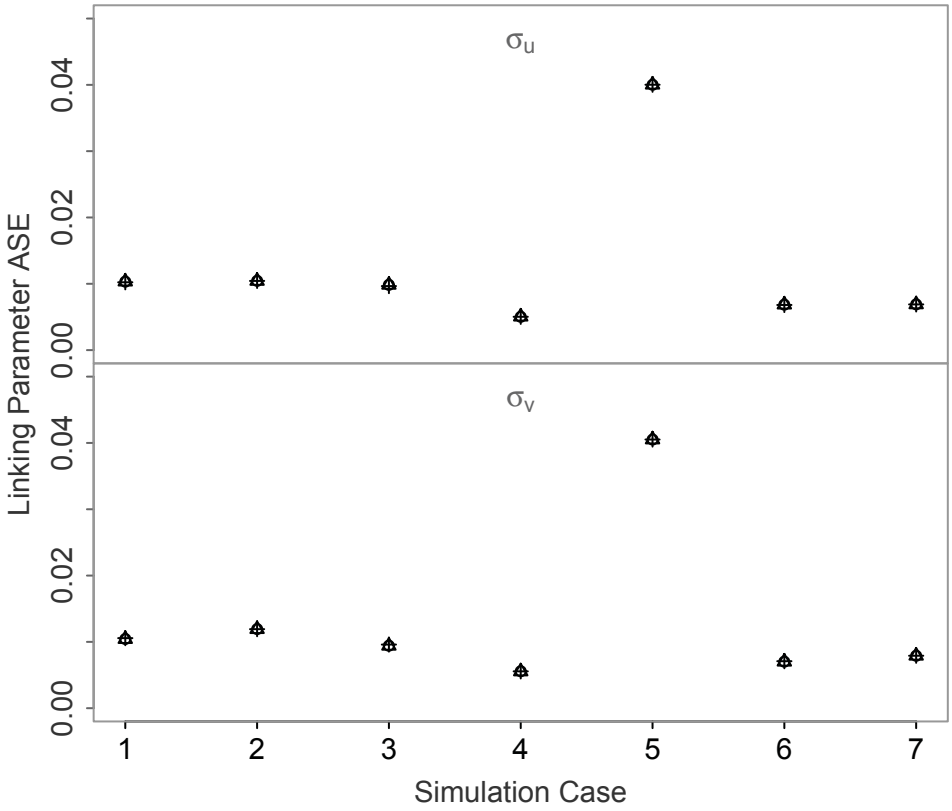

Figure 4.3. ASE of optimal linking design solutions on high-school mathematics exam for each of the simulated ability distributions for second administration (Cases 1-7) and weighting scenarios ($\omega_u = 1, \omega_v = 1$ (circle); $\omega_u = 1, \omega_v = 2$ (triangle); $\omega_u = 2, \omega_v = 1$ (cross)) in Example 2. Note. The distinction among the different weighting scenarios is hardly discernible in all cases.

4.4.3 Example 3. Minimizing σ_u while constraining σ_v for a High-School Mathematics Exam

Our next example illustrates variations of the model in (4.28)–(4.29). The objective of the models was to minimize error in one of the linking parameters (e.g., σ_u) while constraining the error in the other parameter to be below an upper bound (e.g., λ_v). For each of these models, it was necessary to consider the value of λ for the upper bound on the error we might find reasonable for the linking parameters given the items available for linking and the purpose of the test.

We first consider the case of minimizing σ_u while constraining σ_v for the mathematics exam described above. With 44 available linking items, as linking error is monotonically decreasing with the number of linking items, minimum possible error would result from the inclusion of

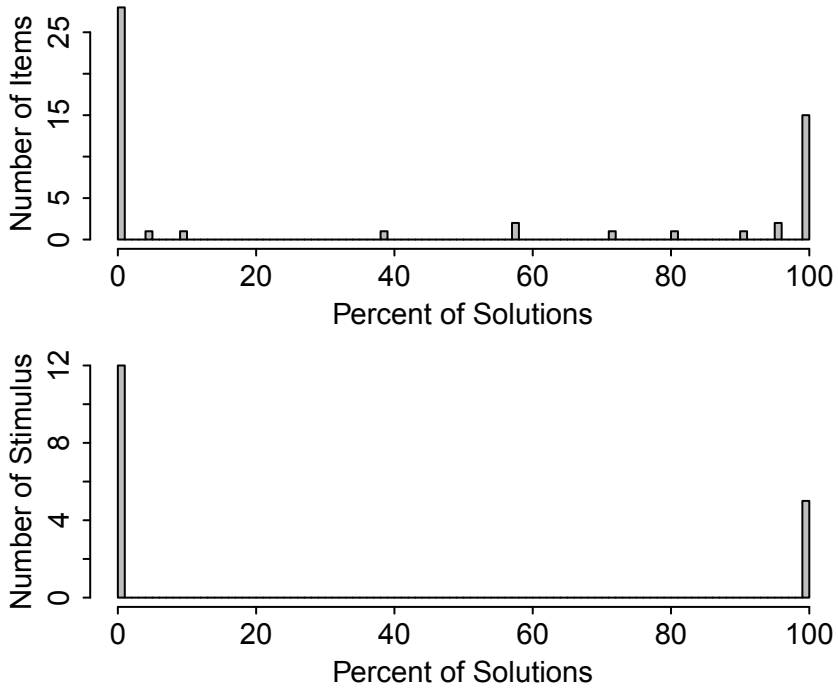

Figure 4.4. Percent of the high-school mathematics exam optimal linking design solutions in which each of the 53 available items and 17 available shared stimulus appear in Example 2.

all of them; however we constrained the solution to include 15 items as was desired by the practitioners assembling the test forms. When a simulation of the second administration was conducted with an $N(0,1)$ distribution to obtain $\zeta_m^{(1)}$–$\zeta_m^{(4)}$, Table 4.4 shows the results of this model for a variety of λ_v to constrain σ_v. As expected, for this data, σ_u of this solution increases as the upper bound on σ_v is decreased, and both σ_u and σ_v of the solution increase when additional constraints from Table 4.1 are added to the model.

As illustrated in this example, the interplay among constraints was generally complicated and iteration with a few varieties of the model is generally recommended to establish reasonable λ_u or λ_v. Observe that without any constraints other than $\lambda_m = 15$, the minimum obtainable $\sigma_u = 0.00769$; for this item pool, there was no solution with 15 items that could produce a lower σ_u. Additional constraints as in Table 4.1 increased the minimum obtainable σ_u to 0.01007 (without a constraint on σ_v). When σ_v was then constrained, a solution was possible for $\lambda_v > 0.01400$ without other constraints and a solution was possible for $\lambda_v > 0.01900$ with the constraints in Table 4.1. If the linking error of $\sigma_u = 0.01007$ or $\sigma_v = 0.01900$ was unacceptable, the constraints in Table 4.1 would need to be relaxed; however, not imposing any

Table 4.4. *Minimizing σ_u while constraining σ_v ($m = 15$) for a high-school mathematics exam (Example 3)*

	No additional constraints			Including all constraints in Table 4.1		
λ_v	Status	σ_u	σ_v	Status	σ_u	σ_v
–	F	0.00769	0.01402	F	0.01007	0.02127
0.02200	F	0.00769	0.01402	F	0.01007	0.02127
0.02100	F	0.00769	0.01402	F	0.01028	0.01955
0.02000	F	0.00769	0.01402	F	0.01028	0.01955
0.01900	F	0.00769	0.01402	I	–	–
0.01800	F	0.00769	0.01402	I	–	–
0.01700	F	0.00769	0.01402	I	–	–
0.01600	F	0.00769	0.01402	I	–	–
0.01500	F	0.00769	0.01402	I	–	–
0.01400	F	0.00787	0.01396	I	–	–
0.01300	I	–	–	I	–	–

Note. F=feasible, I=infeasible.

other constraints, they could not be set lower than the minimally obtainable σ_u or σ_v and still maintain feasibility.

4.4.4 Example 4. Minimizing m while constraining σ_u and σ_v for a High-School Mathematics Exam

Consider now the scenario where we want to minimize the number of linking items m subject to constraints on the linking error for one or both of the linking parameters. In this case, the objective in (4.28) is replaced with

$$\min \sum_{m=1}^{M} x_m \quad \text{(objective function)} \tag{4.46}$$

and a constraint for σ_u analogous to (4.29) is included. The steps described above will support the effort to build this model when determining reasonable values of λ_u and λ_v. This model may be desired when item exposure from one administration to the next is a concern, but there is an acceptable upper bound on the linking errors. Results for this model with a variety of λ_u and λ_v are presented in Table 4.5.

Note that the lowest linking error with a feasible solution was $\sigma_u = 0.0069$ and $\sigma_v = 0.00130$, which occured with $m = 44$ items and no other constraints. This was expected as linking error decreases with increases in m and 44 was the maximum value of m possible with this item pool.

Table 4.5. *Minimizing mwhile constraining σ_u and σ_v for a high-school mathematics exam (Example 4)*

		No additional constraints				Including all constraints in Table 4.1			
λ_u	λ_v	Status	m	σ_u	σ_v	Status	m	σ_u	σ_v
0.0060	0.0100	I	–	–	–	I	–	–	–
0.0069	0.0130	F	44	0.0069	0.0130	I	–	–	–
0.0070	0.0140	F	30	0.0070	0.0130	I	–	–	–
0.0080	0.0150	F	13	0.0080	0.0148	I	–	–	–
0.0090	0.0160	F	10	0.0088	0.0157	I	–	–	–
0.0100	0.0170	F	8	0.0095	0.0165	F	18	0.0087	0.0170
0.0110	0.0180	F	7	0.0099	0.0173	F	17	0.0100	0.0177
0.0120	0.0190	F	6	0.0114	0.0181	F	16	0.0092	0.0186
0.0130	0.0200	F	5	0.0124	0.0192	F	15	0.0103	0.0196
0.0140	0.0210	F	4	0.0137	0.0203	F	15	0.0108	0.0209
0.0150	0.0220	F	4	0.0132	0.0212	F	15	0.0103	0.0206
0.0160	0.0230	F	3	0.0156	0.0225	F	15	0.0116	0.0220
0.0170	0.0240	F	3	0.0150	0.0236	F	15	0.0112	0.0233
0.0180	0.0250	F	3	0.0150	0.0236	F	15	0.0138	0.0239
0.0190	0.0260	F	3	0.0150	0.0236	F	15	0.0138	0.0239
0.0200	0.0270	F	2	0.0168	0.0270	F	15	0.0140	0.0239
0.0210	0.0280	F	2	0.0204	0.0254	F	15	0.0122	0.0210

Note. F=feasible, I=infeasible.

Accordingly, increases in the upper bounds λ_u and λ_v reduced m. Yet, when the constraints in Table 4.1 were included in the model, m did not reduce lower than 15, regardless of λ_u and λ_v, as constraints on numbers of items for each of the sub-content areas required at least 15 items. Analysis of this type may prove helpful to practitioners in its explicit exploration of trade-offs between linking item set size and linking error.

4.4.5 Example 5. Selecting an Entire Test Form with a Subset of Linking Items While Constraining σ_u and σ_v

Our final example illustrates the use of the model in (4.30)–(4.37). In contrast to the previous examples, this model selects an entire test form, including the subset of linking items. The model is effective in that it prevents a first selection of linking items from causing possible infeasibility of the overall selection due to the presence of constraints that run across all items in the test form.

For this example, we used an item pool associated with a high-school mathematics exam. There were 108 items in the pool, which had been administered in various combinations over

the past several years and were collected in the pool using multi-group concurrent calibration. As the total number of test-takers for the items them ranged from 5,000 to 20,000, parameter estimation error varied substantially across the items. Forty-five of the items in the pool were administered in the most recent test administration to 5,000 test-takers. They constituted the first test form of the two be linked. Obviously, less linking error was to be expected from items calibrated with higher numbers of responses and our optimal selection should favor these items; however, the selection was to be constrained by several additional requirements for the test. Simulation with a $N(0, 1)$ distribution with 5,000 test-takers per item was used to obtain $\zeta_m^{(1)}$–$\zeta_m^{(4)}$.

To establish the model, we first defined decision variables for the $M = 45$ items available as linking items and the $J = 63$ items available to serve as unique items. The constraints for the subset of the M (4.32) and total set of $M + J$ (4.33) items specified by the content experts and psychometricians for the program are indicated in Table 4.6. For this example, while constraining the standard errors of linking σ_u and σ_v, the objective maximized test information at $\theta = 0.5$. However, the reader will recognize that more advanced objective functions are possible, such as matching a target information curve, matching a target test characteristic curve, maximizing information at multiple cut scores, and so on. The *R* code for the MIP model in (4.30)–(4.37) is included in Appendix B. The reader should note, however, that this rendition of our code does not represent the most efficient use of *lpsolveAPI* possible; its primary goal is only to ease interpretation of it against the models used in this example.

Results for this model are included in Table 4.7. As in the previous example, we first established feasible bounds on linking error prior to entering all other constraints. In order to set a baseline for the maximum information possible given the items available in the pool, we first ran a model for which the objective was to maximize information at $\theta = 0.5$ subject only to the constraints that 15 linking items be selected from the M available and that a total of 45 items be selected from the M and J available. In this case, σ_u was equal to 0.0294 and σ_v was equal to 0.0658. As we decreased the bounds λ_u and λ_v, maximum information decreased, and there was no feasible solution once λ_u reached 0.0200 or λ_v reached 0.0320. Once all other constraints as in Table 4.6 were added to the model, we found that without any constraints on σ_u and σ_v, information was decreased by about 50%, and the σ_u of the solution was equal to 0.0402 and σ_v was equal to 0.0653. Only small reductions to these values in constraints on λ_u and λ_v were possible before the model had no feasible solution. If lower linking error or higher information should have been required for the purposes of the test, some of the constraints in Table 4.6 would have to be relaxed. Again, a benefit of using optimal test design is that trade-offs of this

Table 4.6. *High-school mathematics exam, content and psychometric constraints (Example 5)*

V_c^{item}	$J+M$	
	Lower Bound λ_c	Upper Bound λ_c
All items	45	45
Items with a use status other than 'Yes' (e.g., do not use)	0	0
Items from sub-content area Number Sense and Computation	14	24
Items from sub-content area Algebra	12	15
Items from sub-content area Statistics and Probability	3	15
Items from sub-content area Geometry	12	15
Items with answer key = 1	5	18
Items with answer key = 2	5	18
Items with answer key = 3	5	18
Items with answer key = 4	5	18
Items used 2 administrations prior		15
Items used 3 administrations prior		15
Items flagged due to poor item fit	0	0
Items that cannot be brailled	0	2
Items flagged due to high omit rates	0	0
Items with p-values under 0.25	0	6
Items with p-values over 0.9	0	6
Items with point biserial correlation under 0.1	0	0
Items with a point biserial over 0.05 on a distractor	0	4
	M only	
	Lower Bound λ_c	Upper Bound λ_c
All items	15	15
Items from sub-content area Number Sense and Computation	4	9
Items from sub-content area Algebra	3	6
Items from sub-content area Statistics and Probability	1	3
Items from sub-content area Geometry	3	6
Items in positions 1 through 20 in the prior administration	5	10
Items in positions 21 through 40 in the prior administration	5	7
Items in positions 41 through 60 in the prior administration	5	7

sort are made explicit and can be evaluated against their intended uses by those constructing the tests.

4.5 Concluding Remarks

This paper demonstrates that optimal linking design becomes possible and quite practical when the precision-weighted estimate of the linking parameters proposed earlier by its authors is used. As illustrated, the mixed integer programming models in the approach are straightforward

Table 4.7. *Maximizing information at* $\theta = 0.5$ *while constraining* σ_u *and* σ_v *(*$m = 15, a = 45$*) for a high-school mathematics exam (Example 5)*

		No additional constraints				Including all constraints in Table 4.6			
λ_u	λ_v	Status	$I(\theta = 0.5)$	σ_u	σ_v	Status	$I(\theta = 0.5)$	σ_u	σ_v
–	–	F	48.7933	0.0294	0.0658	F	24.2410	0.0402	0.0653
0.0400	0.0650	F	48.7817	0.0274	0.0593	F	23.8102	0.0376	0.0637
0.0300	0.0400	F	48.3583	0.0247	0.0400	I	–	–	–
0.0240	0.0340	F	46.8526	0.0222	0.0340	I	–	–	–
0.0230	0.0330	F	45.1368	0.0214	0.0327	I	–	–	–
0.0220	0.0330	F	45.1368	0.0214	0.0327	I	–	–	–
0.0220	0.0320	I	–	–	–	I	–	–	–
0.0210	0.0330	F	41.2416	0.0208	0.0324	I	–	–	–
0.0200	0.0330	I	–	–	–	I	–	–	–

Note. F=feasible, I=infeasible.

to build and do have an impact on linking item selection. Running them to identify the combination of items with the least amount of linking error subject to test constraints and prove optimality requires only milliseconds. On the other hand, it would be quite difficult for a human to evaluate all possible combinations of items meeting the constraints to find one that is most likely to lead to the least amount of random linking error.

In addition to providing a set of viable linking items, this method also provides an a priori estimate of the linking error given the test constraints, the item pool, and an assumption about the ability distribution of the test-takers in the second administration. This may be useful in practice. For example, if the stakes of the test are high, as in admission testing, the practitioner might ask whether the linking error is small enough to limit possible mis-classification of test-takers sufficiently. If not, the number of linking items may need to be allowed to increase or some of the test constraints may need to be relaxed to allow other items with smaller contributions to linking error in the selection.

That said, the impact of linking error on the estimates of the abilities of individual test-takers or functionals of their distributions remains to be researched and is reserved for our future work. A final evaluation of the practical usefulness of optimal linking design will depend on results from that research. Another limitation of the current study is its focus on dichotomous items only. Test blueprints often require representation of other item types as well, such as technology-enhanced or constructed-response items modeled with partial credit IRT models. Quantifying the individual contributions to linking error from alternative item types is necessary to fully explore optimal linking design for tests requiring such items.

Acknowledgment

The authors are indebted to Cees A. W. Glas for his support during their use of the *MIRT Scaling Program* (version 1.01) software.

References

Barrett, M. D., & van der Linden, W. J. (2015a). *Estimating linking functions for response model parameters*. Manuscript submitted for publication.

Berkelaar, M., Eikland, K., & Notebaert, P. (2004). *lp_solve, version 5.5.2.0*. [Computer software]. Retrieved from `http://lpsolve.sourceforge.net/5.5/`.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_Solve Version 5.5 in R. *Applied Psychological Measurement, 35*, 398–409.

FICO. (2009). *Xpress optimizer: Reference manual*. Warwickshire, UK: Author. Retrieved from `www.fico.com`.

Glas, C. A. W. (2010). *MIRT: Multidimensional item response theory, version 1.01*. [Computer software and manual]. Enschede, The Netherlands: University of Twente. Retrieved from `http://www.utwente.nl/gw/omd/en/employees/employees/glas.doc/`.

Haebara, T. (1980). Equating logistic ability scales by weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

International Business Machine Corporation. (2009). *IBM ILOG OPL, Version 6.3* [Software program and manuals]. Armonk, NY: Author.

Konis, K. (2013). *lpSolveAPI, version 5.5.2.0-9*. [Computer software]. Retrieved from `http://:CRAN.R-project.org/package=lpSolveAPI`.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

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. (2005). *Linear models for optimal test design*. New York: Springer.

van der Linden, W. J., & Barrett, M. D. (in press). Linking item response model parameters. *Psychometrika*. doi: 20.1007/s11336-015-9469-6.

van der Linden, W. J., & Diao, Q. (2011). Automated test form generation. *Journal of Educational Measurement, 50*, 249–285.

van der Linden, W. J., & Diao, Q. (2014). Using a universal shadow-test assembler with multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 108–118). Boca Raton, FL: Chapman & Hall/CRC.

Veldkamp, B. P. (2013). Application of robust optimization to automated test assembly. *Annals of Operations Research, 206*, 595–610.

Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement, 37*, 123–139.

Chapter 5

Linking Polytomous Response Model Parameters[1]

5.1 Introduction

Parameter linking in item response theory (IRT) is generally necessary to adjust for differences between the true values of the parameters estimated in different calibration studies due to their use of different identifiability restrictions. Lack of identifiability occurs when multiple parameter vectors yield an identical likelihood function. The problem of identifiability arises in IRT because of its explanation of the standard parameters of response distributions by a second level of item and ability parameters. For example, in common IRT models for dichotomous responses, although the success parameters of their binomial distribution are identified, they are described by a model that includes parameters for test-taker ability, item difficulty, item discrimination, and in some cases the probability of guessing correctly. It is easy to show that the same success parameters can result from different choices of values for these model parameters, and therefore additional restrictions need to be placed on the parameter space during item calibration to select one of the parameter vectors. In practice it is common to assume a distribution of the θ parameters that is normal with location and variance equal to $\mu_\theta = 0$ and $\sigma_\theta = 1$ and then use a two-stage estimation process with marginal maximum likelihood (MML) estimation of the item parameters and subsequent estimation of the ability parameters. Although we are not aware of any formal proof of the sufficiency of the procedure

[1]Barrett, M. D., & van der Linden, W. J. (2015). *Linking polytomous response model parameters.* Manuscript to be submitted for publication.

to identify the item parameters, the tradition of successful parameter recovery studies both for dichotomous and polytomous models suggest it is.

When comparison of the results of two different calibrations is desired, a linking study to map the set of parameters from the second calibration onto that for the first calibration is required. This is necessary even when the same restrictions $\mu_\theta = 0$ and $\sigma_\theta = 1$ are used in different calibrations as the empirical distributions of test-takers abilities are always different. In other words, formally identical restrictions are not empirically identical when used in different calibrations, and the use of empirically different identifiability restrictions results precisely in the linking problem addressed in this paper.

Based on the concept that linking is required to adjust for differences in identifiability restrictions applied during different calibrations, van der Linden and Barrett (in press) formally derived the linking functions for dichotomous unidimensional logistic item response models without any assumptions as to their shape. More generally, they also proved that the parameters in a response model that specifies the probabilities of a correct response as a monotone continuous function of all parameters are linked by a vector function with components for each parameter that are both monotone and continuous as well. Their work focused specifically on the 3PL model, but the results hold equally well for the 2PL and 1PL/Rasch models.

The 3PL model explains the probability of a correct response $U_{pi} = 1$ on items $i = 1, ..., I$, each with discrimination a_i, difficulty b_i, and lower asymptote c_i, for test-taker $p = 1, ..., P$ with ability $\theta_p \in \mathbb{R}$ as

$$\Pr\{U_{pi} = 1; \theta_p\} \equiv p(\theta_p; a_i, b_i, c_i) \equiv c_i + (1 - c_i)\{1 + \exp[-a_i(\theta_p - b_i)]\}^{-1}. \tag{5.1}$$

For this model, van der Linden and Barrett (in press) derived the linking function as the solution to the following functional equation in four unknowns:

$$\frac{\varphi_\gamma(\gamma)}{1 + \exp\left[\varphi_a(a)(\varphi_\theta(\theta) - \varphi_b(b))\right]} = \frac{\gamma}{1 + \exp[a(\theta - b)]}, \tag{5.2}$$

with transformation of $c = 1 - \gamma$. The solution was

$$\varphi_a(a) = u^{-1}a, \tag{5.3}$$

$$\varphi_b(b) = ub + v, \tag{5.4}$$

$$\varphi_c(c) = c, \tag{5.5}$$

and

$$\varphi_\theta(\theta) = u\theta + v, \tag{5.6}$$

with

$$u \equiv \frac{\varphi_\theta(\theta) - \varphi_b(b)}{\theta - b}, \theta \neq b, \tag{5.7}$$

and

$$v = \varphi_b(b) - ub = \varphi_\theta(\theta) - \theta. \tag{5.8}$$

The new approach was different from previous mean/sigma (Marco, 1977), mean/mean (Loyd & Hoover, 1980), item response function (Haebara, 1980), and test characteristic curve (Stocking & Lord, 1983) linking approaches in that it did not assume any linearity of the functions for each parameter, nor was it motivated by the common notion of a scale adjustment necessary only due to an indeterminate zero and unit for the θ scale. While the general shapes of the solution for the a, b, and θ parameters were identical to those assumed in the previous literature, the new definition of u as the difference between the test-taker's ability and the difficulty of the item in two calibrations provides flexibility to derive solutions for u and v from alternative linking designs. In addition, the derivation of the linking function for the c parameter as an identity function was a new result. Finally, solutions for u and v with linking elements comprised of one common item, a pair of common items, or pair of common test-takers allows the use of a new precision-weighted overall estimator of the linking parameters for an entire linking design. As demonstrated in Barrett and van der Linden (2015b), this approach is helpful as it can be used to choose a linking design with minimal propagation of item parameter estimation error into the linking function.

This paper extends the work described above for polytomous response models. The topic of linking from two different calibrations of polytomous items has been recognized to be of value, as in some instances these models allow for the measurement of different constructs (e.g., Muraki, Humbo & Lee, 2000; Wan & Henly, 2012). The problem of linking functions for these models has already been investigated (Baker, 1992, 1993; Muraki & Chang, 1994; Kim & Hanson, 2002; Koenig & Roberts, 2007; von Davier & von Davier, 2007), although these previous studies were primarily motivated by the notion of adjusting the zero and unit for the scale of the θ parameters only. These studies also estimated the linking functions using item-response function and test-characteristic methods, with the exception of the moment methods explored by Kim and Hanson (2002) and Kim and Lee (2006). The current authors are not aware of any studies in which asymptotic standard errors of linking functions for polytomous models have been addressed to date, with the exception of Kolen and Brennan (2004, p. 263)

who state that bootstrapping methods may be useful but computationally intensive, and that development of approximations to standard errors of IRT linking would be useful for planning sample sizes in linking designs.

We will start with a short introduction to common polytomous IRT models as derived from step functions (e.g., Penfield 2014). The derivation provides the foundation for our linking approach. Next, we will briefly discuss lack of identification of the polytomous models and the resulting need to link parameters from calibrations with different identifiability restrictions. We then derive the linking functions for polytomous models, discuss their estimation, and derive their asymptotic standard errors of linking. Finally, we provide empirical examples of the use of the new linking approach.

5.2 Models

For the case of polytomous items with response categories $k = 1, ..., K$, let U_{pik} be the binary variable used to indicate whether ($U_{pik} = 1$) or not ($U_{pik} = 0$) test-taker $p = 1, ..., P$ produced a response in category k for item $i = 1, ..., I$. The distribution of the response vectors $\mathbf{U}_{pi} = (U_{pi1}, ..., U_{piK})$ by the test-takers on the items belongs to the family of multinomial distributions with probability functions

$$f(\mathbf{u}_{pi}; \boldsymbol{\pi}_{pi}) = \prod_{k=1}^{K} \pi_{pik}^{u_{pik}}, \; p = 1, ..., P; \; i = 1, ..., I; \; k = 1, ..., K, \tag{5.9}$$

where $\boldsymbol{\pi}_{pi} = (\pi_{pi1}, ..., \pi_{piK})$, with $\boldsymbol{\pi}'_{pi} \cdot \mathbf{1} = 1$. Likewise, under the assumption of statistical independence, the family of joint distributions of a complete array of responses $\mathbf{U} = (U_{pik})$ generalizes to the product of $P \times I \times K$ of these multinomial functions, $f(\mathbf{u}; \boldsymbol{\pi}) = \prod_{p} \prod_{i} \prod_{k=1}^{K} \pi_{pik}^{u_{pik}}$, with parameter vector $\boldsymbol{\pi} = (\pi_{11}, ..., \pi_{1P}, ..., \pi_{I1}, ..., \pi_{IP})$. The well-known case of dichotomous items is obtained for $K = 2$ and $u_{pi1} = 1 - u_{pi2}$. It is important to note that parameter vector $\boldsymbol{\pi}$ is identified; for each distinct set of values it produces a distinct distribution of $\mathbf{U}$.

To further model these multinomial distributions, the options include the nominal response model, partial credit models, sequential response models and graded response models (e.g., Glas, 2010; Ostini & Nering, 2006; Penfield, 2014). Item response theory makes use of the notion of a step function to derive these models—a name somewhat misleading in that it suggests solutions to items obtained through a step-wise cognitive process, an interpretation that does not generally hold for the models reviewed here. The differences among the models

stem from their definition of what constitutes a "step." Each of these models has a parallel in categorical data analysis in the form of a linear baseline, adjacent-category, continuation-ratio, and cumulative logit model (e.g., Agresti, 1990, 2013, chap. 8), but as the current context is IRT we use its own terminology and specifications.

The general choice for the step function is

$$\Psi_{pik} = \frac{\exp[a_i(\theta_p - b_{ik})]}{1 + \exp[a_i(\theta_p - b_{ik})]}, \tag{5.10}$$

with $a_i > 0$ and $b_{ik}, \theta_p \in \mathbb{R}$. An item with K categories will have $K - 1$ step functions, one for each comparison between the K categories as defined for each of the models below. (Note that we actually should have indexed the step functions differently. But since each of them focuses on one of the categories $k = 1, ..., K$ relative to others, use of the same notation helps us to remember their main association.) For the case of two categories, (5.10) reduces to the response function of the 2PL model, and for the additional choice of $a_i = 1$ for all i to that of the 1PL or Rasch model. For the case of more than two categories, the probability of guessing is assumed to be absorbed in the probabilities Ψ_{pik}; hence the absence of a guessing parameter in (5.10). Also, observe that, except for $\theta_p = b_{ik}$, the step function is monotone in each of the parameters θ, a, and b when all others are fixed.

5.2.1 Nominal Response Models

The nominal model defines the step function for pairs of categories using one of them as baseline or nominal category. The choice of baseline category is arbitrary. We choose category K as our baseline. The step function for category k is then

$$\begin{aligned} \Psi_{pik} &= \frac{\pi_{pik}}{\pi_{pik} + \pi_{piK}} \\ &= \frac{\exp\left[a_i(\theta_p - b_{ik})\right]}{1 + \exp\left[a_i(\theta_p - b_{ik})\right]},\ k = 1, ..., K-1. \end{aligned} \tag{5.11}$$

From (5.11),

$$\pi_{pik} = \pi_{piK} \exp\left[a_i(\theta_p - b_{ik})\right],\ k = 1, ..., K. \tag{5.12}$$

The category-response functions for the nominal response model typically recognized in IRT applications (Bock, 1972, 1997) follow from (5.12) as the probability of a response in category

k relative to any of the categories:

$$\begin{aligned}
\pi_{pik} &= \frac{\pi_{pik}}{\pi_{pi1} + \pi_{pi2} + ... + \pi_{pik} + ... + \pi_{pi(K)}} \\
&= \frac{\pi_{pik}}{\pi_{piK}\left[\exp a_i(\theta_p - b_{iK})\right] + \sum_{h=1}^{K-1} \pi_{piK}\left[\exp a_i(\theta_p - b_{ih})\right]} \\
&= \frac{\pi_{pik}}{\pi_{piK}[1 + \sum_{h=1}^{K-1} \exp[a_i(\theta_p - b_{ih})]]} \\
&= \begin{cases} \dfrac{\exp\left[a_i(\theta_p - b_{ik})\right]}{1 + \sum_{h=1}^{K-1} \exp[a_i(\theta_p - b_{ih})]}, & k = 1, ..., K-1, \\ \\ \dfrac{1}{1 + \sum_{h=1}^{K-1} \exp[a_i(\theta_p - b_{ih})]}, & k = K, \end{cases}
\end{aligned} \tag{5.13}$$

where $a_i > 0$ and $b_{ik}, \theta_p \in R$. Observe that b_{iK} in (5.12) may be arbitrarily defined as it is unidentified and cancels itself out of the numerator and denominator of (5.13) for all k.

We did share some of the algebraic steps to illustrate the differences between the step functions and the category-response functions. Specifically, note that the probabilities in (5.13) are *not* monotone in the parameters but the step functions in (5.11) from which they mathematically follow do have this feature. The same holds for the models below.

5.2.2 Partial Credit Models

In contrast to the nominal models, partial credit models define the step function for the responses in category k and one of its adjacent categories. In IRT typically adjacent category $k-1$ is chosen. The step functions now become

$$\begin{aligned}
\Psi_{pik} &= \frac{\pi_{pik}}{\pi_{pik} + \pi_{pi(k-1)}} \\
&= \frac{\exp\left[a_i(\theta_p - b_{ik})\right]}{1 + \exp\left[a_i(\theta_p - b_{ik})\right]}, \; k = 2, ..., K.
\end{aligned} \tag{5.14}$$

Alternatively the equation can be written as

$$\log \frac{\pi_{pik}}{\pi_{pi(k-1)}} = a_i(\theta_p - b_{ik}).$$

For an arbitrary category k, it follows that

$$\log \frac{\pi_{pik}}{\pi_{pi1}} = \log \frac{\pi_{pik}}{\pi_{pi(k-1)}} + \log \frac{\pi_{pi(k-1)}}{\pi_{pi(k-2)}} + ... + \log \frac{\pi_{pi(2)}}{\pi_{pi(1)}}.$$

Thus,

$$\begin{aligned} \frac{\pi_{pik}}{\pi_{pi1}} &= \exp(\log \frac{\pi_{pik}}{\pi_{pi(k-1)}} + \log \frac{\pi_{pi(k-1)}}{\pi_{pi(k-2)}} + ... + \log \frac{\pi_{pi(2)}}{\pi_{pi(1)}}) \\ &= \exp(\sum\nolimits_{h=1}^{k} a_i(\theta_p - b_{ih})), \end{aligned}$$

and therefore

$$\pi_{pik} = \pi_{pi1} \exp(\sum\nolimits_{h=1}^{k} a_i(\theta_p - b_{ih})),\ k = 1, ...K. \tag{5.15}$$

The response functions for this category can now be written as

$$\begin{aligned} \pi_{pik} &= \frac{\pi_{pik}}{\pi_{pi1} + \pi_{pi2} + ... + \pi_{pik} + ... + \pi_{pi(K)}} \\ &= \begin{cases} \dfrac{1}{1 + \sum_{d=2}^{K} \exp(\sum_{h=2}^{d} a_i(\theta_p - b_{ih}))}, & k = 1, \\ \dfrac{\exp(\sum_{h=2}^{k} a_i(\theta_p - b_{ih}))}{1 + \sum_{d=2}^{K} \exp(\sum_{h=2}^{d} a_i(\theta_p - b_{ih}))}, & k = 2, ..., K. \end{cases} \end{aligned} \tag{5.16}$$

The result is the generalized partial credit model (GPCM) (Muraki, 1992, 1997). Master's (1982) partial credit model (PCM) follows upon the assumption of $a_i = 1$ for all i. Observe that b_{i1} in (5.15) may be arbitrarily defined as it is unidentified and cancels itself out of both the numerator and denominator of this expression for all k.

5.2.3 Sequential Response Models

These models arise when the response categories are ordered in the sense that the test-taker can only take the step to the next category when all the preceding steps are successfully completed. It thus assumes that the response process stops as soon as the test-taker fails on a step. The step function is defined for responses in category k and any of the categories $k+1$ through K:

$$\begin{aligned} \Psi_{pik} &= \frac{\pi_{pik}}{\pi_{pik} + \pi_{pi(k+1)} + ... + \pi_{piK}} \\ &= \frac{\exp\left[a_i(\theta_p - b_{ik})\right]}{1 + \exp\left[a_i(\theta_p - b_{ik})\right]},\ k = 2, ..., K. \end{aligned} \tag{5.17}$$

The probability of a response in category k is equal to the joint probability of success at steps $1, ..., k$ but failure at step $k+1$; that is,

$$\pi_{pik} = \begin{cases} 1 - \Psi_{pi2}, & k = 1, \\ \Psi_{pi2} \times ... \times \Psi_{pik} \times [1 - \Psi_{pi(k+1)}], & k = 2, ..., K-1, \\ \Psi_{pi2} \times ... \times \Psi_{piK}, & k = K. \end{cases}$$

Or

$$\pi_{pik} = \begin{cases} \dfrac{1}{1 + \exp[a_i(\theta_p - b_{i2})]}, & k = 1, \\ \dfrac{\prod_{h=2}^{k} \exp[a_i(\theta_p - b_{ih})]}{\prod_{h=2}^{k+1} \{1 + \exp[a_i(\theta_p - b_{ih})]\}}, & k = 2, ..., K-1, \\ \dfrac{\prod_{h=2}^{K} \exp[a_i(\theta_p - b_{ih})]}{\prod_{h=2}^{K} \{1 + \exp[a_i(\theta_p - b_{ih})]\}}, & k = K. \end{cases} \tag{5.18}$$

For $a_i = 1$ for all i, the model was introduced as the sequential model for ordered responses by Tutz (1990, 1997) and Verhelst, Glas, and de Vries (1997).

5.2.4 Graded Response Models

This type of model assumes a step function equal to

$$\Psi_{pik} = \pi_{pik+1} + \pi_{pik+2} + ... + \pi_{piK} = \frac{\exp\left[a_i(\theta_p - b_{ik})\right]}{1 + \exp\left[a_i(\theta_p - b_{ik})\right]}, k = 2, ..., K. \tag{5.19}$$

The probability of a response in category k is now just the difference between the two step functions for categories k and $k+1$; that is,

$$\pi_{pik} = \begin{cases} 1 - \Psi_{pi2}, & k = 1, \\ \Psi_{pik} - \Psi_{pik+1} & k = 2, ..., K-1, \\ \Psi_{piK} & k = K. \end{cases}$$

Thus,

$$\pi_{pik}=\begin{cases} \dfrac{1}{1+\exp[a_i(\theta_p-b_{i2})]}, & k=1, \\ \dfrac{\exp a_i(\theta_p-b_{ik+1})-\exp a_i(\theta_p-b_{ik})}{[1+\exp a_i(\theta_p-b_{ik})]\,[1+\exp a_i(\theta_p-b_{ik+1})]}, & k=2,...,K-1, \\ \dfrac{\exp\left[a_i(\theta_p-b_{iK})\right]}{1+\exp\left[a_i(\theta_p-b_{iK})\right]}, & k=K. \end{cases} \tag{5.20}$$

For various additional specifications of this type of model, see Samejima (1969, 1997).

5.3 The Need to Link

As already noted, the problem of linking arises because of the lack of identifiability of IRT models. Although the standard success parameters of binomial or multinomial response distributions they explain are identified, they do so by introducing a second level of item and ability parameters that are generally not. The purpose of this paper is not to address the issue of identifiability for any of the models above in much detail, nor is it to derive sufficient restrictions to identify these models. Discussions of those topics may be found, for instance, in Bechger, Verhelst, and Verstralen (2001), Bechger, Verstralen, and Verhelst (2002), Fischer (2004), Maris (2002), Maris and Bechger (2004, 2009), Reiersøl (1950), Revuelta (2009), San Martín, Gonzáles, and Tuerlinckx (2009), San Martín, Jara, Rolin, and Mouchart (2011), Tsai (2000), and Volodin and Adams (2002). However, as parameter linking, which is the focus of this paper, is required due to lack of identifiability, we present some concepts related to identifiability and then present a few examples that demonstrate the lack of identifiability of polytomous IRT models.

In polytomous response models, multiple sets of values for the parameters θ_p, a_i,and b_{ik} present in the expression $a_i(\theta_p - b_{ik})$ on the right-hand side of each of the model equations in (5.13), (5.16), (5.18), or (5.20) can imply the same probabilities π_{pik} on their left-hand side. A transformation of a_i by a factor of $1/u$, $u \neq 0$, can be absorbed by θ_p and b_{ik} by multiplying them by a factor of u. Likewise, translating θ_p by a constant $v \in R$ can be offset the same translation

of b_{ik} and vice versa. More formally, it thus holds that

$$a_i(\theta_p - b_{ik}) = \frac{a_i}{u}(u(\theta_p - v) - u(b_{ik} - v)), \tag{5.21}$$

which is a well-known characteristic both of dichotomous and the above polytomous IRT models.

As discussed earlier, in practice MML estimation with the assumption of $\theta_p \sim N(0, 1)$ is used to calibrate the items. The tradition of successful parameter recovery studies for this procedure does suggest sufficiency of the assumption, but a formal proof of it is still lacking.

5.4 True Linking Functions

Consider the case in which two different calibrations have been conducted with empirically different identifiability restrictions but some common items and/or test-takers shared among the two test administrations. For these common elements, although all success probabilities associated with them are identified, their true parameter values for each calibration are not identical; each of them represents a different selection from their sets of observationally equivalent values due to the imposition of different restrictions. In order to compare them, we therefore must map the parameters from one calibration onto the parameters of the other. Linking of this nature is common in the testing industry. While concurrent calibration of the response data from both administrations may eliminate the issue (von Davier & von Davier, 2011), this is rarely practical for testing programs that have to link their parameters continuously across multiple test administrations. We will first work with the mathematical aspects of the problem at the level of the true model parameters, even though in practice these parameters are unknown and must be estimated. The issue of estimation of the linking functions will be addressed subsequently.

As shown by van der Linden and Barrett (in press, theorem 3), for a response model with a structure of fixed-effect parameters that (i) specifies success probability π as a monotone continuous function of $\boldsymbol{\xi}$ and (ii) has been used in two different calibration studies with identified parameters $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}$, the two sets of parameters are linked by a vector function

$$\boldsymbol{\xi}^* = \varphi(\boldsymbol{\xi}) = (\varphi_1(\xi_1), ..., \varphi_d(\xi_d)) \tag{5.22}$$

with components that are monotone and continuous as well.

The polytomous models presented by the category-response functions in (5.13), (5.16), (5.18), and (5.20) do not specify $\boldsymbol{\pi}_{ik}$ as a monotone function of $\boldsymbol{\xi} = (\boldsymbol{\theta}, \mathbf{a}, \mathbf{b})$. This is easily illustrated

in Figure 5.1; except for $k = 1$ and K, the probabilities of success $\boldsymbol{\pi}_{ik}$ first increase and then decrease with θ. Consequently, the result in (5.22) can *not* be used to derive the linking functions for any of the four models using a functional equation as in (5.2). Neither can they be derived replacing the category-response functions by their item characteristic curves; that is, their expected scores $\sum k\pi_{ik}$ as a function of θ. These functions are not always monotone in θ either, for instance, (5.13) does not involve any ordering of categories.

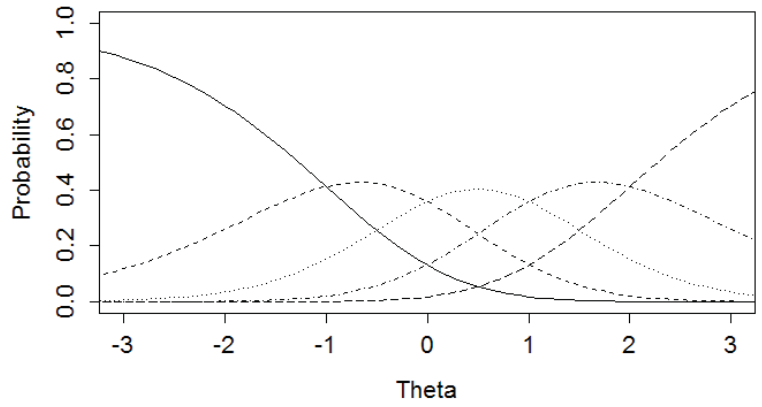

Figure 5.1. Example of a polytomous model with π_{ik} as a function of θ_p for $k = 1, 2, ..., 5$.

However, each of the basic step functions in (5.11), (5.14), (5.17) and (5.19) do have monotonicity in (θ_p, a_i, b_{ik}), and their models were mathematically derived or defined by them for exactly the same parameters. Consequently, the linking functions for the polytomous models can be derived from their underlying step functions, entirely analogous to the case of the dichotomous models addressed by van der Linden and Barrett (in press). The approach is illustrated in Figure 5.2: The plots on the left-hand and right-hand sides are for the same item in two different calibrations with different true parameter values due to the use of different identifiability restrictions. As the same true parameter values figure both in the step and the category-response functions, linking functions that map the step functions onto each other do the same with their associated response functions. Of course, the equivalence only holds at the current level of the true parameters. As these parameters are always estimated from likelihood equations defined by the category-response probabilities, the statistical properties of the estimated linking functions are dependent on the response functions solely. Indeed, as will become clear below, the major difference between linking for dichotomous and polytomous model parameters exists in statistical features, *not* in the true linking function that is estimated.

We document the conclusion as follows:

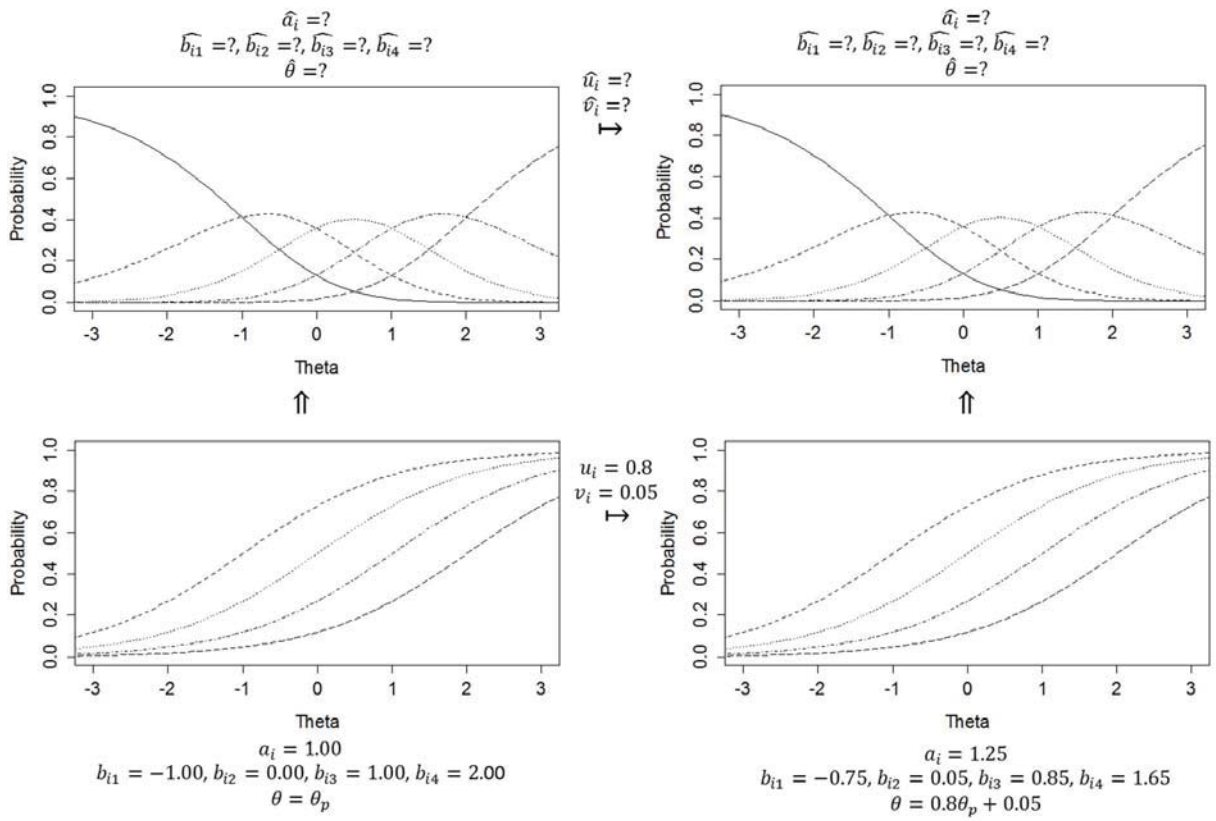

Figure 5.2. Polytomous response model linking as implied by step function linking, with the item from the first calibration (left-hand panes) linked to those in the second calibration (right-hand panes). The true parameter values for the step functions in the first calibration are mapped onto those in the second by linking equations with parameters $u = .8$ and $v = .05$. The same holds for the category response functions. The parameters are estimated, however, from likelihood equations defined by the latter.

Theorem 1. The linking functions for each of the nominal, partial credit, sequential response, and graded response models are identical to those for the 2PL response functions used as step functions in their derivation. The only differences exist in the statistical properties of their estimates.

For each of the four polytomous models we can thus use the earlier linking functions in (5.3)–(5.8), ignoring the one for the c_i in (5.5). Observe, however, that for the result to hold both the step and response functions need to have the same basic parameterization. In our earlier derivation of the models, the parameterization was $a_i(\theta_p - b_{ik})$. It is not uncommon, however, to present some of these models using a slope-intercept parameterization, $a_i\theta_p + b_{ik}$. The linking functions in (5.3)–(5.8) then do not hold but adopt a more complicated form and can

only be estimated from a practically infeasible type of linking design with common test-takers responding to common items in both calibrations (van der Linden & Barrett, 2015, theorem 6).

5.5 Linking Design

Using (5.3)–(5.8), we can define a system of equations of the unknown linking parameters u and v in the model parameters for common elements in the two calibrations. In order for this system to be identifiable, an appropriate linking design must be selected. Minimal linking elements for linking designs for polytomous models that lead to identifiable systems are given below. As we will see, the major difference with dichotomous linking exist in the presence of multiple equivalent solutions of the linking parameters due to multiple response categories.

In the remainder of this paper, we use $i = 1, ...I$, $p = 1, ..., P$, and $k = 1, ..., K-1$ to denote the common items, test-takers, and steps in the design, and $t = 1, 2$ to denote the two calibrations. Minimal linking elements in a linking design will be denoted as $m = 1, ..., M$.

5.5.1 One Common Item

Linking parameters u and v are identified by a linking element m consisting of a single common item $i = 1$ in the two calibrations. The system of linking equations follows from (5.3) and (5.8) as

$$u = \frac{a_{1_1}}{a_{1_2}},\ a_{1_2} > 0, \tag{5.23}$$

$$v = b_{1k_2} - ub_{1k_1},\ k = 1, ..., K-1 \tag{5.24}$$

Note that v is already identified for the choice of a single category $k = 1, ..., K-1$. Each of these choices will return the same true value for v. In practice, as all parameters are estimated, multiple estimates of v will be available. However, rather than choosing between these different estimates statistically it makes more sense to combine them into a better estimate. This is addressed subsequently.

5.5.2 Two Common Items

Alternatively, linking parameters u and v are identified by a linking element m comprised of two common items $i = 1, 2$ in the two calibrations. The system of linking equations then follows

from (5.8) with subsequent elimination of v as

$$u = \frac{b_{1k_2} - b_{2k_2}}{b_{1k_1} - b_{2k_1}},\ b_{1k_1} \neq b_{2k_1},\ k = 1, ..., K-1; \tag{5.25}$$

$$v = b_{ik_2} - ub_{ik_1},\ i = 1, 2,\ k = 1, ..., K-1. \tag{5.26}$$

Note that in this case multiple solutions exist both for u and v, one for each of the possible response categories. Again, each of these solutions returns the same true value for u and v and, just as for the case of one common items, and their different estimates need to be combined.

5.5.3 One Common Test-Taker

Linking parameters u and v are not identified when the linking design has only one common test-taker $p = 1$. It is impossible to derive a system of equations for u and v from (5.3)–(5.8) for the two θ values for one common test-taker.

5.5.4 Two Common Test-Takers

This case is entirely analogous to that of a linking design with a minimal linking element m comprised of two common test-takers $p = 1, 2$ for the dichotomous 2PL model. Linking parameters u and v can be obtained from (5.3)–(5.8) for a pair of test-takers as

$$u = \frac{\theta_{1_2} - \theta_{2_2}}{\theta_{1_1} - \theta_{2_1}},\ \theta_{1_1} \neq \theta_{1_2} \tag{5.27}$$

and

$$v = \theta_{p_2} - u\theta_{p_1},\ p = 1, 2. \tag{5.28}$$

Again either choice of $p = 1, 2$ will return the same true value of v for the design.

5.6 Estimating Linking Functions

In practice, response model parameters are estimated. So, when calculating linking parameters u and v, random error in the model parameter estimates propagates into them. Analogous to the dichotomous case (Barrett & van der Linden, 2015a), the asymptotic standard errors for

these linking parameters can be approximated using the first-order multivariate delta method (e.g., Casella & Berger, 2002). There is, however, a critical distinction for the polytomous response models. As their covariance matrices show, the estimates of the category difficulty parameters for the common items, $\widehat{b}_{ik}$, $k = 1, ..., K-1$, are not independent. Consequently, for each of them, we have $K-1$ dependent estimators $\widehat{v}_{ik}$ of v, whereas a design with common pair of items yields $K-1$ dependent estimators $\widehat{u}_{ik}$ and $2(K-1)$ dependent estimators $\widehat{v}_{ik}$ for each pair.

Just as for the case of common pairs of dichotomous items, where a comparable issue arose for $\widehat{v}$ due to an estimate $\widehat{u}$ shared by both, we use a simple average of these estimators and allow the delta method to account for the dependencies. (Right here, it might already look tempting to improve on the simple average using the same idea of precision weighing applied later in this paper. But we do not have any standard errors to weigh the separate estimators yet; in fact, the current averages are used to derive them for the linking elements.)

5.6.1 One Common Item

For the case of a single common item the delta method will be applied to the linking equation for u in (5.23) with the one for v in (5.24) replaced by

$$v = (K-1)^{-1} \sum\nolimits_{k=1}^{K-1} (b_{1k_2} - u b_{1k_1}). \tag{5.29}$$

Let $\boldsymbol{\eta} = (\eta) = (u, v)$ denote the vector of the estimated linking parameters and $\boldsymbol{\xi}_{1_t} = (a_{1_t}, b_{1k_t})$ the values of estimated parameters a_1 and b_{1k}, $k = 1, ..., K-1$, of the common item in calibration $t = 1, 2$, and $\boldsymbol{\xi} = \left(\boldsymbol{\xi}_{1_t}\right)$. The system of linking equations in (5.23) and (5.29) defines a vector function $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$.

The delta method enables us to approximate the 2×2 covariance matrix of $\widehat{\boldsymbol{\eta}}$ as

$$\mathrm{Cov}(\widehat{\boldsymbol{\eta}}|\boldsymbol{\eta}) = \boldsymbol{J}_{\varphi}^{'} \mathrm{Cov}(\widehat{\boldsymbol{\xi}}|\boldsymbol{\xi}) \boldsymbol{J}_{\varphi}, \tag{5.30}$$

where

$$\boldsymbol{J}_{\varphi} = \left(\frac{\partial \eta}{\partial \xi}\right) \tag{5.31}$$

is the $2K \times 2$ Jacobian matrix associated with $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$ and $\mathrm{Cov}(\widehat{\boldsymbol{\xi}}|\boldsymbol{\xi})$ is the $2K \times 2K$ covariance matrix for the estimators of the elements of $\boldsymbol{\xi}$. Assuming different test-takers in each

administration, $\mathrm{Cov}(\widehat{\boldsymbol{\xi}}|\boldsymbol{\xi})$ is block diagonal with the two $K \times K$ covariance matrices

$$\begin{bmatrix} \sigma^2_{a_{1_t}} & \sigma_{a_{1_t}b_{1k_t}} & \cdots & \sigma_{a_{1_t}b_{1(K-1)_t}} \\ \sigma_{a_{1_t}b_{1k_t}} & \sigma^2_{b_{1k_t}} & \cdots & \sigma_{b_{1k_t}b_{1(K-1)_t}} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{a_{1_t}b_{1(K-1)_t}} & \sigma_{b_{1k_t}b_{1(K-1)_t}} & \cdots & \sigma^2_{b_{1(K-1)_t}} \end{bmatrix}, \; k = 1, ..., K-1, \; t = 1, 2, \tag{5.32}$$

as blocks. We are interested in the standard errors of the estimators of $\boldsymbol{\eta} = (u, v)$; that is, the diagonal elements of the left-hand side of (5.30). The elements follow from (5.32), using the corresponding elements of (5.31), as

$$\begin{aligned} \mathrm{Var}(\widehat{u}_m|\boldsymbol{\xi}) \;=\; & \sum\nolimits_{t=1}^{2} \left\{ \left(\frac{\partial u}{\partial a_{1t}} \right)^2 \sigma^2_{a_{1_t}} + 2 \frac{\partial u}{\partial a_{1t}} \left(\sum\nolimits_{k=1}^{K-1} \frac{\partial u}{\partial b_{1k_t}} \sigma_{a_{1_t}b_{1k_t}} \right) \right. \\ & \left. + \sum\nolimits_{k=1}^{K-1} \left[\frac{\partial u}{\partial b_{1k_t}} \left(\sum\nolimits_{j=1}^{K-1} \frac{\partial u}{\partial b_{1j_t}} \sigma_{b_{1jt}b_{1k_t}} \right) \right] \right\} \end{aligned} \tag{5.33}$$

and

$$\begin{aligned} \mathrm{Var}(\widehat{v}_m|\boldsymbol{\xi}) \;=\; & \sum\nolimits_{t=1}^{2} \left\{ \left(\frac{\partial v}{\partial a_{1t}} \right)^2 \sigma^2_{a_{1_t}} + 2 \frac{\partial v}{\partial a_{1t}} \left(\sum\nolimits_{k=1}^{K-1} \frac{\partial v}{\partial b_{1k_t}} \sigma_{a_{1_t}b_{1k_t}} \right) \right. \\ & \left. + \sum\nolimits_{k=1}^{K-1} \left[\frac{\partial v}{\partial b_{1k_t}} \left(\sum\nolimits_{j=1}^{K-1} \frac{\partial v}{\partial b_{1j_t}} \sigma_{b_{1jt}b_{1k_t}} \right) \right] \right\}. \end{aligned} \tag{5.34}$$

With the function $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$ defined by equations (5.23) and (5.29), the partial derivatives in (5.31) are

$$\frac{\partial u}{\partial a_{1_1}} = \frac{1}{a_{1_2}} \tag{5.35}$$

$$\frac{\partial u}{\partial b_{1_{k_t}}} = 0, t = 1, 2, k = 1, ..., K-1 \tag{5.36}$$

$$\frac{\partial v}{\partial a_{1_1}} = \frac{-1}{a_{1_2}\,(K-1)} \sum\nolimits_{k=1}^{K-1} b_{1_{k_1}} \tag{5.37}$$

$$\frac{\partial v}{\partial a_{1_2}} = \frac{u}{a_{1_2}(K-1)} \sum\nolimits_{k=1}^{K-1} b_{1_{k_1}} \tag{5.38}$$

$$\frac{\partial v}{\partial b_{1_{k_1}}} = \frac{-a_{1_1}}{a_{1_2}(K-1)} \tag{5.39}$$

$$\frac{\partial v}{\partial b_{1_{k_2}}} = \frac{a_{1_2}}{a_{1_2}(K-1)} \tag{5.40}$$

The standard error of u_m is obtained from (5.33) entirely analogous to the dichotomous case, as all partial derivatives of u with respect to the b_k parameters are equal to zero. Thus,

$$\sigma_{u_m} = \left(\frac{\sigma^2_{a_{i_1}} + u^2 \sigma^2_{a_{i_2}}}{a^2_{i_2}} \right)^{1/2}, \tag{5.41}$$

with $a_{i_2} \neq 0$. However, the standard error of v_m becomes

$$\begin{aligned} \sigma_{v_m} \;=\; & \frac{1}{a_{1_2}(K-1)} \Bigg\{ \left(\sum\nolimits_{k=1}^{K-1} b_{1k_1} \right)^2 \left(\sigma^2_{a_{1_1}} + u \sigma^2_{a_{1_2}} \right) \\ & + 2a_{1_1} \left(\sum\nolimits_{k=1}^{K-1} b_{1k_1} \right) \left(\sum\nolimits_{k=1}^{K-1} \sigma_{a_{1_1} b_{1k_1}} + \sum\nolimits_{k=1}^{K-1} \sigma_{a_{1_2} b_{1k_2}} \right) \\ & + a^2_{1_2} \left(u^2 \sum\nolimits_{k=1}^{K-1} \sum\nolimits_{j=1}^{K-1} \sigma_{b_{1j_1} b_{1k_1}} + \sum\nolimits_{k=1}^{K-1} \sum\nolimits_{j=1}^{K-1} \sigma_{b_{1j_2} b_{1k_2}} \right) \Bigg\}^{1/2}, \end{aligned} \tag{5.42}$$

with $a_{i_2} \neq 0$. For $k = 2$, the latter simplifies to the standard error for the dichotomous case presented in Barrett and van der Linden (2015a).

5.6.2 Two Common Items

For the case of two common items, we have $K-1$ dependent estimators u_{ik} and $2(K-1)$ dependent estimators v_{ik} for a single minimal linking element m. Hence, we will use the delta method for

$$u = (K-1)^{-1} \sum\nolimits_{k=1}^{K-1} \left(\frac{b_{1_{k_2}} - b_{2_{k_2}}}{b_{1_{k_1}} - b_{2_{k_1}}} \right), \tag{5.43}$$

with $b_{1_{k_1}} - b_{2_{k_1}} \neq 0$ and

$$v = \frac{1}{2(K-1)} \sum_{i=1}^{2} \sum_{k=1}^{K-1} \left(b_{i_{k_2}} - u b_{i_{k_1}} \right). \tag{5.44}$$

This system of equations defines a vector function $\boldsymbol{\eta} = \varphi(\boldsymbol{\xi})$ with $\boldsymbol{\xi} = (\boldsymbol{\xi}_t)$, where $\boldsymbol{\xi}_t = (b_{1_{k_t}}, b_{2_{k_t}})$, $k = 1, ..., K-1$. $\boldsymbol{J}_\varphi$ is now of dimension $4(K-1) \times 2$ and, because of local independence, $\mathrm{Cov}(\widehat{\boldsymbol{\xi}} | \boldsymbol{\xi})$ becomes a $4(K-1) \times 4(K-1)$ diagonal matrix with blocks

$$\begin{bmatrix} \sigma^2_{b_{i_{k_t}}} & \dots & \sigma_{b_{i_{k_t}} b_{i_{(K-1)_t}}} \\ \dots & \dots & \dots \\ \sigma_{b_{i_{k_t}} b_{i_{(K-1)_t}}} & \dots & \sigma^2_{b_{i_{(K-1)_t}}} \end{bmatrix}, \; k = 1, ..., K-1, \; i = 1, 2, \; t = 1, 2. \tag{5.45}$$

as elements.

Again, we are interested in the diagonal elements of $\mathrm{Cov}(\widehat{\boldsymbol{\eta}}|\boldsymbol{\eta})$, which follow from (5.45) using the corresponding elements of (5.31), as

$$Var(\widehat{u}_m \mid \boldsymbol{\xi}) = \sum_{t=1}^{2} \left\{ \sum_{i=1}^{2} \left[\sum_{k=1}^{K-1} \frac{\partial u}{\partial b_{ik_t}} \left(\sum_{j=1}^{K-1} \frac{\partial u}{\partial b_{ij_t}} \sigma_{b_{ik_t} b_{ij_t}} \right) \right] \right\} \tag{5.46}$$

and

$$Var(\widehat{v}_m \mid \boldsymbol{\xi}) = \sum_{t=1}^{2} \left\{ \sum_{i=1}^{2} \left[\sum_{k=1}^{K-1} \frac{\partial v}{\partial b_{ik_t}} \left(\sum_{j=1}^{K-1} \frac{\partial v}{\partial b_{ij_t}} \sigma_{b_{ik_t} b_{ij_t}} \right) \right] \right\}. \tag{5.47}$$

The partial derivatives $\boldsymbol{J}_{\varphi}$ required in (5.46)–(5.47) are

$$\frac{\partial u}{\partial b_{1k_1}} = \frac{-b_{1k_2} + b_{2k_2}}{(K-1)(b_{1k_1} - b_{2k_1})^2} \tag{5.48}$$

$$\frac{\partial u}{\partial b_{1k_2}} = \frac{1}{(K-1)\,(b_{1k_1} - b_{2k_1})} \tag{5.49}$$

$$\frac{\partial u}{\partial b_{2k_1}} = \frac{b_{1k_2} - b_{2k_2}}{(K-1)\,(b_{1k_1} - b_{2k_1})^2} \tag{5.50}$$

$$\frac{\partial u}{\partial b_{2k_2}} = \frac{-1}{(K-1)\,(b_{1k_1} - b_{2k_1})} \tag{5.51}$$

$$\frac{\partial v}{\partial b_{1k_1}} = \frac{-1}{2(K-1)} \left[u + \frac{\partial u}{\partial b_{1k_1}} (b_{1k_1} + b_{2k_1}) \right] \tag{5.52}$$

$$\frac{\partial v}{\partial b_{1k_2}} = \frac{-1}{2(K-1)} \left[-1 + \frac{\partial u}{\partial b_{1k_2}} (b_{1k_1} + b_{2k_1}) \right] \tag{5.53}$$

$$\frac{\partial v}{\partial b_{2k_1}} = \frac{-1}{2(K-1)} \left[u + \frac{\partial u}{\partial b_{2k_1}} (b_{1k_1} + b_{2k_1}) \right] \tag{5.54}$$

$$\frac{\partial v}{\partial b_{2k_2}} = \frac{-1}{2(K-1)} \left[-1 + \frac{\partial u}{\partial b_{2k_2}} (b_{1k_1} + b_{2k_1}) \right]. \tag{5.55}$$

With some simplification, the standard errors can now be written as

$$\begin{aligned}\sigma_{u_m} &= \frac{1}{(K-1)} \times \\ &\left\{ \sum_{k=1}^{K-1} \frac{-(b_{1k_2} - b_{2k_2})}{(b_{1k_1} - b_{2k_1})^2} \left(\sum_{j=1}^{K-1} \frac{-(b_{1j_2} - b_{2j_2})}{(b_{1j_1} - b_{2j_1})^2} \sigma_{b_{1k_1} b_{1j_1}} \right) \right. \\ &+ \sum_{k=1}^{K-1} \frac{b_{1k_2} - b_{2k_2}}{(b_{1k_1} - b_{2k_1})^2} \left(\sum_{j=1}^{K-1} \frac{b_{1j_2} - b_{2j_2}}{(b_{1j_1} - b_{2j_1})^2} \sigma_{b_{2k_1} b_{2j_1}} \right) \\ &+ \sum_{k=1}^{K-1} \frac{1}{(b_{1k_1} - b_{2k_1})} \left(\sum_{j=1}^{K-1} \frac{1}{(b_{1j_1} - b_{2j_1})} \sigma_{b_{1k_2} b_{1j2}} \right) \\ &\left. + \sum_{k=1}^{K-1} \frac{-1}{(b_{1k_1} - b_{2k_1})} \left(\sum_{j=1}^{K-1} \frac{-1}{(b_{1j_1} - b_{2j_1})} \sigma_{b_{2k_2} b_{2j_2}} \right) \right\}^{1/2}\end{aligned} \tag{5.56}$$

and

$$\begin{aligned}\sigma_{v_m} &= \frac{1}{2(K-1)} \times \\ &\left\{ \sum_{k=1}^{K-1} \left(\left[u + \frac{\partial u}{\partial b_{1k_1}} (b_{1k_1} + b_{2k_1}) \right] \sum_{j=1}^{K-1} \left(\left[u + \frac{\partial u}{\partial b_{1j_1}} (b_{1j_1} + b_{2j_1}) \right] \sigma_{b_{1k_1} b_{1j_1}} \right) \right) \right. \\ &+ \sum_{k=1}^{K-1} \left(\left[u + \frac{\partial u}{\partial b_{2k_1}} (b_{1k_1} + b_{2k_1}) \right] \sum_{j=1}^{K-1} \left(\left[u + \frac{\partial u}{\partial b_{2j_1}} (b_{1j_1} + b_{2j_1}) \right] \sigma_{b_{2k_1} b_{2j_1}} \right) \right) \\ &+ \sum_{k=1}^{K-1} \left(\left[-1 + \frac{\partial u}{\partial b_{1k_2}} (b_{1k_1} + b_{2k_1}) \right] \sum_{j=1}^{K-1} \left(\left[-1 + \frac{\partial u}{\partial b_{1j_2}} (b_{1j_1} + b_{2j_1}) \right] \sigma_{b_{1k_2} b_{1j2}} \right) \right) \\ &\left. + \sum_{k=1}^{K-1} \left(\left[-1 + \frac{\partial u}{\partial b_{2k_2}} (b_{1k_1} + b_{2k_1}) \right] \sum_{j=1}^{K-1} \left(\left[-1 + \frac{\partial u}{\partial b_{2j_2}} (b_{1j_1} + b_{2j_1}) \right] \sigma_{b_{2k_2} b_{2j_2}} \right) \right) \right\}^{1/2}.\end{aligned} \tag{5.57}$$

5.6.3 Two Common Test-Takers

The standard errors for the estimation of the linking parameters for a design with a pair of common test-takers for the polytomous models are identical to those for the dichotomous 2PL model presented in Barrett and van der Linden (2015a).

5.6.4 Multiple Common Items or Test-Takers

Linking designs usually have multiple common elements, which then have different estimates of the same linking parameters u and v for each element $m = 1, ..., M$. Although we had to deal with multiple dependent estimators *within* each of these elements in the previous section, *across* elements the estimators are independent. As we now know the standard errors for each of these independent estimators, we can pool them using their precision-weighted average. Estimators of linking parameters for a minimal linking element are denoted $\widehat{u}_m$ and $\widehat{v}_m$, and standard errors denoted as σ_{u_m} and σ_{v_m}, respectively.

The pooled estimator of u then becomes

$$\widehat{u} = \frac{\left(\sum_{m=1}^{M} \frac{\widehat{u}_m}{\widehat{\sigma}^2_{u_m}} \right)}{\left(\sum_{m=1}^{M} \frac{1}{\widehat{\sigma}^2_{u_m}} \right)} \tag{5.58}$$

with estimated standard error

$$\widehat{\sigma}_u = \frac{1}{\left(\sum_{m=1}^{M} \frac{1}{\widehat{\sigma}^2_{u_m}} \right)^{1/2}}, \tag{5.59}$$

where the latter is obtained by substituting estimates for the parameters in the right-hand side of (5.41) or (5.56). The pooled estimator and standard error of v are defined entirely analogously.

5.7 Empirical Examples

The empirical examples in this section allowed us to study the properties of the precision-weighted estimators and their standard errors for polytomous items in (5.58) and (5.59). As it is unusual for testing companies to link through common test-takers due possible to memory effects or logistic considerations, we only present empirical examples of common-item designs. Also, comparisons between the errors for the precision-weighted method and those for the current test-characteristic (Stocking & Lord, 1983; Muraki & Chang, 1994) and item-response curve methods (Haebara, 1980) remain reserved for future research. In fact, as the linking functions for the former are not explicit functions of the common item parameters, and they are calculated directly from all category-response functions for a specific polytomous model,

calculation of their asymptotic standard errors may be quite complicated. At least, while Ogasawara (2001) derived the asymptotic standard error for these methods for the 3PL model, we are not aware of any extensions of the same type of analysis to polytomous models in the literature. Finally, use of the test-characteristic curve method precludes the derivation of the standard errors for the minimal linking elements in the design, eliminating the opportunity to weigh the linking estimates by statistical precision and select the best linking items using optimal design principles (Barrett & van der Linden, 2015b).

Empirical data from two real-world testing programs were used for these examples. Both programs administered a different test form once a year, with common items in subsequent forms for linking purposes. While the forms included both dichotomous and polytomous common items, the polytomous items alone were used for the purposes of these examples. The first program was a high-school mathematics exam that included four common polytomous items between the two forms used in the study; 36,963 test-takers responded to the first form, and 37,740 test-takers responded to the second. The first item had four possible response categories, while the other three had five response categories. The second program was a high-school reading exam with three common polytomous items between the forms, each with four possible response categories; 32,931 test-takers responded to the first form and 36,157 test-takers to the second.

The *MIRT Scaling Program*, version 1.01, (Glas, 2010) with a modification to export the complete information matrix for polytomous items, was used to calibrate both test forms. MML estimation for the GPCM in (5.16) and the common identifiability restriction of a normal theta distribution with $\mu_\theta = 0$ and $\sigma_\theta = 1$ were used during each calibration. The estimated item parameters and covariance matrices for all test forms are included in Tables 5.1–5.8. In addition, plots of the estimated step and category-response functions for each item are shown in Figure 5.3–5.9. The reader will note that some items performed better than others, and that for a few of them (Mathematics Items 1 and 4 and Reading Item 1), the response functions for some of categories were quite dominated by others. In addition, note that for the reading test, the b_3 parameters for Item 1 and 3 in the first calibration were close to one another, while the one for the first item was larger than for the second item in the first calibration but the opposite was true in the second calibration. As will be seen below, these observations impacted the linking results for the method with the pairs of items.

To explore the behavior of the standard errors for the overall estimates as items are added to the linking design, we first considered the case of adding identical items to the design for the linking functions in (5.23) and (5.29). For each of the common items in our empirical data,

Table 5.1. *High-school mathematics exam GPCM item parameters and standard errors for* $t = 1$

Item	$\widehat{a}$	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{b}_4$	$\widehat{\sigma}_a$	$\widehat{\sigma}_{b_1}$	$\widehat{\sigma}_{b_2}$	$\widehat{\sigma}_{b_3}$	$\widehat{\sigma}_{b_4}$
1	1.145	0.379	1.254	-0.815	–	0.023	0.069	0.052	0.019	–
2	1.413	0.802	0.941	–	–	0.024	0.032	0.019	–	–
3	1.232	1.337	2.671	–	–	0.023	0.062	0.045	–	–
4	0.723	0.648	-0.495	2.069	-0.209	0.012	0.104	0.083	0.067	0.032

Table 5.2. *High-school mathematics exam GPCM parameter covariance for* $t = 1$

Item	$\widehat{\sigma}_{ab_1}$	$\widehat{\sigma}_{ab_2}$	$\widehat{\sigma}_{ab_3}$	$\widehat{\sigma}_{ab_4}$	$\widehat{\sigma}_{b_1b_2}$	$\widehat{\sigma}_{b_1b_3}$	$\widehat{\sigma}_{b_1b_4}$	$\widehat{\sigma}_{b_2b_3}$	$\widehat{\sigma}_{b_2b_4}$	$\widehat{\sigma}_{b_3b_4}$
1	-0.001	-0.001	0.000	–	0.004	0.001	–	0.001	–	–
2	-0.001	0.000	–	–	0.001	–	–	–	–	–
3	-0.001	-0.001	–	–	0.003	–	–	–	–	–
4	-0.001	-0.001	0.000	0.000	0.008	0.007	0.002	0.005	0.002	0.002

Table 5.3. *High-school mathematics exam GPCM item parameters and standard errors for* $t = 2$

Item	$\widehat{a}$	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{b}_4$	$\widehat{\sigma}_a$	$\widehat{\sigma}_{b_1}$	$\widehat{\sigma}_{b_2}$	$\widehat{\sigma}_{b_3}$	$\widehat{\sigma}_{b_4}$
1	1.164	0.023	1.181	-0.848	–	0.024	0.060	0.047	0.018	–
2	1.435	0.758	1.006	–	–	0.025	0.033	0.019	–	–
3	1.060	1.537	2.820	–	–	0.020	0.076	0.054	–	–
4	0.666	0.489	-0.659	2.272	-0.167	0.011	0.104	0.083	0.068	0.033

Table 5.4. *High-school mathematics exam GPCM parameter covariance for* $t = 2$

Item	$\widehat{\sigma}_{ab_1}$	$\widehat{\sigma}_{ab_2}$	$\widehat{\sigma}_{ab_3}$	$\widehat{\sigma}_{ab_4}$	$\widehat{\sigma}_{b_1b_2}$	$\widehat{\sigma}_{b_1b_3}$	$\widehat{\sigma}_{b_1b_4}$	$\widehat{\sigma}_{b_2b_3}$	$\widehat{\sigma}_{b_2b_4}$	$\widehat{\sigma}_{b_3b_4}$
1	-0.001	-0.001	0.000	–	0.003	0.001	–	0.001	–	–
2	-0.001	0.000	–	–	0.001	–	–	–	–	–
3	-0.001	-0.001	–	–	0.004	–	–	–	–	–
4	-0.001	0.000	0.000	0.000	0.008	0.007	0.003	0.006	0.002	0.002

Table 5.5. *High-school reading exam GPCM item parameters and standard errors for* $t = 1$

Item	$\widehat{a}$	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{\sigma}_a$	$\widehat{\sigma}_{b_1}$	$\widehat{\sigma}_{b_2}$	$\widehat{\sigma}_{b_3}$
1	0.369	0.081	-0.698	1.482	0.008	0.115	0.081	0.052
2	1.144	-1.553	-0.181	1.378	0.022	0.062	0.045	0.025
3	1.588	-0.923	0.089	1.483	0.041	0.044	0.034	0.022

Table 5.6. *High-school reading exam GPCM parameter covariance for* $t = 1$

Item	$\widehat{\sigma}_{ab_1}$	$\widehat{\sigma}_{ab_2}$	$\widehat{\sigma}_{ab_3}$	$\widehat{\sigma}_{b_1b_2}$	$\widehat{\sigma}_{b_1b_3}$	$\widehat{\sigma}_{b_2b_3}$
1	0.000	0.000	0.000	0.009	0.005	0.004
2	0.000	0.000	0.000	0.003	0.001	0.001
3	0.000	0.000	0.000	0.001	0.001	0.001

Table 5.7. *High-school reading exam GPCM item parameters and standard errors for* $t = 2$

Item	$\widehat{a}$	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{\sigma}_a$	$\widehat{\sigma}_{b_1}$	$\widehat{\sigma}_{b_2}$	$\widehat{\sigma}_{b_3}$
1	0.373	-0.215	-0.652	1.140	0.009	0.108	0.077	0.046
2	0.923	-2.278	-0.133	1.425	0.020	0.090	0.061	0.031
3	1.167	-1.223	-0.302	1.772	0.030	0.056	0.041	0.026

Table 5.8. *High-school reading exam GPCM parameter covariance for* $t = 2$

Item	$\widehat{\sigma}_{ab_1}$	$\widehat{\sigma}_{ab_2}$	$\widehat{\sigma}_{ab_3}$	$\widehat{\sigma}_{b_1b_2}$	$\widehat{\sigma}_{b_1b_3}$	$\widehat{\sigma}_{b_2b_3}$
1	0.000	0.000	0.000	0.008	0.004	0.003
2	0.001	0.001	0.000	0.005	0.002	0.002
3	0.000	0.000	0.000	0.002	0.001	0.001

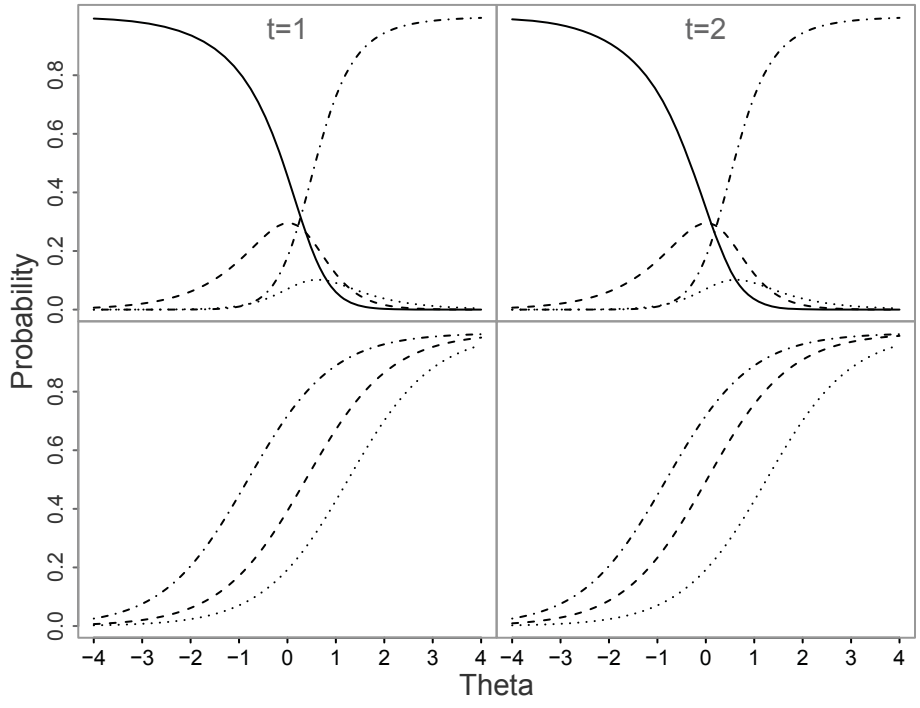

Figure 5.3. Response curves and step functions for the first high-school mathematics linking item. Response curves are in the upper panes and step functions in the lower panes, while the results from the first calibration and second calibration are in the left-hand and right-hand panes, respectively.

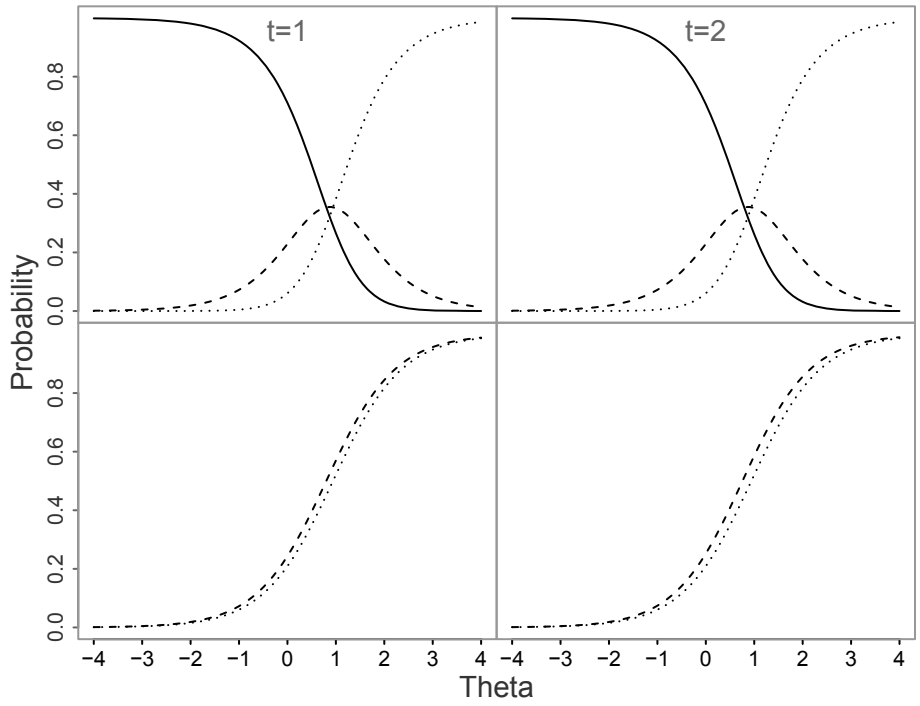

Figure 5.4. Response curves and step functions for the second high-school mathematics linking item.

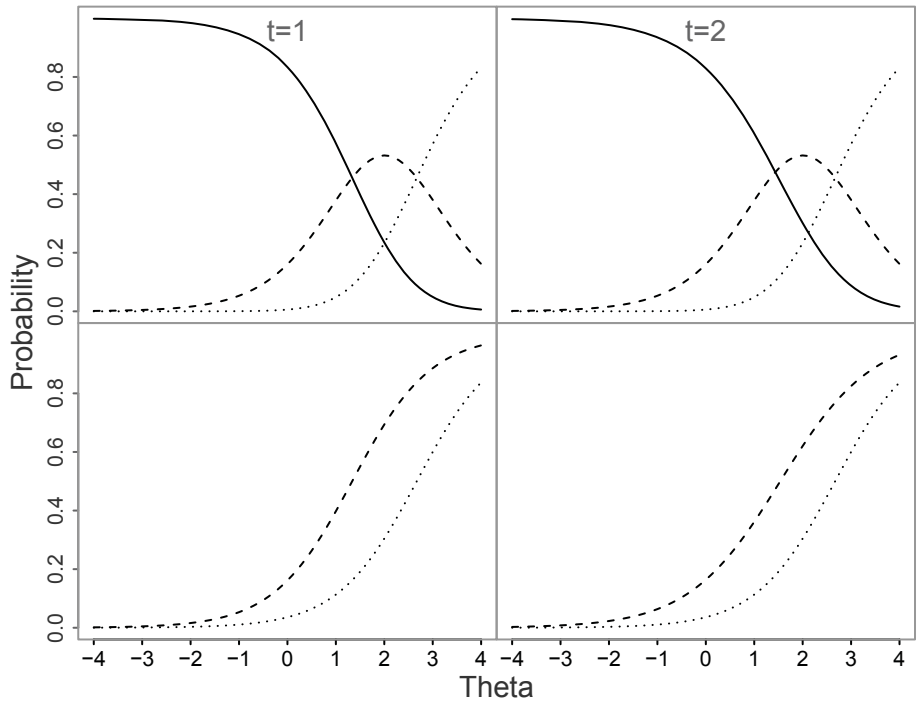

Figure 5.5. Response curves and step functions for the third high-school mathematics linking item.

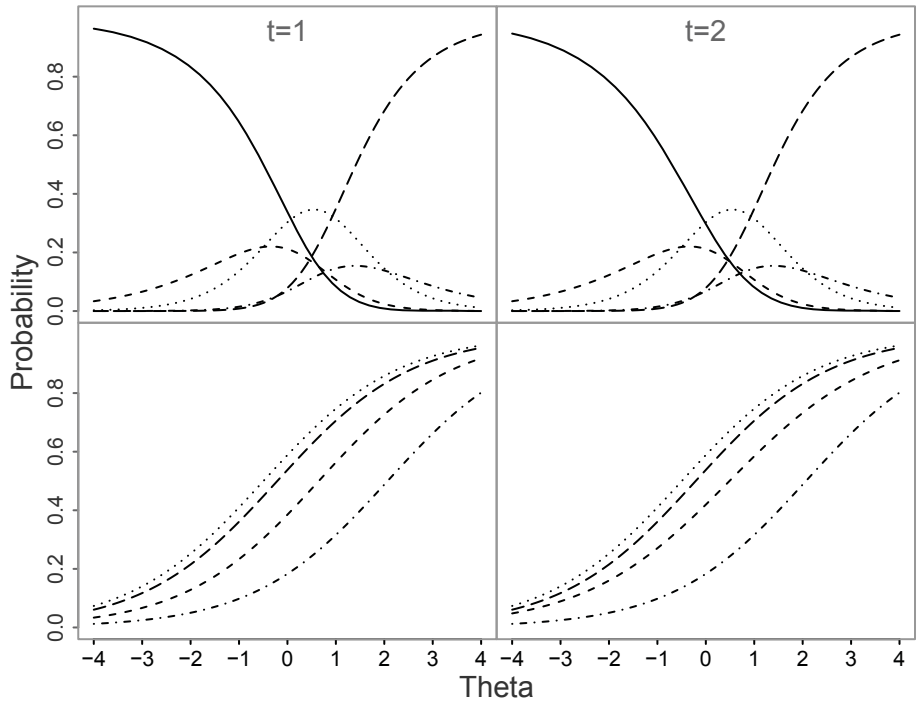

Figure 5.6. Response curves and step functions for the fourth high-school mathematics linking item.

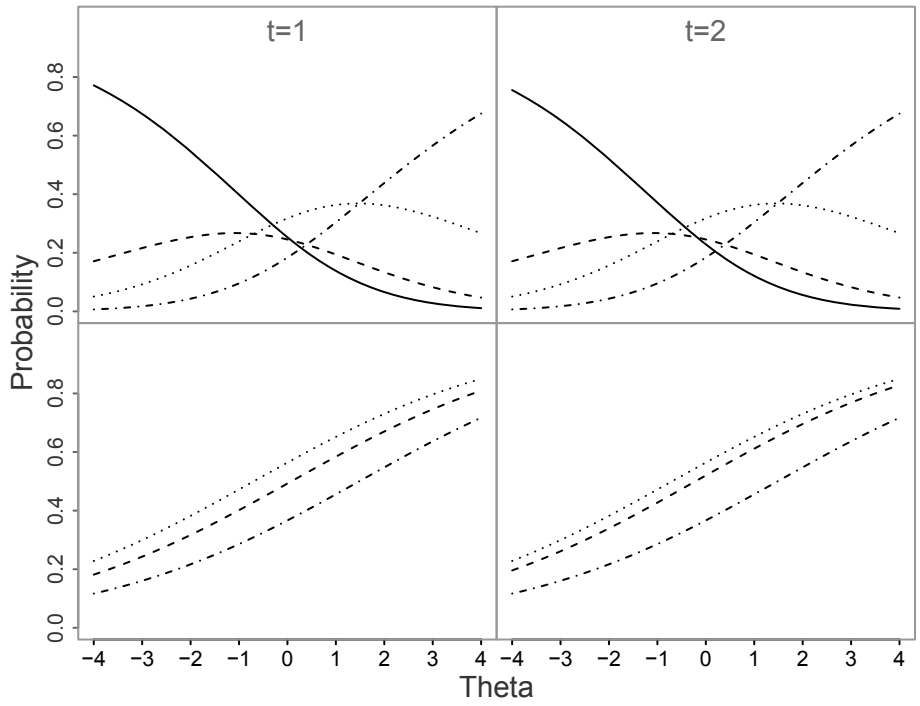

Figure 5.7. Response curves and step functions for the first high-school reading linking item.

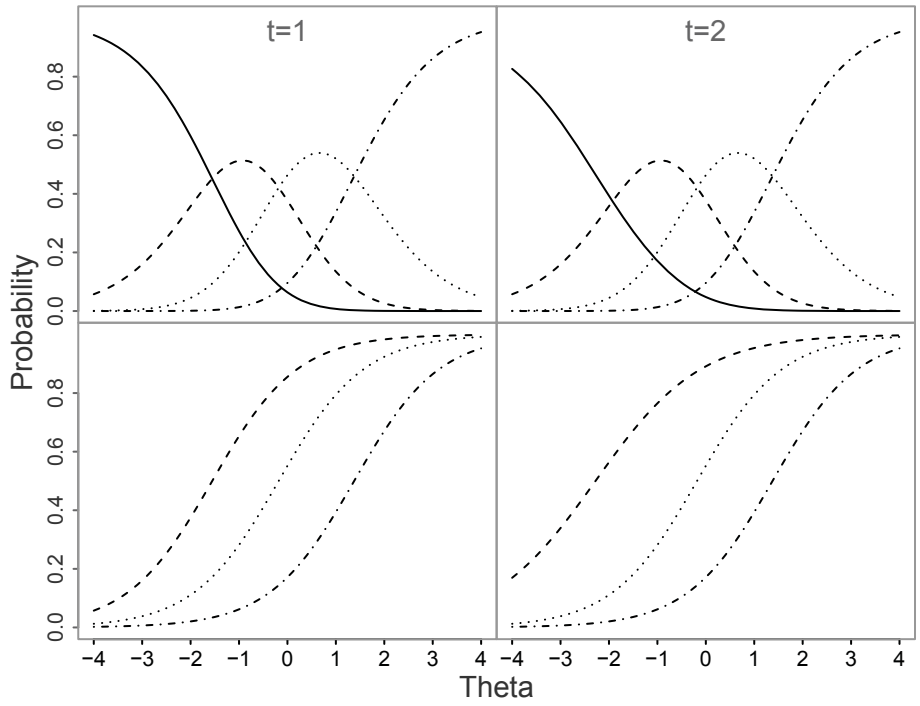

Figure 5.8. Response curves and step functions for the second high-school reading linking item.

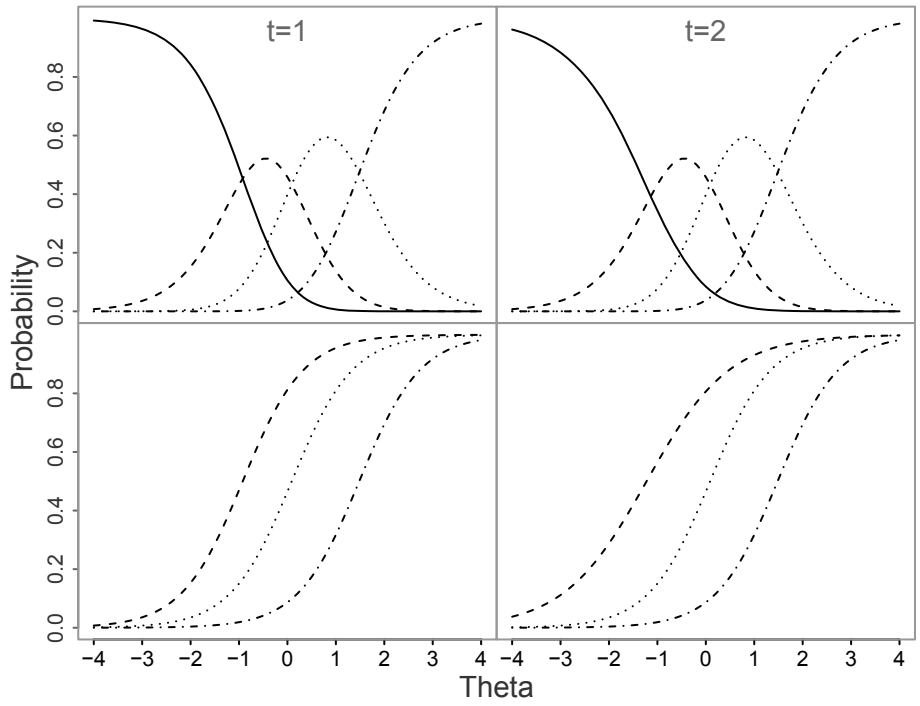

Figure 5.9. Response curves and step functions for the third high-school reading linking item.

we computed the linking estimates and their associated standard errors (Table 5.9). Then, we plotted the error that would occur if more identical elements were added to the linking design. The results are shown in Figure 5.10 and 5.10. As expected, linking error decreased monotonically in all cases. Also, for each of the items, the standard error for the estimate of u was less than that for v. And, similar to the observation we were able to make for the case of dichotomous items (Barrett & van der Linden, 2015a), the size of the errors halved after addition of about five common items and then decreased much more slowly with the additional items.

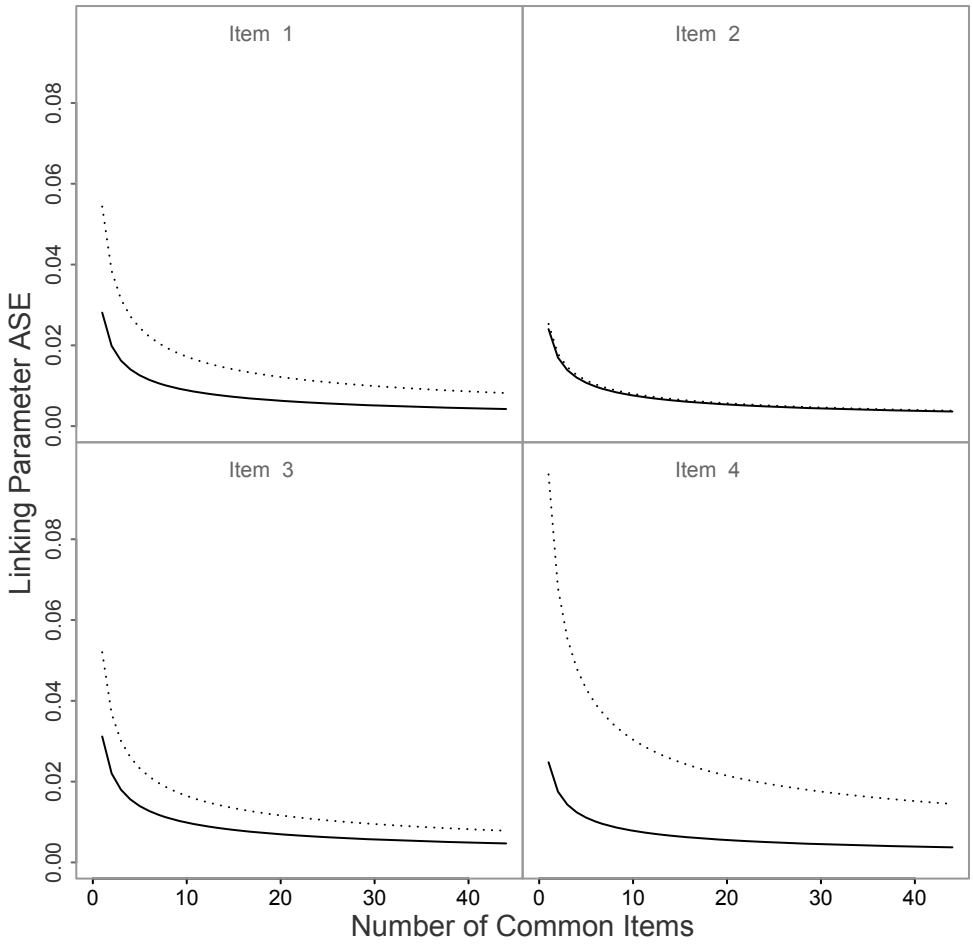

Figure 5.10. ASE of precision-weighted linking parameter estimates u (solid) and v (dash) as a function of the number of identical high-school mathematics polytomous items in single-item linking elements.

Next, to examine the case of estimating linking parameters derived from pairs of common items, as in (5.43) and (5.44), we computed their estimates and associated standard errors for the available item pairs. The results are given in Table 5.9. As the dataset had a limited number of pairs with the same number of response categories, we conducted the analysis for

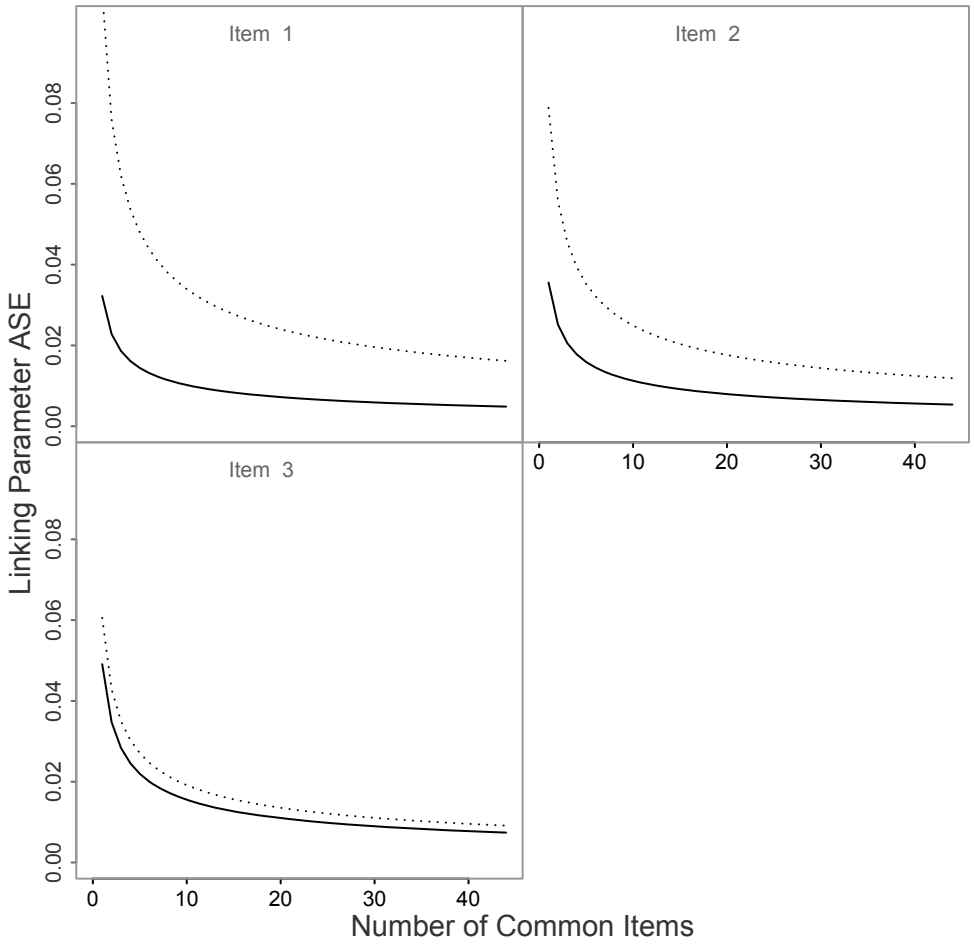

Figure 5.11. ASE of precision-weighted linking parameter estimates u (solid) and v (dash) as a function of the number of identical high-school reading polytomous items in single-item linking elements.

the common pair of items with four possible response categories in the mathematics test (Item 2 and 3) and three possible common pairs of items in the language arts test (the reader should note that the latter are not independent.) The results allowed us to explore the impact of different item parameter estimates on the estimates of the linking parameters and gain insight into the sensitivity of the method to the choice of item pairs. For example, note the estimates for the pair of Reading Items 1 and 3; denominator $\widehat{b}_{1_{3_1}} - \widehat{b}_{2_{3_1}}$ in the estimate of u_m is extremely small ($1.482 - 1.483 = -0.001$), causing it to be unreasonably large. As another example, note how the reversal of the b_3 estimates for the pair of Reading Items 1 and 2 resulted in a negative estimate of u_m. While the reversal is likely to be a result of low discrimination and therefore higher error in the b parameter estimates, it caused an unreasonable result for the linking method based on pairs.

Table 5.9. *Linking parameter estimates and standard error for minimal linking elements*

Exam	Design	Element	$\widehat{u_m}$	$\widehat{\sigma}_{u_m}$	$\widehat{v_m}$	$\widehat{\sigma}_{v_m}$
HS MA	Item	1	0.984	0.028	-0.150	0.054
		2	0.985	0.024	0.024	0.025
		3	1.162	0.031	-0.151	0.052
		4	1.086	0.025	-0.063	0.096
	Paired-item	2,3	1.252	0.144	-0.195	0.204
HS RD	Item	1	0.989	0.032	-0.194	0.107
		2	1.239	0.036	-0.182	0.079
		3	1.361	0.049	-0.212	0.061
	Paired-item	1,2	-0.158	0.552	-0.105	0.975
		2,3	1.451	0.454	-0.194	1.091
		1,3	211.150	11876.973	-53.193	11850.633

To examine the behavior of the standard errors as item pairs are added to the linking design, we sequentially added the same item pair to the design and plotted the errors. The results are given in Figure 5.12–5.13. Again, note that Item 2 is included in both panes of Figure 5.13. The pair with Item 1 and 3 was excluded from this figure for the reason discussed above.

Again, as expected, the linking error decreased monotonically in all cases. For all item pairs, the error in the estimate of u was smaller than that of v. And, similar to the observation above, the error halved after about five common pairs and then decreased more slowly with the additional pairs. The reader should note that, for the same set of items, the error for the precision-weighed method based on linking elements of common pairs of items was considerably larger than the error for the method based on linking elements of single items. For example, for each of Items 2 and 3 of the mathematics test in Figure 5.10 the error was only a fraction of that for the pair of same items in Figure 5.12. The same results were obtained for the case of dichotomous items by Barrett and van der Linden (2015a), which these authors explained by pointing out that the pair-based method with the solutions for u and v in (5.25)–(5.26) ignores the unique information in the item discrimination parameters. However, due to the small number of pairs possible in these empirical datasets, additional verification of this result is necessary.

Finally, we plotted the standard errors for the linking design with single item linking elements using all polytomous common items available for the linking of these two tests, adding them to the design in the order they appeared in the first test. The results are given in Table 5.10 and plotted in Figures 5.14–5.15. As expected, reduction of the linking error with additional items was noted, and the magnitude of the error seemed reasonable.

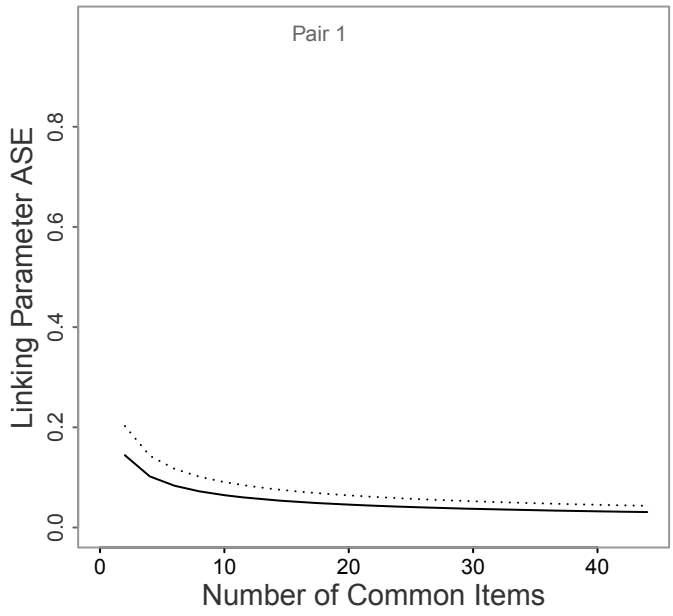

Figure 5.12. ASE of precision-weighted linking parameter estimates u (solid) and v (dash) as a function of the number of identical high-school mathematics polytomous items in paired-item linking elements.

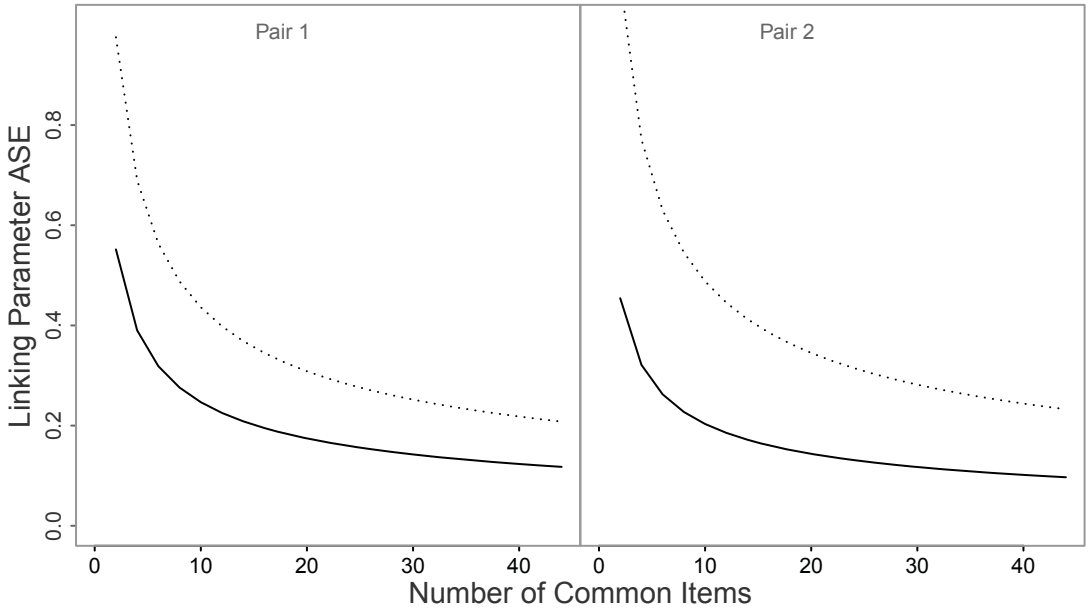

Figure 5.13. ASE of precision-weighted linking parameter estimates u (solid) and v (dash) as a function of the number of identical high-school reading polytomous items in paired-item linking elements.

Table 5.10. *Linking parameter estimates and standard error as minimal linking elements of single common items are added to linking design*

Exam	Number of Items	$\widehat{u}$	$\widehat{\sigma}_u$	$\widehat{v}$	$\widehat{\sigma}_v$
HS MA	1	0.984	0.028	-0.150	0.054
	2	0.984	0.018	-0.007	0.023
	3	1.030	0.016	-0.030	0.021
	4	1.046	0.013	-0.032	0.021
HS RD	1	0.989	0.032	-0.194	0.107
	2	1.102	0.024	-0.186	0.064
	3	1.152	0.021	-0.200	0.044

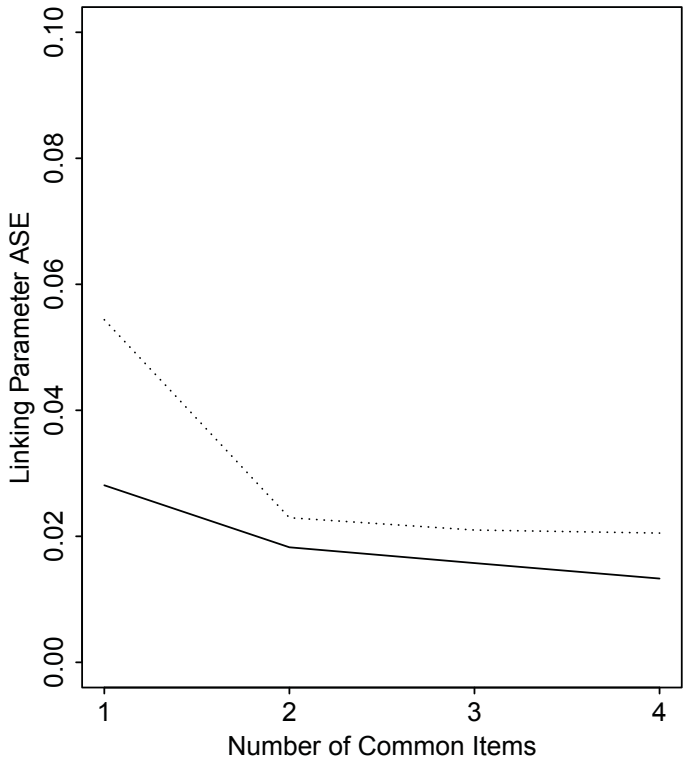

Figure 5.14. ASE of precision-weighted linking parameter estimates u (solid) and v (dash) for high-school mathematics exam as common polytomous items are added to the linking design.

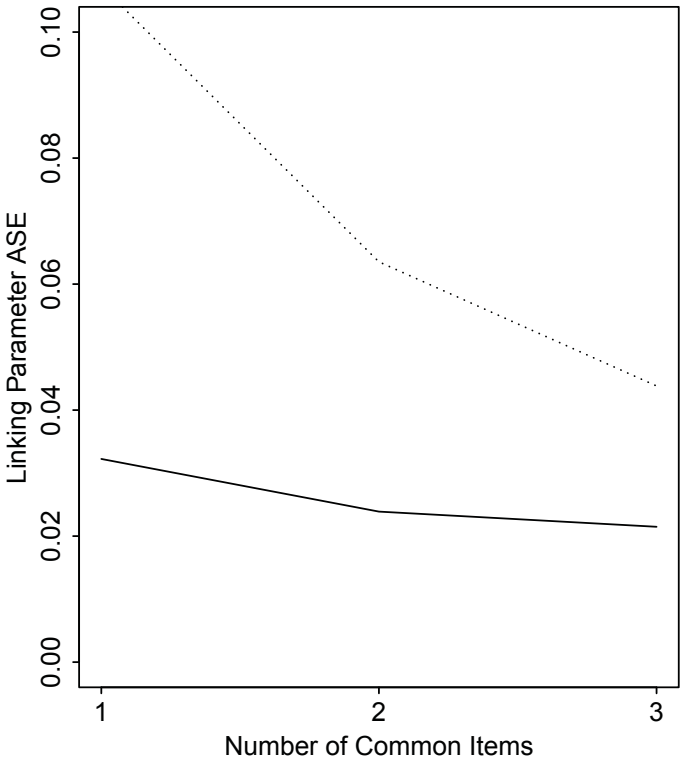

Figure 5.15. ASE of precision-weighted linking parameter estimates u (solid) and v (dash) for high-school reading exam as common polytomous items are added to the linking design.

A similar analysis was not possible for the design with item pairs as there was only one unique pair with the same number of response categories per test. The limitation of this type of linking design to unique pairs of items with the same numbers of response categories seriously reduces its utility.

5.8 Concluding Remarks

Two important findings resulted in the course of this work. First, two test forms with polytomous items can be linked just using any one of their response categories. The best way to deal with this unexpected richness is to average across all categories and calculate the standard error for the average. Second, because of their common step functions, the same linking function may be used for each of the nominal, partial credit, sequential response and graded response models, provided each of them has the same type parameterization. In fact, we found that when using

$a_i(\theta_p - b_{i_k})$ as our parameterization, the linking of polytomous response model parameters was a generalization of the linking the parameters for the dichotomous logistic models, albeit with more complicated calculation of the standard errors due to statistical dependence between the estimates of the response category parameters for the same item.

The fact that the same type of linking function may be used for both types of items allows us to compare the relative contribution to linking error from dichotomous and polytomous items in the same linking design in future research. As a result of our current study, we already know that possible differences can be due only due to the statistical properties of the linking parameters estimates.

In addition, to our knowledge, while test-characteristic curve and item-response function methods are used in practice to link with common polytomous items, to date no closed-form expressions for their linking error estimates have been derived; linking error is not typically reported. We therefore find our research and results promising, in that they may increase transparency of linking methods and linking error when polytomous items are included in linking designs.

The research presented in this paper assumed only random linking error due to response model parameter estimation. In practice, systematic error may also be present. For example, there may be flaws in implementation of linking designs or a lack of model fit for some items or test-takers. Therefore, it is important that linking designs are robust, which should be an additional consideration when selecting a design. We also acknowledge that this research did not address the case of the randomly-equivalent group linking design, which includes sampling error in addition to parameter estimation error.

Acknowledgment

The authors are indebted to Cees A. W. Glas for his support during their use of his *MIRT Scaling Program* (version 1.01) software.

References

Agresti, A. (2013). *Categorical data analysis*. New Jersey: John Wiley & Sons.

Baker, F. B. (1992). Equating under the graded response model. *Applied Psychological Measurement, 16*, 87–96.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17*, 239–251.

Baldwin, P. (2011). A strategy for developing a common metric in item response theory when parameter posterior distributions are known. *Journal of Educational Measurement, 48*, 1–11.

Barrett, M. D., & van der Linden, W. J. (2015a). *Estimating linking functions for response model parameters.* Manuscript submitted for publication.

Barrett, M. D., & van der Linden, W. J. (2015b). *Optimal linking design for response model parameters.* Manuscript submitted for publication.

Bartels, R. (1985). Identification in econometrics. *The American Statistician, 39*, 102–104.

Bechger, T. M., Verhelst, N. D., & Verstralen, H. H. F. M. (2001). Identifiability of nonlinear logistic models. *Psychometrika, 66*, 357–372.

Bechger, T. M., Verstralen, H. H. F. M., Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika, 67*, 123–136.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33–49). New York: Springer.

Chang, H. & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59*, 391–404.

Childs, R .A, & Chen, W. (1999). Software Note: Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement, 23*, 371–379.

Cohen, A. S., & Kim, S.-H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22*, 116–130.

Fischer, G. H. (2004). Remarks on "Equivalent linear logistic test models" by Bechger, Verstralen, and Verhelst (2002). *Psychometrika, 69*, 305–315.

Glas, C. A. W. (2010). *Multidimensional item response theory* (MIRT 1.0) [Computer software and manual]. Retrieved October 27, 2014, from `http://www.utwente.nl/gw/omd/Medewerkers/medewerkers/glas/`.

Kim, J. S., & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement, 26,* 255–270.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed-format texts. *Journal of Educational Measurement, 43*, 53–76.

Koenig, J. A., & Roberts, J. S. (2007). Linking parameters estimated with the generalized graded unfolding model: a comparison of the accuracy of characteristic curve methods. *Applied Psychological Measurement, 31*, 504–523.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices (2nd ed.).* New York: Springer.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Maris, G. (2002). *Concerning the identification of the 3PL model* (Measurement and Research Department Reports 2002-3). Arnhem, The Netherlands: Cito.

Maris, G., & Bechger, T. (2004). Equivalent MIRD models. *Psychometrika, 69*, 627–239.

Maris, G., & Bechger, T. (2009). On interpreting model parameters for the three-parameter logistic model. *Measurement: Interdisciplinary Research and Perspectives, 7*, 75–88.

Masters, G. M. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–173.

Masters, G. M., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York: Springer.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164) New York: Springer.

Muraki, E., & Chang, H. (1994). *Horizontal and vertical test equating methods based on the generalized partial credit model*. (ETS Internal Report). Princeton NJ: Educational Testing Service.

Muraki, E., Hombo, C. M., & Lee, Y. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*, 325–337.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage Publications.

Penfield, R. D. (2014). An NCME Instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice, 33*, 36–48.

Revuelta, J. (2009). Identifiability and Equivalence of GLLIRM Models. *Psychometrika*, *74*, 257-272.

Reiersol, O. (1950). On the identifiability of parameters in Thurstone's multiple factor analysis. *Psychometrika, 15*, 121–149.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, *39*, 577–591.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100) New York: Springer.

San Martín, E., González, J., & Tuerlinckz, F. (2009). Identified parameters, parameters of interest and their relationships. *Measurement: Interdisciplinary Research and Perspective, 7*, 97–105.

San Martín, E., González, J., & Tuerlinckz, F. (in press). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*. DOI: 10.1007/s11336-014-9494-2.

San Martín, E., Jara, A., Ronlin, J.-M., & Mouchart, M. (2011). On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika, 76*, 341–379.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49,* 501–519.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer.

Tsai, R.-C. (2000). Remarks on the identifiability of the Thurstonian ranking models: Case V, Case III, or neither? *Psychometrika, 65*, 201–210.

Tutz, G. (1997). Sequential models for ordered responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139–152) New York: Springer.

van der Linden, W. J., & Barrett, M. D. (in press). Linking item response model parameters. *Psychometrika*. doi: 20.1007/s11336-015-9469-6.

Verhelst, N.D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). New York: Springer.

Verhelst, N. D., & Verstralen, H. H. F. M. (2008). Some considerations on the partial credit model. *Psicologica, 29*, 229–254.

Volodin, N., & Adams, R. J. (2002). *The estimation of polytomous item response models with many dimensions* (Internal Report). Parkville, Victoria, Australia: University of Melbourne, Faculty of Education.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology, 3*, 115–124.

Wan, L. & Henly, G. A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education, 25*, 58–78.

Chapter 6

Optimal Linking Design with Response Models for Mixed-Format Tests[1]

6.1 Introduction

Item response models explain the probabilities with which test-takers respond to an item using separate ability and item parameters. The models are not identified in that more than one set of values for the ability and item parameters can imply the same probabilities. During parameter estimation, identifiability restrictions are imposed to select one set of values from these multiple observationally equivalent sets. When the item or ability parameters from two calibration studies are to be compared, it is necessary to map the parameters from one calibration onto the other to adjust for differences in identifiability restrictions using a linking function. In practice, linking functions are obtained from common item and/or ability parameters with estimation error propagating into linking error. This paper investigates the relative contributions to linking error by common items with varying numbers of response categories.

Several estimators of linking function parameters have been proposed and researched, including the mean/mean (Loyd & Hoover, 1980), the mean/sigma (Marco, 1977), the item characteristic curve (Haebara, 1980), the test characteristic curve (Stocking & Lord, 1983), and more recently, the precision-weighted average estimators (van der Linden and Barrett, in press). Ogasawara (2000; 2001; 2011) presented the asymptotic standard errors (ASEs) for the first four of these estimators for the dichotomous logistic response models. Barrett and van der Linden (2015a,

[1]Barrett, M. D. & van der Linden, W. J. (2015). *Optimal linking design with response models for mixed-format tests.* Manuscript to be submitted for publication.

2015c) presented the ASEs for the precision-weighted estimators for dichotomous and polytomous response models. They also demonstrated their ASEs to be smaller than those for the mean/mean and mean/sigma approaches because of the lower weights for the contributions by items with a larger estimation error. In addition, because these contributions are explicit, these estimators allow for the application of optimal linking design principles as illustrated by Barrett and van der Linden (2015b).

When dichotomous and polytomous common items are both included in a linking study, important practical questions remain though: How does the contribution to linking error for an item with more response categories compare to that of one with fewer categories? Or more specifically, what precise factors determine when one type of item has a smaller contribution to the standard error of linking than the other? And, in an optimal linking design study, what types of items will generally be preferred?

In accordance with the existing literature, we refer to assessments that combine items with dichotomous and polytomous response formats as "mixed-format" assessments. However, the term stems from a time when the only response formats available were selected response and constructed response. Since then, a variety of additional formats has been introduced due to advances in user interface and computing technologies, often referred to as "technology-enhanced" items. Technology-enhanced items may have two response categories, similar to dichotomously scored multiple-choice items, or they may have more complicated rubrics that are scored in multiple categories, therefore calling for the application of polytomous response models (e.g., Wan & Henly, 2012). We will use "mixed-format" to simply indicate a test that has items scored with both dichotomous and polytomous response models, rather than a substantive combination of response formats (e.g., free text and selected response).

Previous research on linking of mixed-format assessments tended to focus mainly on two areas: possible violations of unidimensionality assumptions and the impact of rater instability with constructed responses. Only a few earlier papers presented empirical results from comparisons of different linking methods with mixed-format tests.

6.1.1 Possible Violation of Unidimensionality Assumptions

For traditional observed-score equating, studies on the impact of mixed-format items and possible multidimensionality stemming from mixed-format item types on the equating function have been conducted by several authors. These studies examine, for example, the equating

functions resulting from tests with varying degrees of correlation among scores on the multiple-choice and constructed-response items (Hagge & Kolen, 2011; Lee et al., 2012), the impact of the presence of subgroups with differential performance on the different item formats (Kim & Walker, 2012), and the impact of test length and different combinations of selected response and constructed response items on the accuracy of the equating functions (Fitzpatrick & Yen, 2001; Kim, Walker & McHale, 2010a). Fitzpatrick and Yen (2001) found that given the same total number of response categories in a test, using more items with fewer categories rather than fewer items with more categories resulted in more accurate equating. However, as the study was not originally designed to examine this question, the conclusions may be limited to a single empirical dataset in which this effect was observed.

Within the IRT framework, studies of the impact of mixed formats on unidimensionality indicate that for many mixed-format tests, unidimensionality assumptions are appropriate. Ercikan et al (1998) presented empirical examples of calibration with 3PL models alone, 2PPC models alone, and a mix of 3PL and 2PPC models, for tests in which both item types were intended to measure the same construct. They found that impact on item fit was minimal and concluded that simultaneous calibration of both item types for their empirical data was appropriate. Wang, Lee, and Kolen (2012) also found evidence that an assumption of unidimensionality was appropriate for mixed item formats, and that item fit was not negatively impacted by the introduction of mixed item formats on several *Advanced Placement* examinations.

As the presence of items with mixed format does not necessarily appear to lead to violations of unidimensionality, our current goal of studying the properties of linking functions for a mixture of unidimensional response models seems relevant.

6.1.2 Possible Item-Rater Instability

A concern raised in relation to the use of human-scored constructed-response items in linking is the introduction of systematic error in the item scores due to rater drift that then propagates into the item parameter estimates. In an observed-score equating context, some authors conducted trend analyses as part of their investigation (Kim & Walker, 2012; Kim, Walker, and McHale, 2010a, 2010b). As expected, in cases where no substantial scoring difference between the two sets of raters were found, there was little impact of rater drift but in cases where rater drift was prevalent, the accuracy of the equating function was compromised. In an IRT context, a procedure to take error due to rater instability in conjunction with the mean/sigma linking

approach has been proposed (Kamata and Tate, 2005; Tate, 2000, 2003). They concluded that the procedure held promise, although it has not been widely adopted.

While it is possible for differential rating to occur across test administrations, for the purpose of the current study, it was assumed that scoring of constructed response items is consistent from one administration to the next. This is not unrealistic, as with the recent introduction of automated scoring of technology-enhanced polytomous items and essays, rater drift may cease to be a concern.

6.1.3 Empirical Comparisons of Linking Methods for Mixed-Format Tests

A empirical comparison of mean/sigma linking with fixed common item and concurrent parameter estimation procedures for tests with mixed-format items calibrated with the 3PL, 2PL, and graded response models was conducted by Jodoin, Keller and Swaminathan (2003). The authors implemented the mean/sigma approach using the location of the polytomous items on the θ scale along with the b parameter of the dichotomous items and equally weighting of the contributions of both types of items to the linking function. They found systematic differences between the linked polytomous-item locations and the b parameters, and therefore recommended that research be conducted to determine how these locations and parameters should be combined in the current linking methods.

Kim and Lee (2006) presented comparisons of mean/mean, mean/sigma, item characteristic curve, and test-characteristic curve linking for mixed-format tests and concluded that the characteristic-curve methods were superior. However, they highlight several considerations for the mean/mean and mean/sigma methods not addressed by their procedures that may have impacted their results, such as how to weigh the parameters for the polytomous and dichotomous items?, whether and how to account for the identifiability restrictions used during the calibrations when calculating the means and standard deviations of the a_i and b_{ik}?, and how combinations of the $(\alpha\theta - \beta)$ parameterization of polytomous models and the typical $a(\theta - b)$ parameterization of dichotomous models would impact linking results? Kim and Kolen (2006) found that, in the presence of multidimensionality, test-characteristic curve methods outperformed the mean/mean and mean/sigma methods, but noticed that the considerations formulated in the preceding study applied to this study as well.

Yao and Boughton (2009) reported a simulation study that included constructed-response and multiple-choice items calibrated under multidimensional versions of the 2-parameter PCM and the 3PL model, respectively. They found that, for the same total number of response categories, the addition of constructed-response items decreased the root mean square error beyond that of a linking set comprised of multiple-choice items only. As linking method, their study used a multidimensional version the test-characteristic curve method.

Keller and Hambleton (2013) questioned the impact of mixed-format items on parameters linking across multiple administrations with a variety of changes in the shape of the test-taker ability distributions. Specifically, they evaluated the differences between fixed-parameter calibration and test-characteristic curve linking using mis-classification rates as their criterion. However, the notion that linking error is a result of the propagation of item-parameter estimation error was not considered to explain results (but very well could explain why a skewed ability distribution in one of the administrations resulted in higher rates of mis-classification). Nor did the authors compare the relative contributions to linking error between the various item formats in their study.

In general, the research cited above used empirical or simulated datasets to investigate a few issues related to the linking of parameters in response models for mixed item formats. None of them provided a theoretical or analytic approach to the issue. After a review of the precision-weighted linking method, this paper will address the problem of the contributions to linking error as a function of the number of response categories in mixed-format common-item linking analytically, illustrating the results through empirical examples. It will then describe and illustrate how the results can facilitate applications of optimal design principles to mixed-format linking.

6.2 Linking Functions for Dichotomous and Polytomous Models

The 3PL gives the probability of a correct response on an item as

$$\pi_{pi} = c_i + (1 - c_i)\frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}, \tag{6.1}$$

with $a_i > 0, b_i \in \mathbb{R}$, $c_i \in [0, 1]$, and $\theta_p \in \mathbb{R}$. The 2PL results from setting all $c_i = 0$, and the 1PL results from setting all $a_i = 1$ and all $c_i = 0$.

Without any assumptions about their shape, the linking functions for these dichotomous models may be derived as the solution to a functional equation directly based on the response function in (6.1); for details, see van der Linden and Barrett (2015). The solution is

$$\varphi_a(a) = u^{-1}a, \tag{6.2}$$

$$\varphi_b(b) = ub + v, \tag{6.3}$$

$$\varphi_c(c) = c, \tag{6.4}$$

$$\varphi_\theta(\theta) = u\theta + v, \tag{6.5}$$

with

$$u \equiv \frac{\varphi_\theta(\theta) - \varphi_b(b)}{\theta - b}, \theta \neq b \tag{6.6}$$

and

$$v = \varphi_b(b) - ub = \varphi_\theta(\theta) - \theta \tag{6.7}$$

As the linking function for the c parameters is the identity function, these parameters are automatically linked once identified. As a result, for the 3PL model there are no linking parameters over those required for a 2PL model . In addition, linking for the 1PL model is the special case of (6.2)–(6.7) for $u = 1$.

In an extension of their method to polytomous models, Barrett and van der Linden (2015c) demonstrated that the true linking functions for the nominal (Bock, 1972, 1997), generalized partial credit (Masters, 1982; Masters & Wright, 1997; Muraki, 1997), graded (Samejima, 1969, 1997), and sequential response models (Tutz, 1990, 1997; Verhelst, Glas, & de Vries, 1997) are the same special case of those for the 2PL dichotomous response model [provided for each of them the traditional $a_i(\theta_p - b_{ik})$ parameterization is used]. Their argument was based on the equivalence of the problems of linking the parameters in these models and in their underlying step functions. Generally, a step function provides a conditional probability of success by a test-taker $p = 1, ..., P$ at the k^{th} step on item $i = 1, ..., I$ with categories $k = 1, ..., K$. The choice of step function for each of the polytomous models above is

$$\Psi_{pik} = \frac{\exp[a_i(\theta_p - b_{ik})]}{1 + \exp[a_i(\theta_p - b_{ik})]}, \tag{6.8}$$

$a_i > 0$, $b_{ik} \in \mathbb{R}$, $\theta_p \in \mathbb{R}$. Note that the right-hand side of the expression for the step function is the same as that of the response function for the 2PL model, which involves one "step" (from a response category of 0 to 1).

Each of the polytomous models listed above follows from a different definition of the steps. For example, the generalized partial credit model (GPCM) (Muraki, 1997) defines a step as the event of a response in category k rather than $k-1$ and specifies the probability of a response in category k as

$$\begin{aligned}\pi_{pik} &= \frac{\pi_{pik}}{\pi_{pi1}+\pi_{pi2}+...+\pi_{pik}+...+\pi_{pi(K)}} \\ &= \begin{cases} \dfrac{1}{1+\sum_{d=2}^{K}\exp(\sum_{h=2}^{d}a_i(\theta_p-b_{ih}))}, & k=1; \\ \dfrac{\exp(\sum_{h=2}^{k}a_i(\theta_p-b_{ih}))}{1+\sum_{d=2}^{K}\exp(\sum_{h=2}^{d}a_i(\theta_p-b_{ih}))}, & k=2,...,K. \end{cases}\end{aligned} \tag{6.9}$$

This model will be assumed to hold throughout the remainder of this paper.

However, although the true linking function for each of the polytomous models above is simply that of the 2PL step function, the choice of polytomous model does impact the estimation of the step parameters—and hence the estimated linking function.

6.3 Precision-Weighted Average Estimation Method

We begin our review of the method using a linking design with one common item $i=1$; for the general case, see Barrett and van der Linden (2015a, 2015c). For a polytomous item, the linking parameters that are to be estimated are defined as

$$u = \frac{a_{1_1}}{a_{1_2}} \tag{6.10}$$

$$v = (K-1)^{-1}\sum_{k=1}^{K-1}(b_{1k_2}-ub_{1k_1}), \tag{6.11}$$

where a_{1_1} and a_{1_2} are the item-discrimination parameters and b_{1k_1} and b_{1k_2} the difficulty parameters for step k in the first and second calibration, respectively. Plugging estimates of these parameters into (6.10) and (6.11) yields estimators of u and v with standard errors

$$\sigma_{u_i} = \left(\frac{\sigma^2_{a_{1_1}}+u^2\sigma^2_{a_{1_2}}}{a^2_{1_2}}\right)^{1/2}, \tag{6.12}$$

and

$$\begin{aligned}\sigma_{v_i} &= \frac{1}{a_{1_2}(K-1)}\Bigg\{\left(\sum_{k=1}^{K-1} b_{1k_1}\right)^2\left(\sigma^2_{a_{1_1}} + u\sigma^2_{a_{1_2}}\right) \\ &\quad +2a_{1_1}\left(\sum_{k=1}^{K-1} b_{1k_1}\right)\left(\sum_{k=1}^{K-1}\sigma_{a_{1_1}b_{1k_1}} + \sum_{k=1}^{K-1}\sigma_{a_{1_2}b_{1k_2}}\right) \\ &\quad +a^2_{1_2}\left(u^2\sum_{k=1}^{K-1}\sum_{j=1}^{K-1}\sigma_{b_{1j_1}b_{1k_1}} + \sum_{k=1}^{K-1}\sum_{j=1}^{K-1}\sigma_{b_{1j_2}b_{1k_2}}\right)\Bigg\}^{1/2},\end{aligned} \tag{6.13}$$

respectively, where the σs are the elements of covariance matrix of the item parameter estimates.

For the case of multiple common items $i = 1, .., M$, there is an (independent) estimate of u and v for each item. The estimates can be combined using the idea of precision-weighted averaging. The result is

$$\widehat{u} = \frac{\left(\sum_{i=1}^{M}\frac{\widehat{u}_i}{\widehat{\sigma}^2_{u_i}}\right)}{\left(\sum_{i=1}^{M}\frac{1}{\widehat{\sigma}^2_{u_i}}\right)} \tag{6.14}$$

with estimated standard error

$$\widehat{\sigma}_u = \frac{1}{\left(\sum_{i=1}^{M}\frac{1}{\widehat{\sigma}^2_{u_i}}\right)^{1/2}}, \tag{6.15}$$

with an analogous result for v. Because of the choice of weights, the contributions by the common items to the estimator in (6.14) with relatively larger estimation error are automatically reduced. Further, as desired, its standard error in (6.15) decreases monotonically with the number of linking elements (items) added to the linking design.

6.4 Relationship Between Linking Error and Number of Response Categories

Of interest is the relationship between the standard errors of linking for the common items in (6.12) and (6.13) and the number of response categories, K. The expressions for these standard errors reveal several factors with an impact on them. First, σ_{u_i} is dependent on the true values of the discrimination parameters (a_{i_1}, a_{i_2}) in both calibrations as well as their standard error of estimation. In addition, σ_{v_i} is dependent on the sum of the category parameters in the first administration $(\sum_{k=1}^{K-1} b_{ik_1})$, the number of these categories (K), and the full covariance matrix for all item parameters in both calibrations. This matrix, in turn, depends on the ability

distributions of the test-takers, the specific response model used during parameter estimation, as well as its true parameter values for the item.

First consider the impact of K on σ_{u_i}. Assume two items, one with $K = 2$ and one for which $K > 2$, both taken by the same P test-takers. When both the values of discrimination parameter a and its standard error are equal for these two items in the two calibrations, σ_{u_i} will be equal as well regardless of the number of categories, K. If the standard error of a varies with K, one would expect any changes in σ_{u_i} to be commensurate.

Next consider the impact of increasing K on σ_{v_i}. Assume again the same two items taken by the same P test-takers. If in addition the average of the b parameters is equal in both calibrations, it still holds that estimation error in the b parameters as well as in the (co)variances in the expression of σ_{v_i} depends on the distribution of the P test-takers across all response categories. For example, when $K = 2$, all test-takers are either scored 0 or 1, and each of their responses contributes to the estimation of the location of the step between the two response categories. When $K > 2$, however, the test-takers distribute across multiple categories, and only a portion of them may be used to estimate the location of each of the steps. The standard errors for the b parameters increase with the number of response categories K, and so will the standard error for the intercept parameter in the linking function, σ_{v_i}. In addition, an increase in K implies an increase in the size of the covariance matrix for the item, which is also expected to reinforce the increase in σ_{v_i}.

Two important hypotheses follow from our analyses:

1. The standard error of slope parameter u in the linking function hardly changes with the number of response categories of the items.
2. The standard error of intercept parameter v in the linking function generally increases with the number of response categories of the items.

6.4.1 Empirical Results

To further explore the validity of these hypotheses, a simulation study was conducted with the design explained in Table 3.1. The factors in this design are the ones with an expected impact on the linking error in σ_{u_i} (6.12) and σ_{v_i} (6.13) as discussed above. The design resulted in 48 different items (two values for discrimination a parameter; three values for the average

Table 6.1. *Simulation design to study impact of number of response categories on linking error*

a	Average b	g	K	N items
0.5, 1.0	-1.0	0.2	2-5	8
0.5, 1.0	-1.0	0.6	2-5	8
0.5, 1.0	0.0	0.2	2-5	8
0.5, 1.0	0.0	0.6	2-5	8
0.5, 1.0	1.0	0.2	2-5	8
0.5, 1.0	1.0	0.6	2-5	8

of the item's b parameters; two values for the equal distances between the adjacent category parameters b; and four values of K).

Responses on the items were simulated for the GPCM in (6.9) for 5,000 test-takers randomly drawn from a U[-3.0, 3.0] ability distribution. A rather wide uniform distribution was chosen to prevent the occurrence of relatively sparse responses in the outer categories. The *MIRT 1.0* software program (Glas, 2010) was used for parameter estimation with the method of marginal maximal likelihood (MML) estimation and the common identifiability restrictions of a normal theta distribution with $\mu_\theta = 0$ and $\sigma_\theta = 1$. The software produced the covariance matrix for each of the items associated with the parameter estimates. Figures 6.1–6.6 illustrate the standard errors and covariance obtained for each of the 48 items.

Figure 6.1 reveals a small decrease in the standard errors for the a parameters with an increase in K. Also, the items with the higher discrimination parameter generally had slightly higher standard errors overall and exhibited hardly any dependency on K.

Figure 6.2 provides a picture of the standard errors for the b parameters of each of the items. Within each panel, the vertical collections of points represent the standard errors for each of these parameters for a single item. The same pattern was found for all items. Clearly, as K increased, the average standard errors of the b parameters increased. The error in the b parameter for the items with $K = 2$ was always the least. In addition, the standard error for the b parameters was generally higher for items with lower discrimination parameters.

Figures 6.3–6.6 show the covariances between the estimators of the item parameters found in this study. The covariances for the a and b parameters were generally small in magnitude with a similar pattern for most items. Also, these covariances tended to decrease with K (many times into negative values). On the other hand, the covariances for the pairs of b parameters were generally positive and increased with K. As expected, it is easily seen that the standard errors

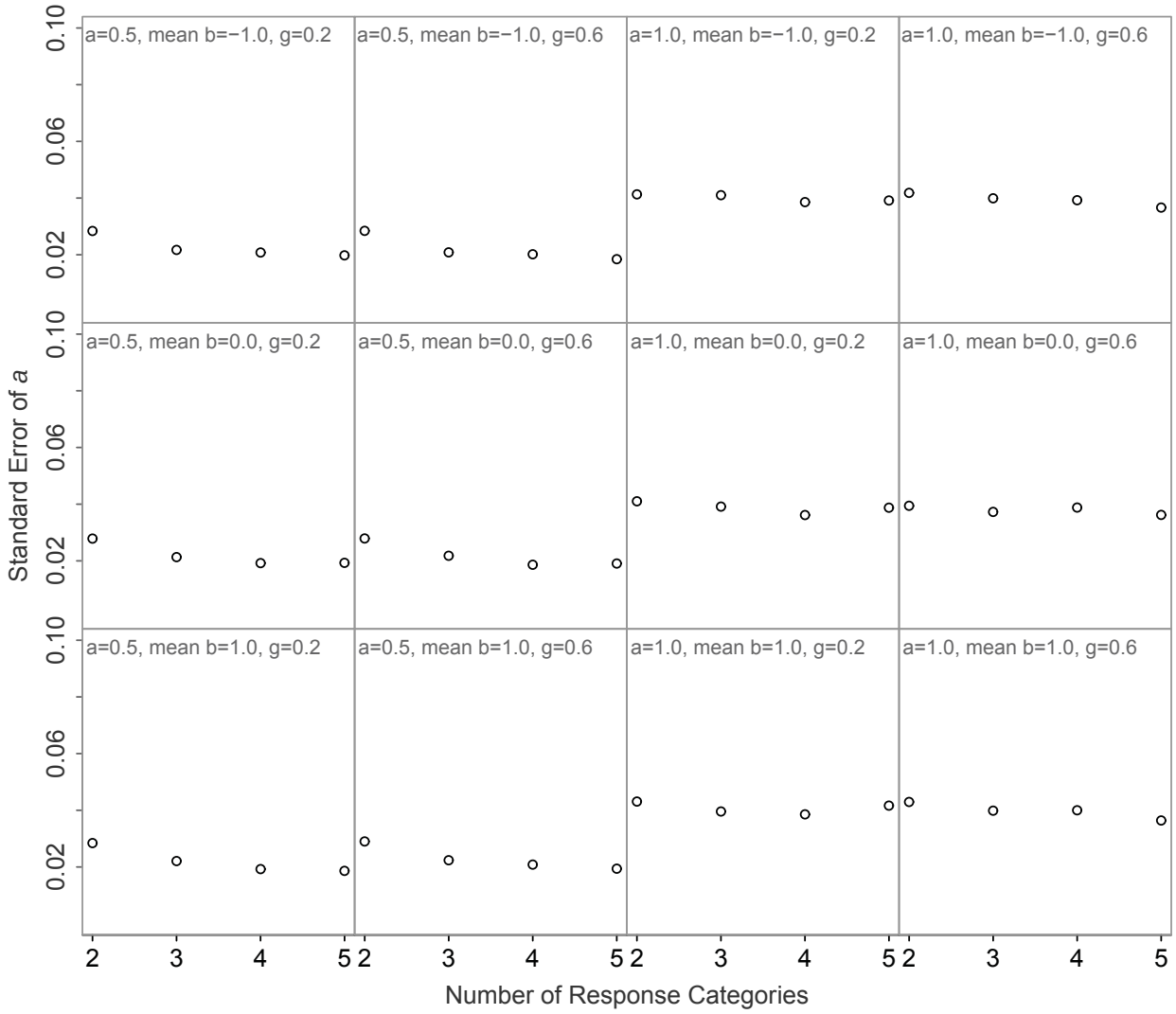

Figure 6.1. Estimation error of a with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). A uniform $[-3.0, 3.0]$ test-taker distribution was used for response simulation.

for the b parameters increased with the number of response categories, as did the covariances; yet, little difference in the standard errors for a was observed.

Figures 6.7 and 6.8 illustrate σ_{u_i} and σ_{v_i} as functions of K for each of the 48 items in the study. As hypothesized earlier, σ_{v_i} increased with K for all combinations of item parameter values due to the increase in estimation error for the b parameters as well as the size of the covariance matrix. For items with lower discrimination, the decrease in σ_{v_m} was even dramatic. As for the size of σ_{u_m}, it appeared to be quite robust against an increase in K indeed. In fact, it even tended to decrease slightly, especially for the items with the lower values for the discrimination

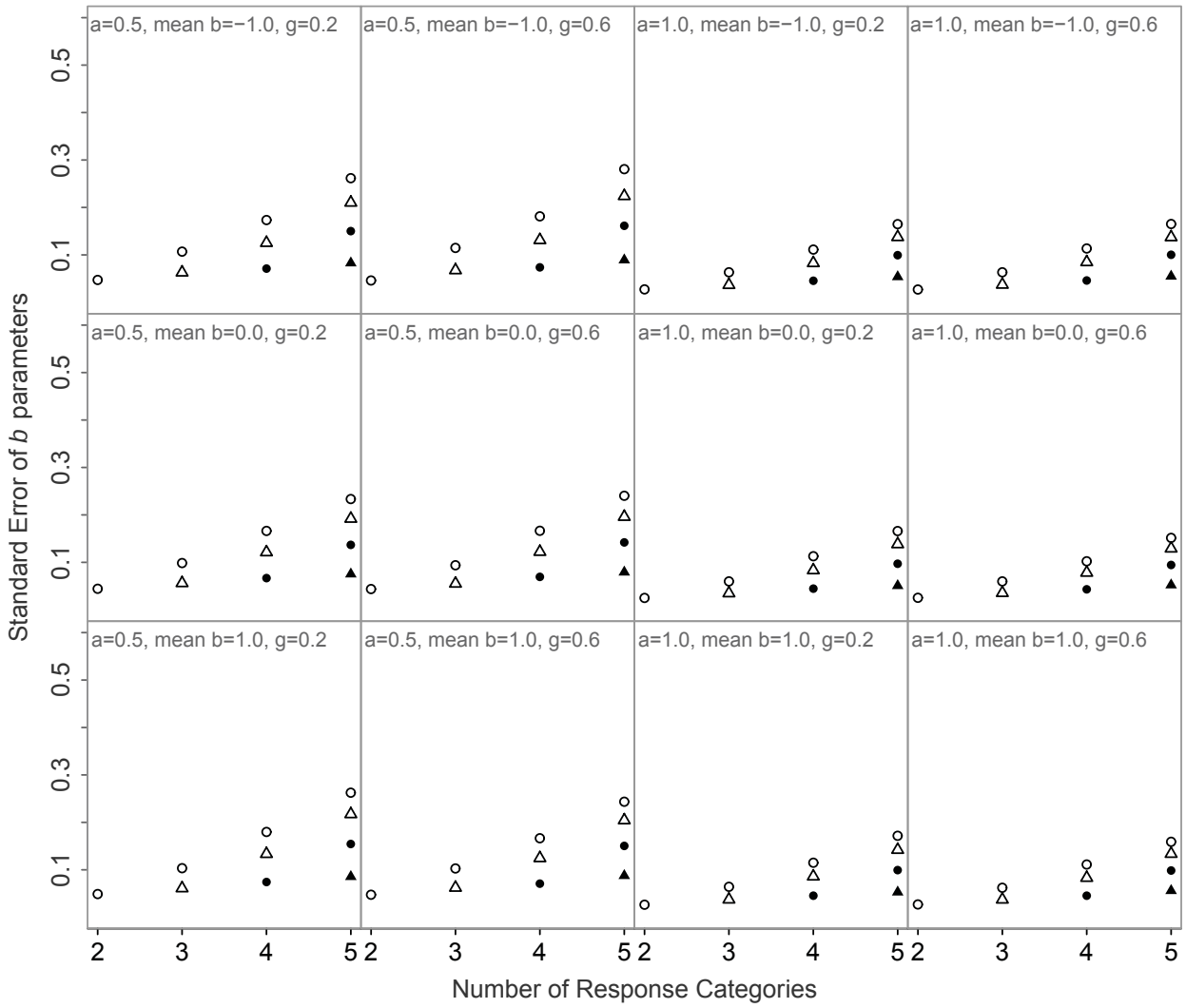

Figure 6.2. Estimation error of b parameters with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). b_1 is notated by the clear circle, b_2 is notated by the clear triangle, b_3 is notated by the solid circle, and b_4 is notated by the solid triangle. A uniform $[-3.0, 3.0]$ test-taker distribution was used for response simulation.

parameters. The preceding results were based on a wide uniform ability distribution to minimize its differential effect on the parameters for the middle and more extreme categories. However, in practice more peaked distributions of test-takers are typically met. Therefore, the same analyses were repeated with a $N(0, 1)$ ability distribution. Results showed larger standard errors for the b parameters and larger standard errors of linking, as would be expected from the lower number of test-takers in the tails of the distribution. Besides, depending on the average of the b parameters, there were some systematic differences between the covariance for the a and b parameters. In spite of these differences, however, the same general patterns were observed: σ_{u_i}

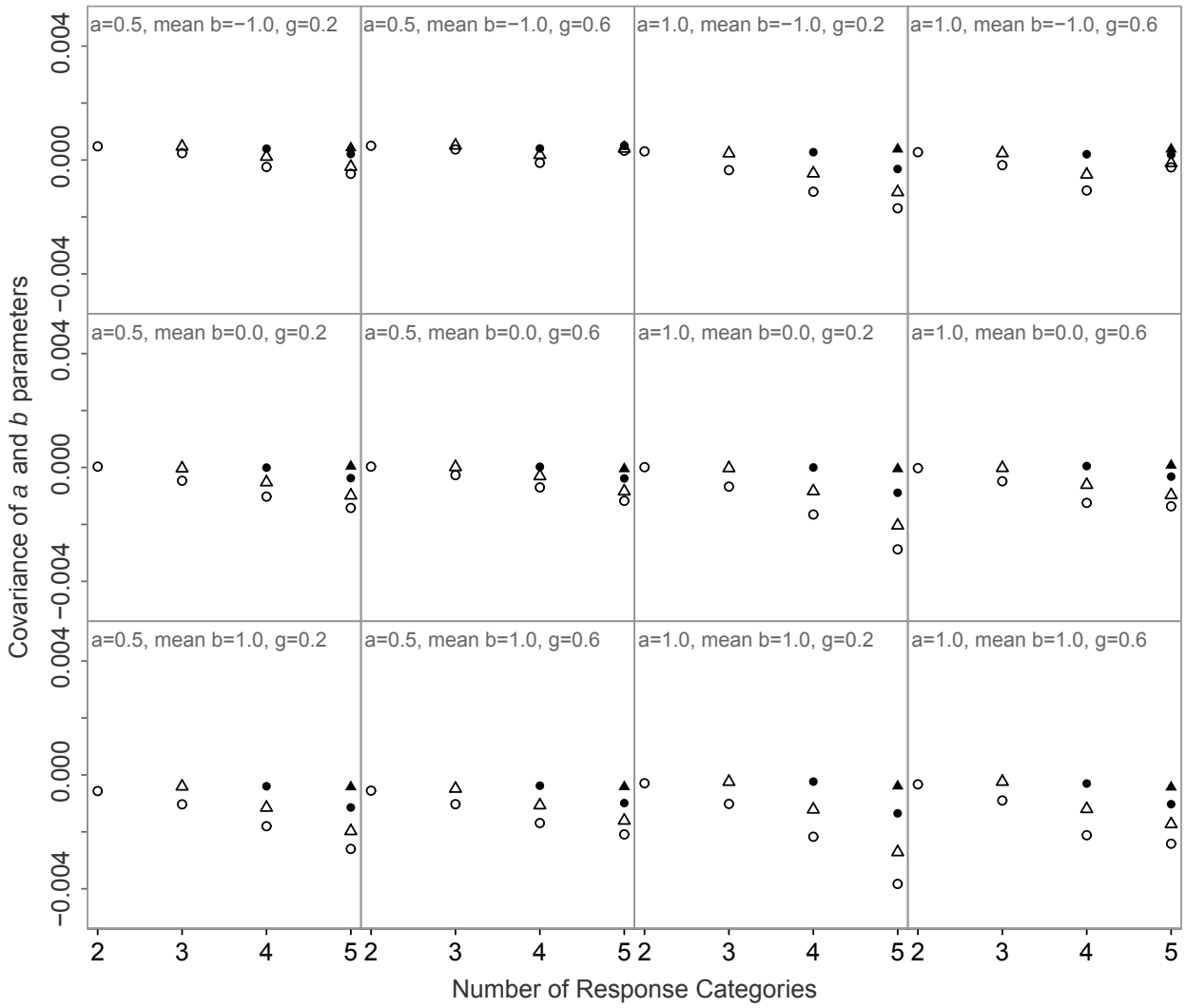

Figure 6.3. Covariance of a and b parameters with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). b_1 is notated by the clear circle, b_2 is notated by the clear triangle, b_3 is notated by the solid circle, and b_4 is notated by the solid triangle. A uniform $[-3.0, 3.0]$ test-taker distribution was used for response simulation.

decreased only slightly with K for the lower discrimination parameters items or not at all, and σ_{v_i} increased with K in all cases. Plots for this case analogous to Figures 6.1–6.8 are located in Appendix C.

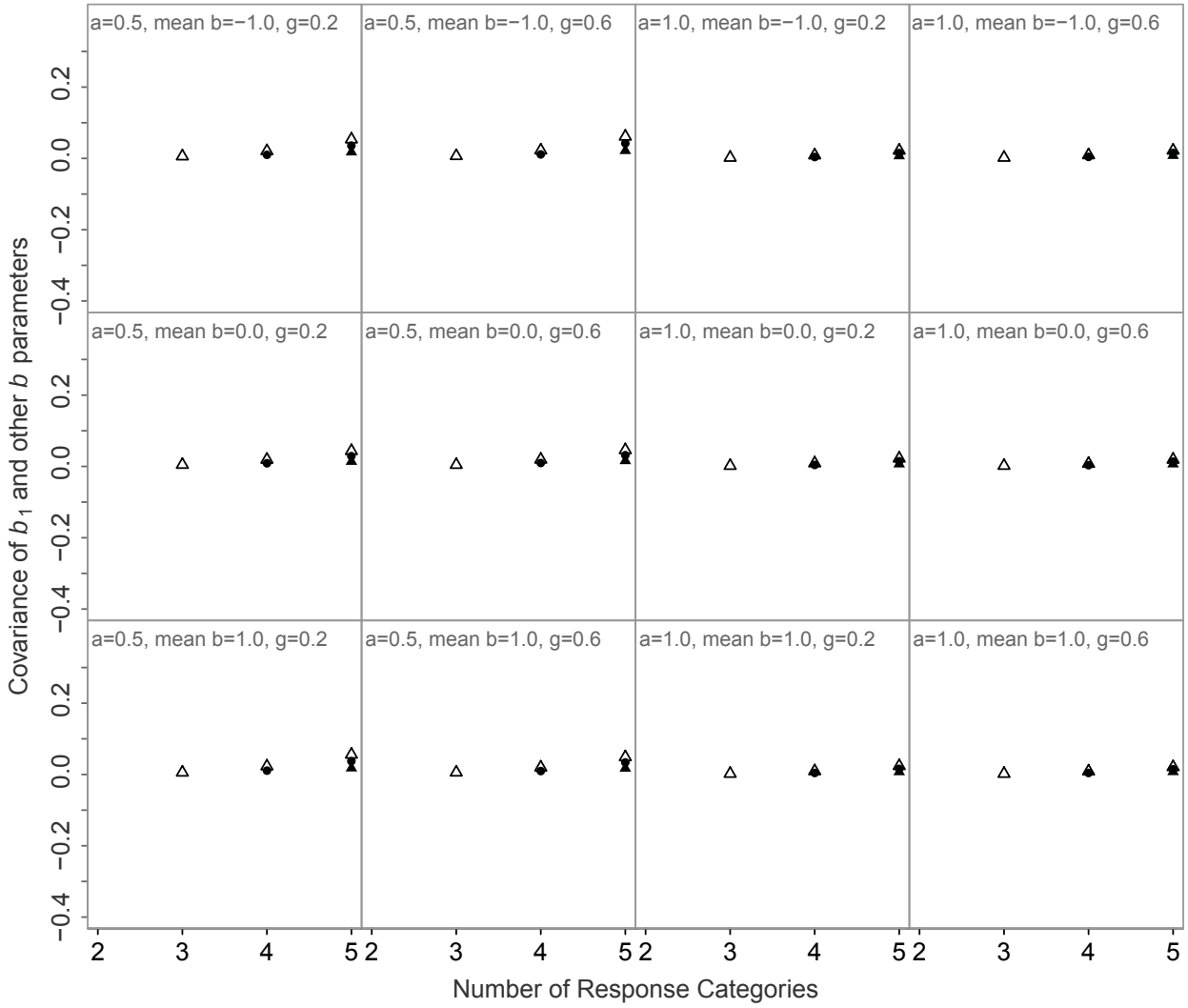

Figure 6.4. Covariance of b_1 and other b parameters with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). b_2 is notated by the clear triangle, b_3 is notated by the solid circle, and b_4 is notated by the solid triangle. A uniform $[-3.0, 3.0]$ test-taker distribution was used for response simulation.

6.5 Optimal Linking Design

When two test forms are to be linked, optimal linking design may be used to identify which of the available items should be included as common items to minimize linking error. Barrett and van der Linden (2015b) illustrated the use of mixed integer programming (MIP) to optimize linking designs for tests with dichotomous items. Their examples included cases such as designs with minimal σ_u, minimal σ_v, and a minimal composite of σ_u and σ_v, as well as cases with σ_u

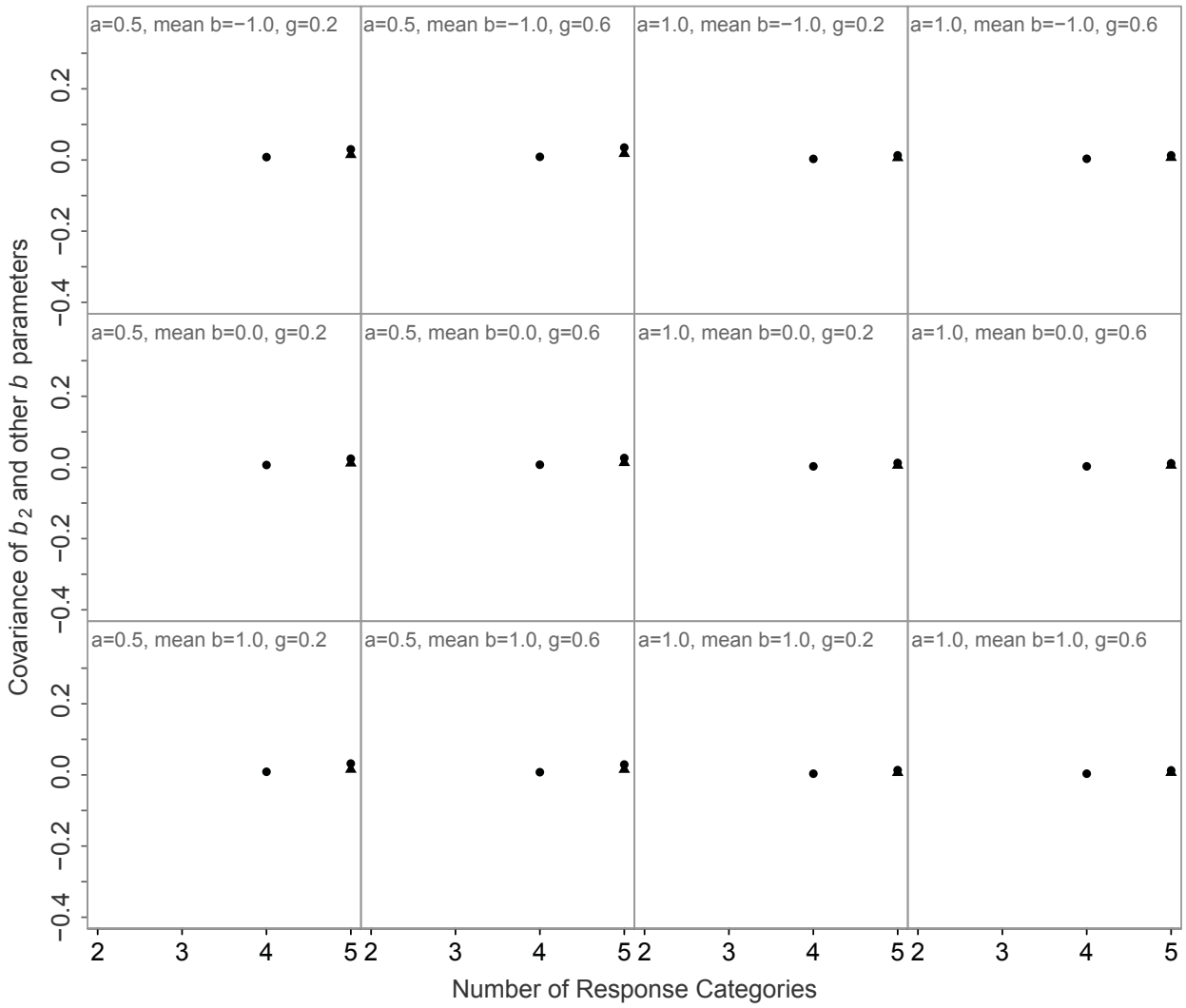

Figure 6.5. Covariance of b_2 and other b parameters with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). b_3 is notated by the solid circle, and b_4 is notated by the solid triangle. A uniform $[-3.0, 3.0]$ test-taker distribution was used for response simulation.

and σ_v constrained by upper bounds and optimization with respect to another objective, such as maximal information in the test scores at a cut score or with respect to another target.

Optimal linking design is possible because in precision-weighted linking, the contribution of each item to overall linking error is explicit. More specifically, the overall standard errors for the linking parameters can be shown to be linear in those for the common items. As a result

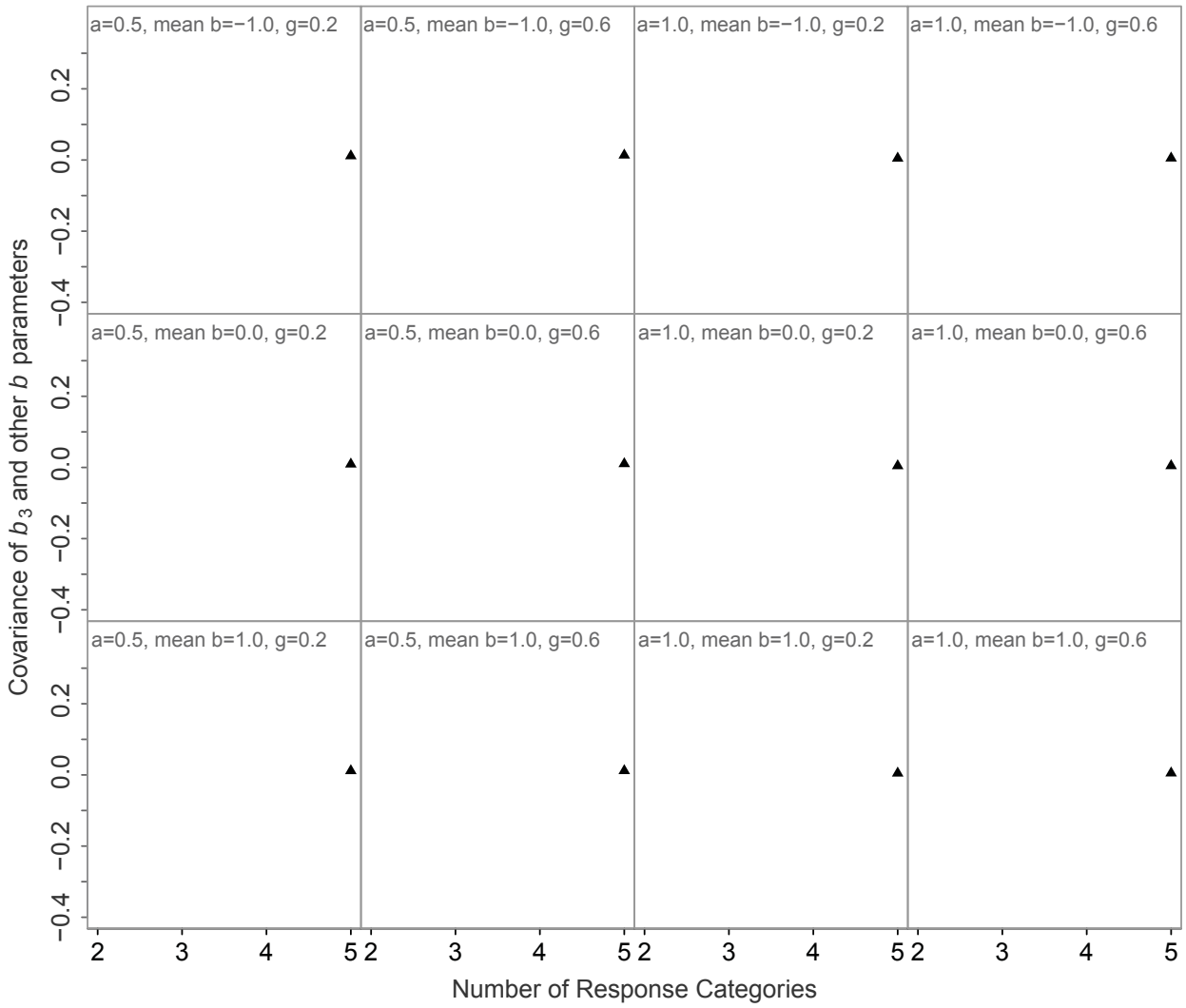

Figure 6.6. Covariance of b_3 and b_4 parameters with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). A uniform $[-3.0, 3.0]$ test-taker distribution was used for response simulation.

the standard error for u parameter in (6.15) is minimized when

$$\sum_{i=1}^{M} \frac{1}{\sigma_{u_i}^2}$$

is maximized, where M the number of common items, with the same relationship for the v parameter. For the mathematical formulations of a variety of optimal linking design models, refer to Barrett and van der Linden (2015b). Barrett and van der Linden, (2015c) extended the precision-weighted linking methodology for application to polytomous models.

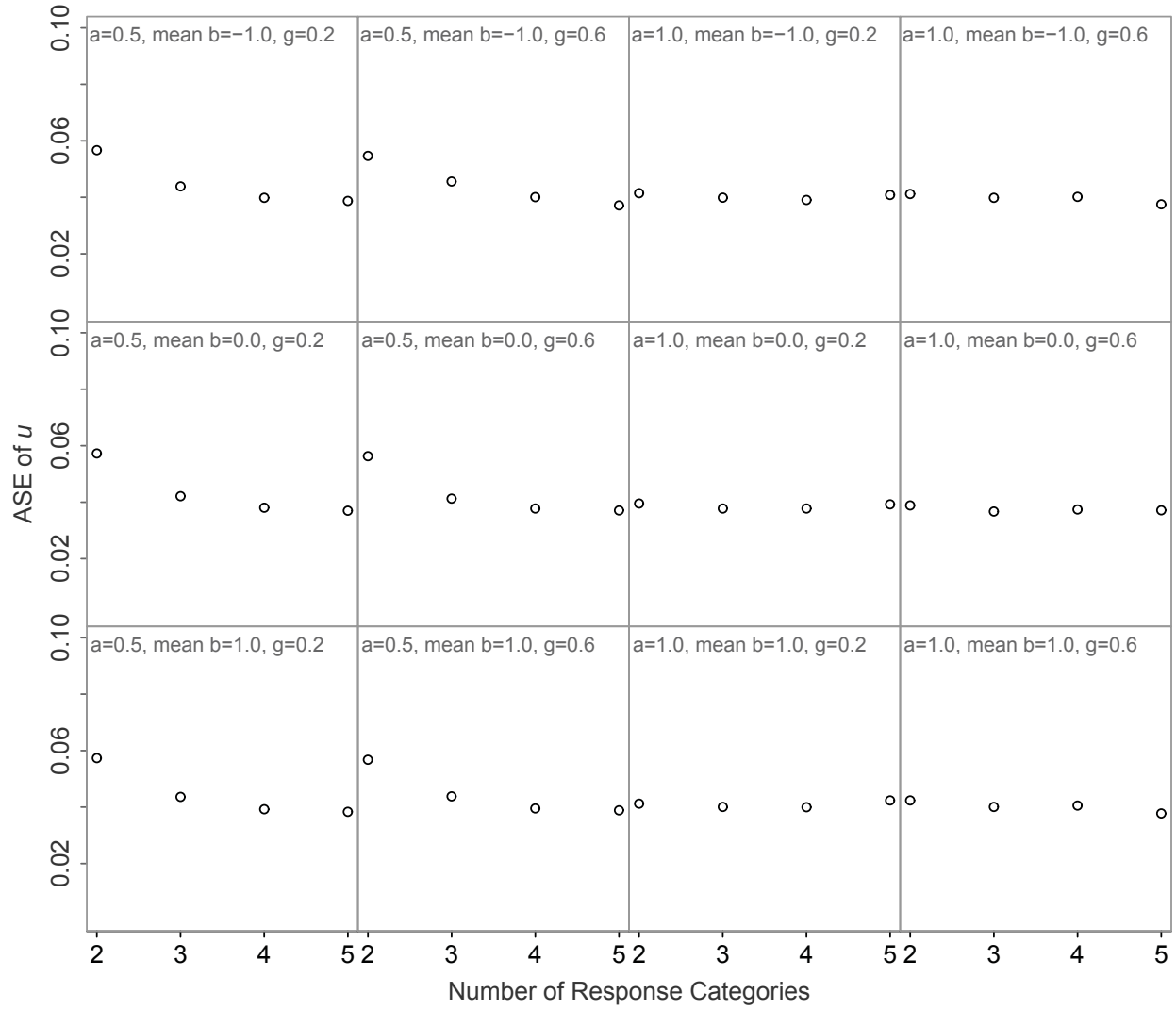

Figure 6.7. ASE of u with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). A uniform $[-3.0, 3.0]$ test-taker distribution was used for response simulation.

As discussed above, an increase in K tended to result in negligibly small decreases in σ_{u_i} but substantial increases in σ_{v_i}. It is therefore likely that minimizing a linking design with respect to σ_u will lead both to the choice of items with $K > 2$ and a higher σ_v. Conversely, when σ_v is minimized, items with $K = 2$ may be favored with as result a higher σ_u. An obvious choice, therefore, is to constrain both σ_u and σ_v with well-chosen upper bounds λ_u and λ_v, respectively. This choice allows us to optimize the design simultaneously with respect to another objective. Berger (1998) proposed optimization criteria (information targets) for efficiency of ability estimation for tests consisting of items with both dichotomous and polytomous response models. Of course, test assembly is typically subjected to numerous other constraints (e.g.,

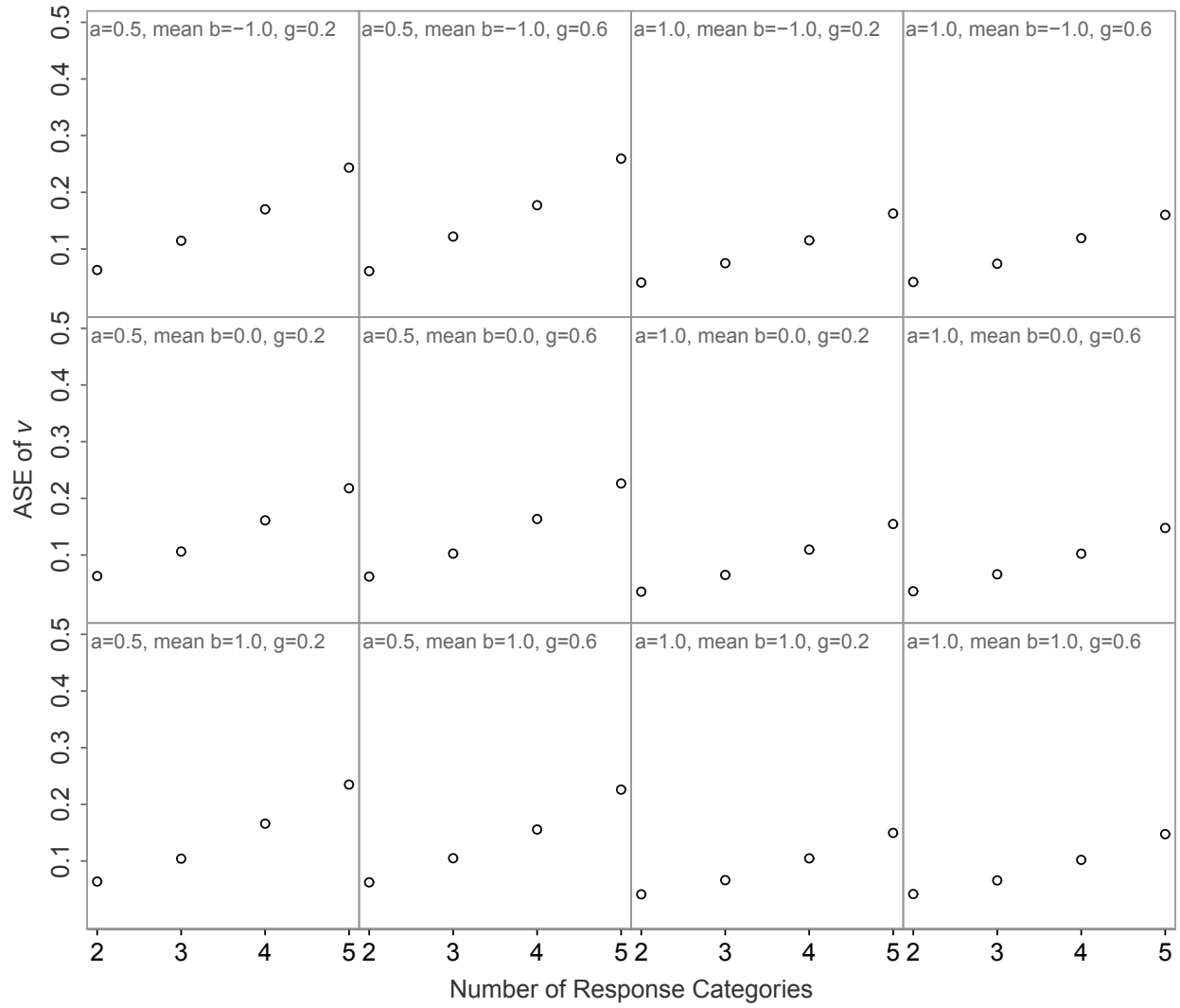

Figure 6.8. ASE of v with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). A uniform $[-3.0, 3.0]$ test-taker distribution was used for response simulation.

to realize blueprint requirements), which will also impact the selection of common items with different numbers of response categories.

Typically, an optimal linking design problem is met when one test form has already been administered and the subsequent form has to be assembled. Although estimates of the item parameters and the covariance matrices for the second form are not yet available, it is still possible to conduct an optimal linking design study by simulating the second administration with the anticipated ability distribution and the item parameter estimates from the first form. The responses can be used to estimate σ_{u_i} and σ_{v_i} for the candidate common items from the

first form. As odd distributions during second administrations due to the presence of repeaters or other reasons can be a concern (Keller & Hambleton, 2013; Qian & von Davier, 2015), this is a powerful approach that enables us to minimize their impact on linking error. For more details on such simulations, see Barrett and van der Linden (2015b).

The empirical example of optimal linking design presented here is from a real-world high-school mathematics program that has both dichotomous and polytomous items in its test forms. The empirical data were for 59,518 test-takers who had taken a 60-item exam with 45 items that were scored dichotomously and 15 that were scored polytomously. Of the polytomous items, $K = 3$ for six, $K = 4$ for six, and $K = 5$ for three items. The goal was to minimize linking error through optimal selection of 15 items from the first test administration that had to serve as common items in the second.

The response data from the first administration were used to calibrate its items under the GPCM using *MIRT 1.0* software (Glas, 2010) (including the items with $K = 2$) with the option of MML estimation and $\theta \sim N(0, 1)$. As the testing program was stable without any anticipated changes in the test-taker population, the second administration was simply simulated with 59,000 test-takers sampled from the same normal distribution. The estimates of both sets of item parameter and covariance matrices were used to calculate the anticipated standard error of the linking parameters for each of the candidate common items from the first test administration.

The test-assembly model was run several times to illustrate the impact of the choices of objective function and constraints. Table 6.2 lists the full set constraints actually imposed by the content experts and psychometricians working on the program. Item metadata was available for each of its item attributes. All runs were made using *lp_solve* 5.5.2.0 (Berkelaar, Eikland & Notebaert, 2004) with *lpSolveAPI* for *R* (Konis, 2013).

First, it is informative to examine the minimum linking error possible given the common items available for selection, without imposing any of the additional constraints in Table 6.2. The results are given in Table 6.3. Each of its rows represents a separate assembly run with the objective function in the far left column. The upper bounds used to constrain the standard error for the linking parameter not in the objective function are provided in the second and third columns. When σ_u was minimized without any constraints, all three of the items with $K = 5$, five of six available items with $K = 4$, and three of six available items with $K = 3$, but only four of the items with $K = 2$ were selected. However, when an upper bound was imposed on σ_v, the proportion of items with lower K categories increased substantially. In fact, each time the upper bound was made more stringent, more items with $K = 2$ were in the

Table 6.2. *High-school mathematics exam, linking item constraints*

V_c^{item}	Lower Bound λ_c	Upper Bound λ_c
All items	15	15
Items with a use status other than 'Yes' (e.g., do not use)	0	0
Items in positions 1 through 20 in the first administration	5	10
Items in positions 21 through 40 in the first administration	5	7
Items in positions 41 through 60 in the first administration	5	7
Items with answer key = 1	3	6
Items with answer key = 2	3	12
Items with answer key = 3	3	6
Items with answer key = 4	3	6
Items flagged due to poor item fit	0	0
Items flagged due to high omit rates	0	0
Items with p-values under 0.25	0	2
Items with p-values over 0.9	0	0
Items with point biserial correlation under 0.2	0	3
Items with a point biserial over 0.05 on a distractor	0	0
Items that cannot be Brailled	0	2
Items from sub-content area Number Sense and Computation	5	8
Items from sub-content area Algebra	4	5
Items from sub-content area Statistics and Probability	1	5
Items from sub-content area Geometry	4	5

optimal selection. In contrast, when σ_v was minimized without any constraints, all 15 items selected had $K = 2$ categories. Each time a more stringent upper bound on σ_u was imposed, the proportion of items with $K > 2$ categories increased. Thus, exactly as hypothesized earlier in this paper, items with more response categories were favored when σ_u was optimized, while items with fewer categories were favored when σ_v was optimized.

Once the content and psychometric constraints were imposed, as seen in Table 6.4, the minimum possible linking error increased. Also, in order to avoid infeasibility, the choice of the upper bounds on the standard error not being optimized became more critical. Again, as σ_u was bounded at lower values, the number of items with $K = 2$ increased. Because of the presence of all other constraints, only one reasonable bound on σ_v was feasible.

Optimal linking design makes explicit the trade-offs that occur among all constraints and the objective. Decisions about whether the linking error is acceptable or whether a constraint on it must be tightened to result in less error can be made using the data produced by this method.

Table 6.3. *Minimizing linking error using optimal linking design without constraints in Table 6.2*

			No additional constraints						
						Number of items selected			
Maximize	λ_u	λ_v	Status	σ_u	σ_v	$K=2$	$K=3$	$K=4$	$K=5$
$1/\sigma_u^2$	–	–	F	0.0034	0.0044	4	3	5	3
$1/\sigma_u^2$	–	0.0040	F	0.0034	0.0040	6	3	3	3
$1/\sigma_u^2$	–	0.0038	F	0.0034	0.0038	7	2	3	3
$1/\sigma_u^2$	–	0.0036	F	0.0035	0.0036	8	3	2	2
$1/\sigma_u^2$	–	0.0034	F	0.0036	0.0034	10	2	1	2
$1/\sigma_u^2$	–	0.0032	F	0.0037	0.0032	12	2	0	1
$1/\sigma_u^2$	–	0.0030	I	–	–	–	–	–	–
$1/\sigma_v^2$	–	–	F	0.0039	0.0030	15	0	0	0
$1/\sigma_v^2$	0.0040	–	F	0.0039	0.0030	15	0	0	0
$1/\sigma_v^2$	0.0038	–	F	0.0038	0.0031	14	1	0	0
$1/\sigma_v^2$	0.0036	–	F	0.0036	0.0033	10	3	2	0
$1/\sigma_v^2$	0.0034	–	F	0.0034	0.0041	5	3	4	3
$1/\sigma_v^2$	0.0032	–	I	–	–	–	–	–	–

Table 6.4. *Minimizing linking error using optimal linking design with constraints in Table 6.2*

			Including all constraints						
						Number of items selected			
Maximize	λ_u	λ_v	Status	σ_u	σ_v	$K=2$	$K=3$	$K=4$	$K=5$
$1/\sigma_u^2$	–	–	F	0.0039	0.0045	10	1	2	2
$1/\sigma_u^2$	–	0.0040	I	–	–	–	–	–	–
$1/\sigma_u^2$	–	0.0038	I	–	–	–	–	–	–
$1/\sigma_u^2$	–	0.0036	I	–	–	–	–	–	–
$1/\sigma_u^2$	–	0.0034	I	–	–	–	–	–	–
$1/\sigma_u^2$	–	0.0032	I	–	–	–	–	–	–
$1/\sigma_u^2$	–	0.0030	I	–	–	–	–	–	–
$1/\sigma_v^2$	–	–	F	0.0043	0.0042	12	2	1	0
$1/\sigma_v^2$	0.0040	–	F	0.0039	0.0045	10	1	2	2
$1/\sigma_v^2$	0.0038	–	I	–	–	–	–	–	–
$1/\sigma_v^2$	0.0036	–	I	–	–	–	–	–	–
$1/\sigma_v^2$	0.0034	–	I	–	–	–	–	–	–
$1/\sigma_v^2$	0.0032	–	I	–	–	–	–	–	–

6.6 Concluding Remarks

This paper examined the relationship between linking error and item characteristics, specifically the number of response categories of an item, when using a precision-weighted linking approach. Both our analytic and empirical results indicated that, when the GPCM is used, estimation error in the u parameter is mildly inversely related to the number of categories when item discrimination is low. In contrast, error in the v parameter appears to increase strongly with the number of categories.

Although all mainstream polytomous models share the same true linking functions, as estimation error in the functions depends on the covariance matrix of the estimators of the item parameters, different results may be obtained for other polytomous models than the GMPM used in this paper. For example, if the standard error of estimation for the a parameter does not appear to decrease with the number of response categories for some of these models, σ_u is more likely to become completely independent of it. Additional research is required to understand how widely our current findings hold.

When selecting common items to serve as a link between different test forms, it is useful to make explicit the various trade-offs among the test assembly constraints and the anticipated linking error. The purpose of the testing program, the nature of the anticipated score distributions, as well as possible other practical factors should inform the choices. Future studies should also examine how linking error propagates into the test scores of individual test-takers. These studies are expected to involve an important change in the objective functions of the optimal design models used in this paper but not in the rest of their structure.

Acknowledgment

The authors are indebted to Cees A. W. Glas for his support during their use of his *MIRT Scaling Program* (version 1.01) software.

References

Barrett, M. D., & van der Linden, W. J. (2015a). *Estimating linking functions for response model parameters.* Manuscript submitted for publication.

Barrett, M. D., & van der Linden, W. J. (2015b). *Optimal linking design for response model parameters.* Manuscript submitted for publication.

Barrett, M. D., & van der Linden, W. J. (2015c). *Linking polytomous response model parameters.* Manuscript to be submitted for publication.

Berger, M. P. F. (1998). Optimal design of tests with dichotomous and polytomous items. *Applied Psychological Measurement, 22*, 248–258.

Berkelaar, M., Eikland, K., & Notebaert, P. (2004). *lp_solve, version 5.5.2.0.* [Computer software]. Retrieved from `http://lpsolve.sourceforge.net/5.5/`.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bock, R. D. (1997). The nominal categories model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer.

Ercikan, K., Schwarz, R., Julian, M. W., Burket, G. R., Weber, M. W., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response test item type. *Journal of Educational Measurement, 35*, 137–154.

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education,* 31–57.

Glas, C. A. W. (2010). MIRT (version 1.01) [Computer software and manual]. Retrieved October 27, 2014, from `http://www.utwente.nl/gw/omd/Medewerkers/medewerkers/glas/`.

Haebara, T. (1980). Equating logistic ability scales by weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Hagge, S. L., & Kolen, M. J. (2011). Equating mixed-format tests with format representative and non-representative common items. *Mixed-format tests: Psychometric properties with a primary focus on equating, 1*, 95-135. Retrieved January 30, 2015 from `http://www.education.uiowa.edu/docs/default-source/casma-monographs/casmamonograph2_1.pdf?sfvrsn=2`.

Jodoin, M. G., Keller, L. A., & Swamination, H. (2003). A comparison of linear, fixed common parameter, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education, 71*, 229–250.

Kamata, A., & Tate, R. (2005). The performance of a method for the long-term equating of mixed-format assessment. *Journal of Educational Measurement, 42*, 193–213.

Keller, L. A., & Hambleton, R. K. (2013). The long-term sustainability of IRT scaling methods in mixed-format tests. *Journal of Educational Measurement, 50*, 390–407.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed-format texts. *Journal of Educational Measurement, 43*, 53–76.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*, 357–381.

Kim, S., Walker, M.E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large scale assessments. *Journal of Educational Measurement, 47*, 36–53.

Kim, S., Walker, M.E., & McHale, F. (2010). Investigating the effectiveness of equating designs for constructed-response tests in large-scale assessments. *Journal of Educational Measurement, 47*, 186–201.

Konis, K. (2013). *lpSolveAPI, version 5.5.2.0-9.* [Computer software]. Retrieved from `http://:CRAN.R-project.org/package=lpSolveAPI`.

Lee, W., He, Yi., Hagge, S., Wang, W., & Kolen, M. J. (2012). Equating mixed-format tests using dichotomous common items. *Mixed-format tests: Psychometric properties with a primary focus on equating, 2*, 13–44. Retrieved February 8, 2015 from `http://www.education.uiowa.edu/docs/default-source/casma-monographs/volume-2.pdf?sfvrsn=2`.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractible testing problems. *Journal of Educational Measurement, 14*, 139–160.

Masters, G. M. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–173.

Masters, G. M., & Wright, B.D. (1997). The partial credit model.In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

Muraki, E., & Chang, H. (1994). *Horizontal and vertical test equating methods based on the generalized partial credit model*. (ETS Internal Report). Princeton NJ: Educational Testing Service.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce), 51*, 1–23.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25,* 53–67.

Ogasawara, H. (2011). Applications of asymptotic expansion in item response theory linking. In A. A. von Davier (Ed.), *Statistical models for test equating scaling, and linking*. New York: Springer.

Powers, S. J., Hagge, S. L., Wang, W., He, Y., Liu, C., & Kolen, M. J. (2011). Effects of group differences on mixed-format equating. *Mixed-format tests: Psychometric properties with a primary focus on equating, 1*, 51–73. Retrieved January 30, 2015 from `http://www.education.uiowa.edu/docs/default-source/casma-monographs/casmamonograph2_1.pdf?sfvrsn=2`.

Qian, J. & von Davier, A. A. (2015). Optimal sampling design for IRT linking with bimodal data. In R. E. Millsap et al. (eds.), *Quantitative Psychology Research, Springer Proceedings in Mathematics & Statistics, 89*, 165–179.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Tate, R. L. (2000). Performance of a proposed method for the linking of mixed-format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*, 329–346.

Tate, R. L. (2003). Equating for long-term scale maintenance of mixed format tests containing multiple choice and constructed response items. *Educational and psychological measurement, 63*, 893–914.

van der Linden, W. J., & Barrett, M. D. (in press). Linking item response model parameters. *Psychometrika*. doi: 20.1007/s11336-015-9469-6.

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer..

Wang, W., Lee, W., & Kolen, M. J. (2012). An investigation of IRT item fit statistics for large-scale mixed format tests. *Mixed-format tests: Psychometric properties with a primary focus on equating, 2*, 143–183. Retrieved February 8, 2015 from `http://www.education.uiowa.edu/docs/default-source/casma-monographs/volume-2.pdf?sfvrsn=2`.

Wan, L. & Henly, G.A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education, 25*, 58–78.

Yao, L. & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement, 46*, 177–197.

Chapter 7

Conclusions

The research reported in this thesis was motivated by a need to clarify and quantify the sources of error in linking function estimators such that error may be reduced and adequately reported in operational testing programs. Accordingly, the research goals were (1) to characterize the formal nature of linking functions for common item response models, (2) to identify the sources of random error in estimators of those linking functions and to quantify that error, and (3) to explore methods to reduce propagation of error. The re-conceptualization of the response model linking problem as addressing the unidentified nature of the statistical model in use, rather than borrowing procedures from the observed-score equating framework, was the foundation of the research.

7.1 Summary of Key Findings

Chapter 2 presents main theorems to characterize the formal nature of linking functions for monotone, continuous response models necessary to adjust for the effects of identifiability restrictions on the model parameters. Specific shapes of the linking functions for the traditional $a_i(\theta_p - b_i)$ and the slope-intercept $\alpha_i\vartheta_p + \beta_i$ parameterizations of the common dichotomous logistic models were derived, and minimal linking elements (common items and/or test-takers) required to identify the function were presented. For the traditional parameterization, the general shape of the linking function matched that currently in use by mean/mean and mean/sigma methods; however, with a new definition for the u and hence the v linking parameter. The linking function for the c_i parameter was shown to be an identity function; therefore, no linking is required for the c_i parameter. Interestingly, if the traditional parameterization is used, response

model parameters may be linked through a single common item, a pair of common items, or a pair of common test-takers, while the slope-intercept parameterization linking function requires both common items *and* common test-takers for the linking functions to be identified. This type of linking design is often practically infeasible.

In *Chapter 3*, asymptotic standard errors (ASE) of the estimated linking function parameters were derived for a minimal linking element of the 3PL model using a multivariate delta method. The analysis clarified the sources of random error present in linking function estimates; only estimation error in the item or ability parameters propagates into the linking function estimates, not any sampling error. Although the 3PL was used in the analysis, results hold equally well for the Rasch and 2PL models when fixing the appropriate parameters. The analysis showed that estimation error in the c_i parameter does not contribute to error in the linking function; therefore, the response-function methods currently in use confound estimation error in the c_i parameter with linking function estimation error. It may be best just to compute a precision-weighted average of these parameters from the two test administrations as an estimate of the true c_i parameter.

Based on the derivations of linking parameter and variance estimates for single linking elements (a common item, a pair of items, or a pair of test-takers), a precision-weighed estimator for the overall linking design was proposed. Empirical examples illustrated monotone decrease in linking error as a function of the number of linking elements. The new approach also resulted in less error than the current mean/mean and mean/sigma methods, especially when some of the item parameters have large estimation error relative to the others. In this case, the current methods appear to violate the obvious requirement of monotone decrease of their standard errors with the number of linking elements in the linking design.

Not only did the precision-weighted estimator exhibit promising characteristics in the empirical examples, it made possible the use of optimal linking design techniques as shown in *Chapter 4*. In this approach, common items are selected for a linking design to minimize overall linking error. Several empirical examples illustrated that it is not only possible but quite practical to minimize the error in the u parameter, minimize error in the v parameter, or minimize a composite of error in the u and v parameters subject to all content, psychometric, and practical constraints usually imposed on the assembly of test forms. Examples also illustrated the use of optimal linking design to select items for an entire test form, maximized to an information target, while bounding error in the u and v parameters. Methods were presented to allow for the selection of common items on two tests, before the second test is administered, such that

linking error will be minimized. These methods were shown to be flexible in handling cases where changes in ability distributions may be expected for the second set of test-takers.

In *Chapter 5*, the methods of *Chapter 3* were extended for polytomous items. It was shown that linking functions for main-stream polytomous response models may be inferred from those for their underlying step functions, which are typically the response functions for the 2PL model. As each polytomous item has multiple step functions, each implying the same linking parameters but with separate (dependent) estimates of them, the proposed estimation method combines these estimates into a single estimator. Closed-form expressions of the ASE for these item-level estimators were derived. The ASEs were then used to combine the item-level estimators using a precision-weighted approach. Empirical examples illustrated the expected reduction in linking error as a function of the number of linking elements. In contrast with the polytomous linking methods currently used in practice, the method made transparent the contribution of each linking element to the overall linking parameter estimate, and favored the elements with the least amount of error in their parameter estimates.

With linking function estimation precision understood at the level of the minimal linking element for polytomous models, it was then possible to examine the relationship between the number of response categories of the common items in a linking design and their contribution to the estimation error in the linking function, the topic of *Chapter 6*. For the generalized partial credit model, analysis was presented that suggested that the asymptotic standard error for the slope parameter in the linking function will hardly change with the number of response categories of an item, while the error for the intercept parameter will increase with the number of categories. Empirical examples were provided that concur with the analytic results. The chapter concludes with an empirical linking design study selecting an optimal set of common items from a pool with varying numbers of score categories, with results that also did confirm the results.

Some limitations to the research should be recognized. The methods proposed quantify and minimize the random error in the linking function. In practice, systematic error may also be present that results in bias. Optimal linking design, as it can minimize error *subject to content and other test requirements*, will help to mitigate the chance of systematic error. That said, there may be factors related to the administration, such as over-exposure of a common item, that can produce biased results. In addition, the methods assume that the models fit the data; lack of model fit or substantial differential item functioning may impact results. Finally, linking through an assumption of randomly equivalent groups of test-takers is not addressed.

7.2 Future Research Directions

This thesis contributes important steps toward laying a theoretical foundation for linking response model parameters by formalizing the nature of the response model linking function for commonly used response models, providing a framework for deriving the linking functions of other response models, presenting a new estimator of the linking function for common dichotomous and polytomous response models that reduces the propagation of model parameter estimation error, and providing and illustrating practical and efficient approaches to selecting a linking design and specific common linking elements to minimize linking error.

From this foundation, many important future research studies become possible and relevant. The findings represented in the thesis focus on propagation of parameter estimation error into linking functions. To understand the ultimate impact of linking error on test-takers and those using test results, it is important to derive and investigate the propagation of linking error into ability estimates of individual test-takers as well as estimators of functionals defined on ability distributions. Minimally, acceptable bounds for linking parameter error should be established to inform related decisions about how many common items or test-takers a linking design requires and which common items are best suited to serve as linking elements. Methods to adequately report the error to test users should be addressed as well.

In *Chapter 2*, the linking function for the c parameter revealed itself to be an identity function. Research to evaluate the impact of confounding estimation error of the c parameter with linking error in response-function linking methods should be extended to inform any future use of these methods.

In cases where multiple test forms are linked sequentially across many administrations, often referred to as chain linking, propagation of linking error may result in unreasonable item parameter values or large error in test-taker ability estimates. Studies should therefore be conducted to inform the practices of chain linking and assessment design; for example, there may be methods for optimal linking design in which linking error across multiple test administrations is minimized.

With more frequent use of Bayesian methods to estimate item parameter distributions, research should be conducted to inform linking methods within the Bayesian framework. Results in the second chapter found that a parameterization commonly used in Bayesian response model parameter estimation requires an often untenable linking design; there may be additional considerations future research would reveal.

Finally, the observed score equating framework is typically also applied to item response models other than those in scope of this dissertation. However, as this thesis demonstrated, a correct theoretical treatment of response model linking as required due to lack of model identifiability is necessary to derive the linking function, identify sources of estimation error, quantify that error, and minimize it for the response model in use. Thus, research on linking functions for other models, such as used for multidimensional IRT, would similarly benefit from a formal consideration of identifiability requirements, derivation of linking functions, analysis of parameter estimation error propagation, optimal linking design, and an ability to report the impact of linking error on test-taker scores.

Appendix A

Example LP File Used for Optimal Linking Design

```
MAX:2.2553340503x1+196.29610441x2+1126.7951695x3+156.60599533x4
+2674.7422362x5+42.115995438x6+2236.8860936x7+6.8271724793x8
+16.190086531x9+67.254424184x10+76.916178375x11+473.82163354x12
+894.11342049x13+277.40761148x14+1.3553584724x15+145.98277011x16
+871.07936961x17+521.01473835x18+851.71739644x19+168.37474761x20
+1263.2322643x21+1710.6502071x22+36.319081465x23+31.676844594x24
+257.08688697x25+1144.2947016x26+1397.2645478x27+260.35778293x28
+38.705539079x29+181.78521416x30+168.58462488x31+418.86928642x32
+2851.9012771x33+277.88238767x34+1597.2884415x35+17.883850958x36
+214.10974608x37+1158.9045356x38+1202.312276x39+436.25134686x40
+358.0289406x41+85.987371106x42+1019.0063648x43+8.1622690971x44;

C2:0<=x2+x7+x10+x17+x18+x26+x29+x34+x35+x36+x37+x44<=0;
C3:5<=x10+x11+x15+x17+x19+x23+x29+x40+x41<=10;
C4:5<=x2+x7+x12+x13+x14+x18+x20+x22+x24+x26+x30+x32+x34+x35+x37
+x42+x43<=7;
C5:3<=x1+x4+x5+x6+x8+x9+x16+x21+x25+x27+x28+x31+x33+x36+x38+x44<=7;
C6:3<=x5+x17+x20+x21+x22+x23+x34+x38+x39+x41+x44<=6;
C7:3<=x4+x14+x18+x25+x26+x27+x28+x30+x32+x37<=12;
C8:3<=x8+x10+x11+x15+x16+x24+x31+x42<=6;
```

```
C9:3<=x1+x2+x3+x6+x7+x9+x12+x13+x19+x29+x33+x35+x36+x40+x43<=6;
C10:0<=x36<=0;
C12:0<=x34+x36<=2;
C14:0<=x24+x28+x29+x33<=3;
C15:0<=x21<=0;
C18:0<=x29+x36<=0;
C19:0<=x2+x8+x13+x14+x15+x16+x19+x20<=2;
C20:5<=x2+x16+x18+x19+x21+x24+x36+x38+x40+x42<=8;
C21:4<=x1+x6+x7+x11+x13+x14+x15+x31+x33+x35+x39+x43<=5;
C22:1<=x4+x5+x8+x12+x20+x22+x25+x26+x27+x30+x32+x34<=5;
C23:4<=x3+x9+x10+x17+x23+x28+x29+x37+x41+x44<=5;
C24:15<=x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17
+x18+x19+x20+x21+x22+x23+x24+x25+x26+x27+x28+x29+x30+x31+x32+x33+x34
+x35+x36+x37+x38+x39+x40+x41+x42+x43+x44<=15;

bin x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13,x14,x15,x16,x17,x18,
x19,x20,x21,x22,x23,x24,x25,x26,x27,x28,x29,x30,x31,x32,x33,x34,x35,
x36,x37,x38,x39,x40,x41,x42,x43,x44
```

Appendix B

Example LPsolveAPI R Code Used for Optimal Linking Design

```
#create model and settings with decision variables for each item in M and I

lprec=make.lp(0,M+I)
lp.control(lprec,sense="max",epsint=0.1,mip.gap=c(0.1,0.05))

#decision variable definitions
#EQ 4.36,4.37

set.type(lprec,columns=c(1:M,1:I),type="binary")
set.semicont(lprec,columns=c(1:M,1:I),sc=FALSE)
set.bounds(lprec,lower=rep(0,M+I),upper=rep(1,M+I))

#add objective function
#EQ 4.38,4.39

#run first without any constraints to get minimally obtainable bounds
#on standard error of linking for common items
#subject to length constraint. This informs decisions about bounds
#when used as a constraint.  Remove # to get the minimally obtainable
#bound for each of the linking parameter estimates
```

```
#set.objfn(lprec,M.dat$tobj.u,indices=1:M)
#set.objfn(lprec,M.dat$tobj.v,indices=1:M)

#run for entire test assembly (maximize information at theta=0.5, as calculated
#in data field info2)

set.objfn(lprec,append(M.dat$info2,I.dat$info2),indices=1:(M+I))

#add test length constraints
#EQ 4.32,4.33

add.constraint(lprec,rep(1,M),"=",g,indices=1:M)
add.constraint(lprec,rep(1,I),"=",h-g,indices=(M+1):(M+I))

#add content constraints
#EQ 4.34,4.35

add.constraint(lprec,rep(1,length(append(M.ItemStatus,I.ItemStatus))),
"=",0,indices=c(M.ItemStatus,M+I.ItemStatus))
add.constraint(lprec,rep(1,(length(append(M.Cont1,I.Cont1)))),
"<=",24,indices=c(M.Cont1,M+I.Cont1),14)
add.constraint(lprec,rep(1,(length(append(M.Cont2,I.Cont2)))),
"<=",15,indices=c(M.Cont2,M+I.Cont2),12)
add.constraint(lprec,rep(1,(length(append(M.Cont3,I.Cont3)))),
"<=",15,indices=c(M.Cont3,M+I.Cont3),3)
add.constraint(lprec,rep(1,(length(append(M.Cont4,I.Cont4)))),
"<=",15,indices=c(M.Cont4,M+I.Cont4),12)
add.constraint(lprec,rep(1,length(M.Cont1)),
"<=",9,indices=(M.Cont1),4)
add.constraint(lprec,rep(1,length(M.Cont2)),
"<=",6,indices=(M.Cont2),3)
add.constraint(lprec,rep(1,length(M.Cont3)),
"<=",3,indices=(M.Cont3),1)
add.constraint(lprec,rep(1,length(M.Cont4)),
"<=",6,indices=(M.Cont4),3)
```

```
add.constraint(lprec,rep(1,(length(append(M.answer.key.1,I.answer.key.1)))),
"<=",18,indices=c(M.answer.key.1,M+I.answer.key.1),5)
add.constraint(lprec,rep(1,(length(append(M.answer.key.2,I.answer.key.2)))),
"<=",18,indices=c(M.answer.key.2,M+I.answer.key.2),5)
add.constraint(lprec,rep(1,(length(append(M.answer.key.3,I.answer.key.3)))),
"<=",18,indices=c(M.answer.key.3,M+I.answer.key.3),5)
add.constraint(lprec,rep(1,(length(append(M.answer.key.4,I.answer.key.4)))),
"<=",18,indices=c(M.answer.key.4,M+I.answer.key.4),5)
add.constraint(lprec,rep(1,length(append(M.prevused2,I.prevused2))),
"<=",15,indices=c(M.prevused2,M+I.prevused2))
add.constraint(lprec,rep(1,length(append(M.prevused3,I.prevused3))),
"<=",15,indices=c(M.prevused3,M+I.prevused3))

#add psychometric constraints
#EQ 4.34,4.35

add.constraint(lprec,rep(1,length(append(M.FitFlag,I.FitFlag))),
"=",0,indices=c(M.FitFlag,M+I.FitFlag))
add.constraint(lprec,rep(1,length(append(M.BrailleFlag,I.BrailleFlag))),
"<=",2,indices=c(M.BrailleFlag,M+I.BrailleFlag))
add.constraint(lprec,rep(1,length(append(M.OmitFlag,I.OmitFlag))),
"=",0,indices=c(M.OmitFlag,M+I.OmitFlag))
add.constraint(lprec,rep(1,length(append(M.LowPvalueFlag,I.LowPvalueFlag))),
"<=",6,indices=c(M.LowPvalueFlag,M+I.LowPvalueFlag))
add.constraint(lprec,rep(1,length(append(M.HighPvalueFlag,I.HighPvalueFlag))),
"<=",6,indices=c(M.HighPvalueFlag,M+I.HighPvalueFlag))
add.constraint(lprec,rep(1,length(append(M.LowPtBiserial,I.LowPtBiserial))),
"<=",0,indices=c(M.LowPtBiserial,M+I.LowPtBiserial))
add.constraint(lprec,rep(1,length(append(M.HiPtBisDist,I.HiPtBisDist))),
"<=",4,indices=c(M.HiPtBisDist,M+I.HiPtBisDist))
add.constraint(lprec,rep(1,length(M.prevusedpos.1)),
"<=",10,indices=(M.prevusedpos.1),5)
add.constraint(lprec,rep(1,length(M.prevusedpos.2)),
"<=",7,indices=(M.prevusedpos.2),5)
add.constraint(lprec,rep(1,length(M.prevusedpos.3)),
"<=",7,indices=(M.prevusedpos.3),5)
```

```
#add linking error constraints
#EQ 4.30, 4.31
#tobj.u and tobj.v are calculated in the data file as
#1/(sigmau^2) and 1/(sigmav^2),
#respectively

add.constraint(lprec,M.dat$tobj.u,
">=",619,indices=(1:M))
add.constraint(lprec,M.dat$tobj.v,
">=",234,indices=(1:M))

#solve the assembly problem

res_flag=solve(lprec)
res_flag

#get output

x_opt$=$get.variables(lprec)
x_opt
qa<-sum(x_opt)
sol<-get.objective(lprec)
con<-get.constr.value(lprec,side$=$c("rhs","lhs"))
val<-get.constraints(lprec)
```

Appendix C

Standard Error Results for Polytomous Model Simulation with N(0,1) Test-Taker Ability Distribution

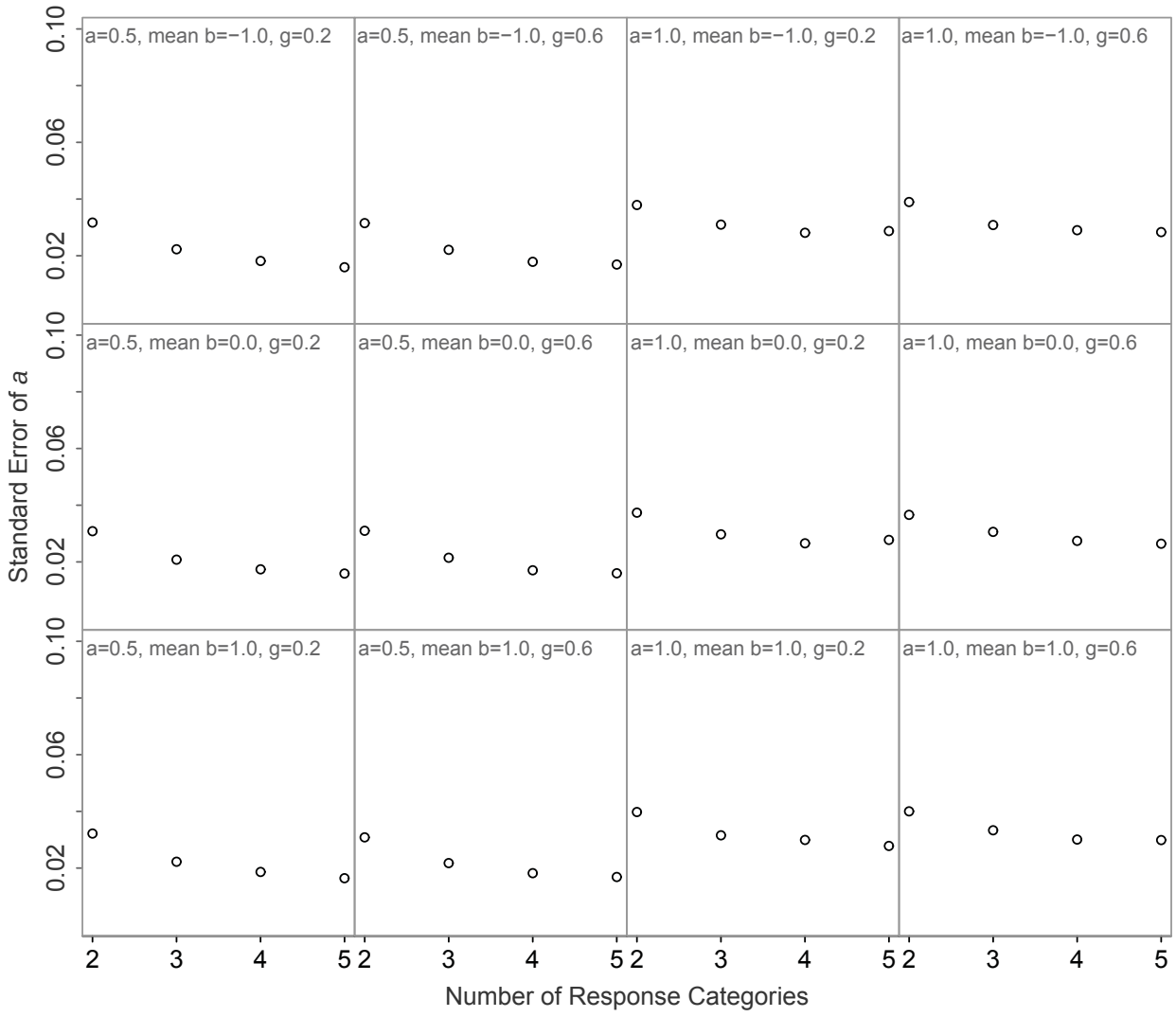

Figure C.1. Estimation error of a with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). A $N(0, 1)$ test-taker distribution was used for response simulation.

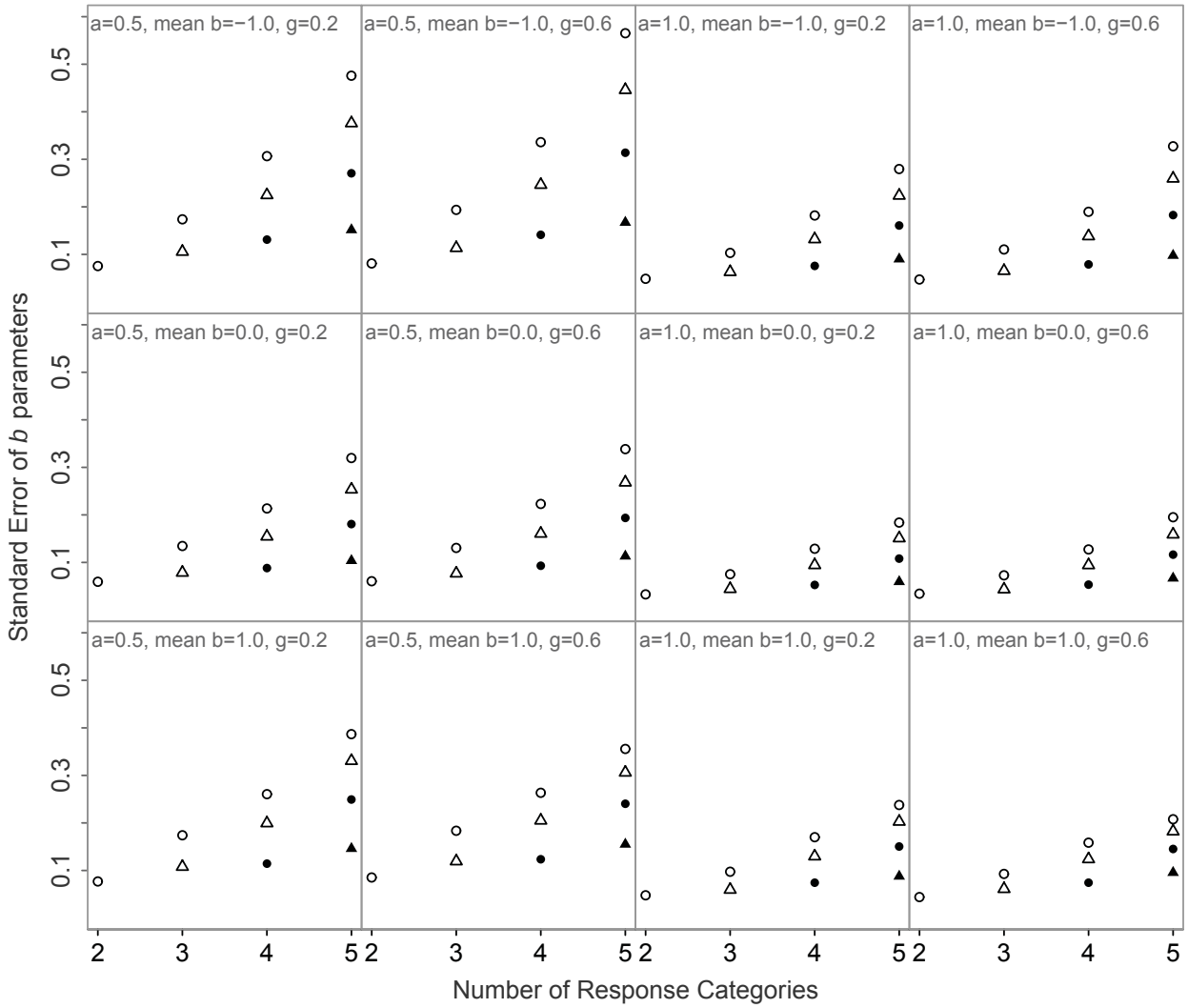

Figure C.2. Estimation error of b parameters with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). b_1 is notated by the clear circle, b_2 is notated by the clear triangle, b_3 is notated by the solid circle, and b_4 is notated by the solid triangle. A $N(0, 1)$ test-taker distribution was used for response simulation.

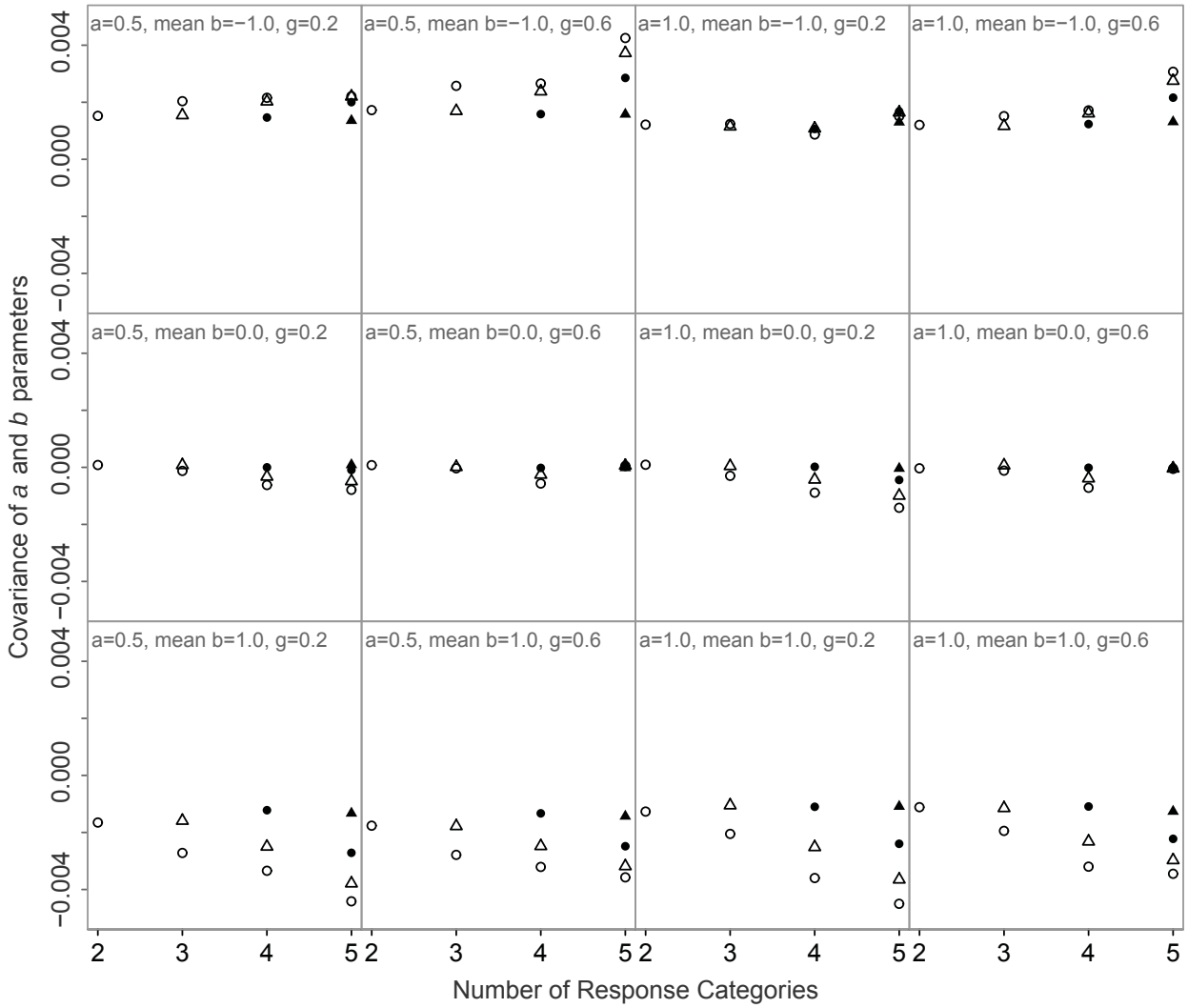

Figure C.3. Covariance of a and b parameters with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). b_1 is notated by the clear circle, b_2 is notated by the clear triangle, b_3 is notated by the solid circle, and b_4 is notated by the solid triangle. A $N(0, 1)$ test-taker distribution was used for response simulation.

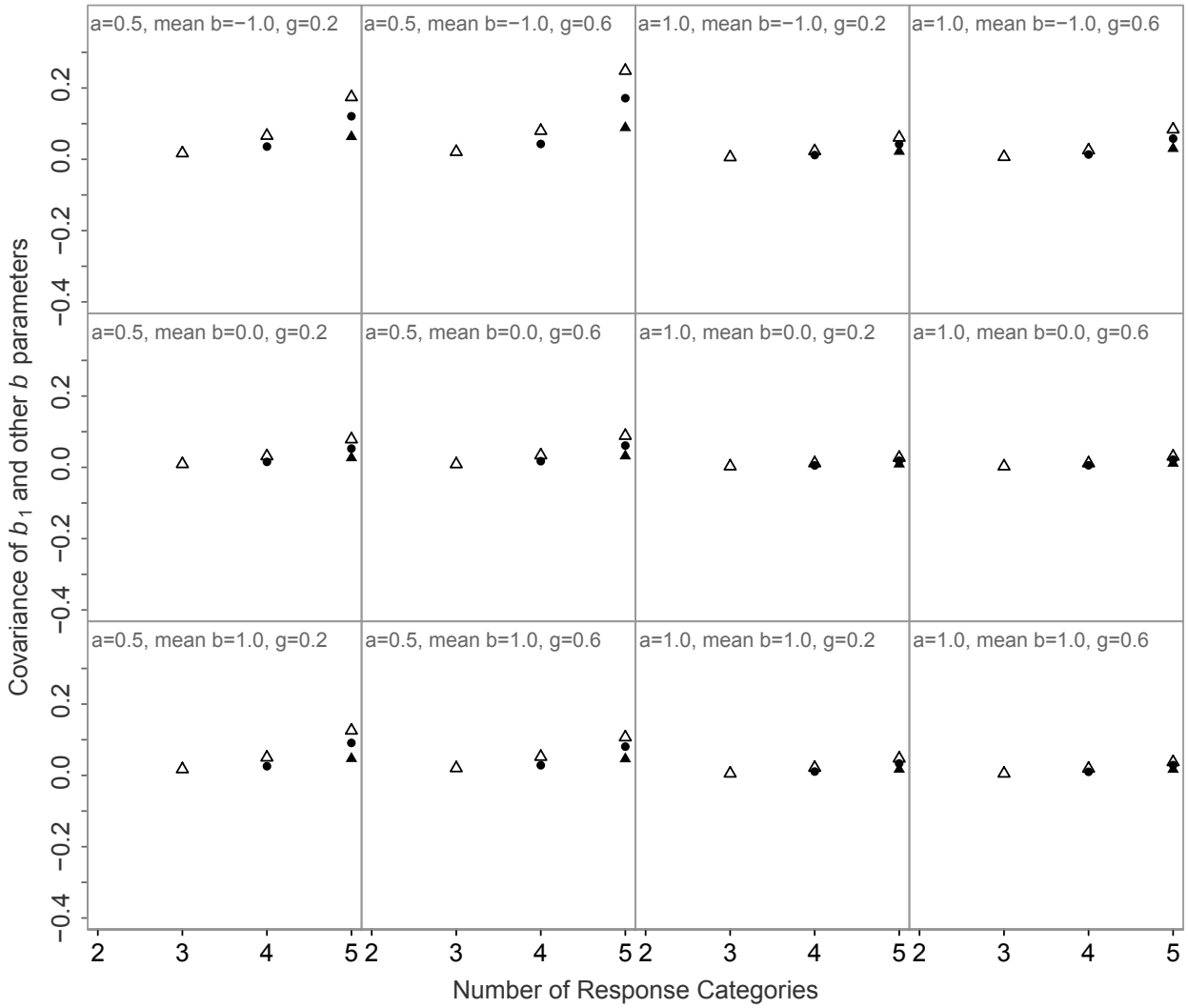

Figure C.4. Covariance of b_1 and other b parameters with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). b_2 is notated by the clear triangle, b_3 is notated by the solid circle, and b_4 is notated by the solid triangle. A $N(0, 1)$ test-taker distribution was used for response simulation.

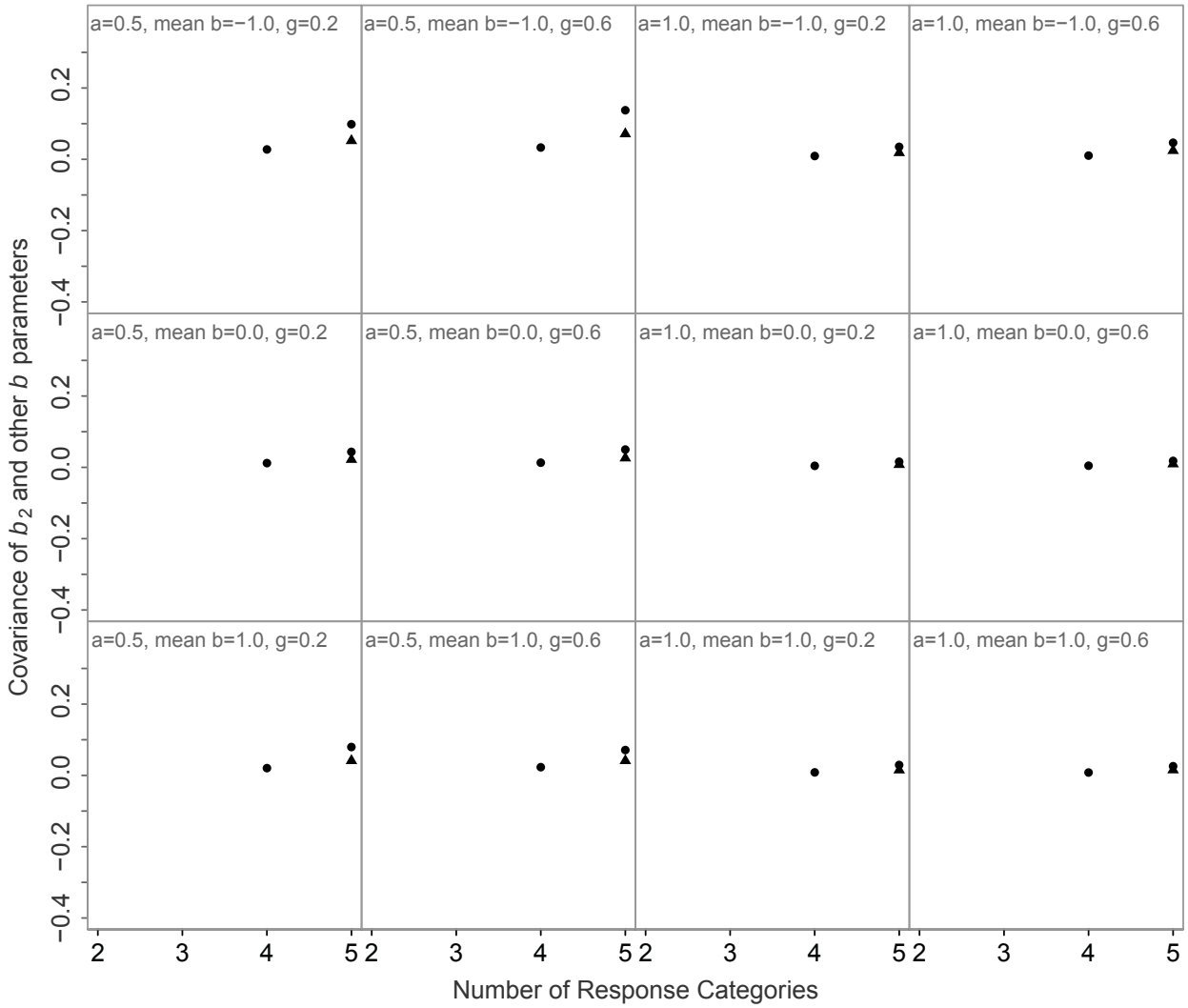

Figure C.5. Covariance of b_2 and other b parameters with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). b_3 is notated by the solid circle, and b_4 is notated by the solid triangle. A $N(0, 1)$ test-taker distribution was used for response simulation.

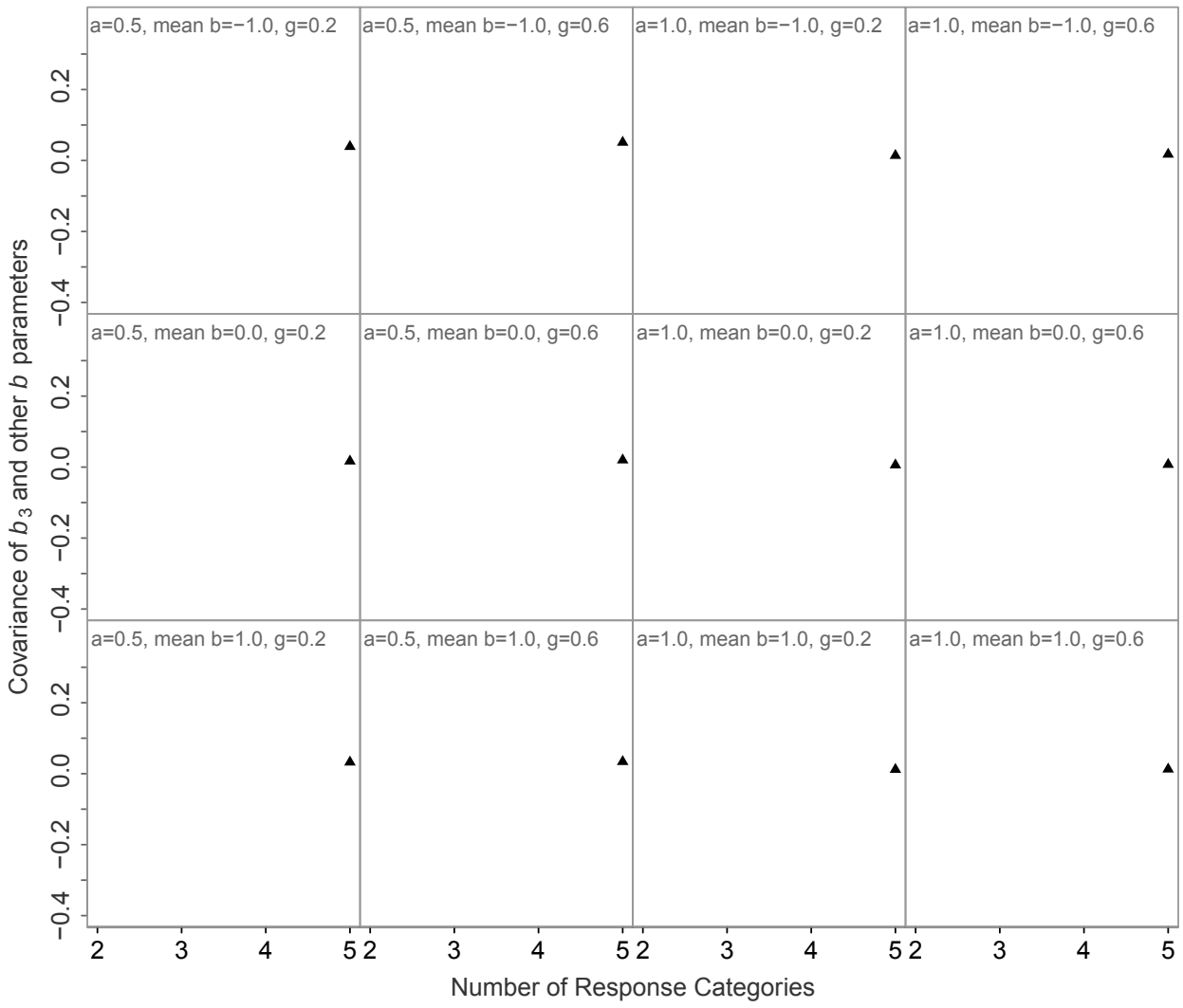

Figure C.6. Covariance of b_3 and b_4 parameters with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). A $N(0, 1)$ test-taker distribution was used for response simulation.

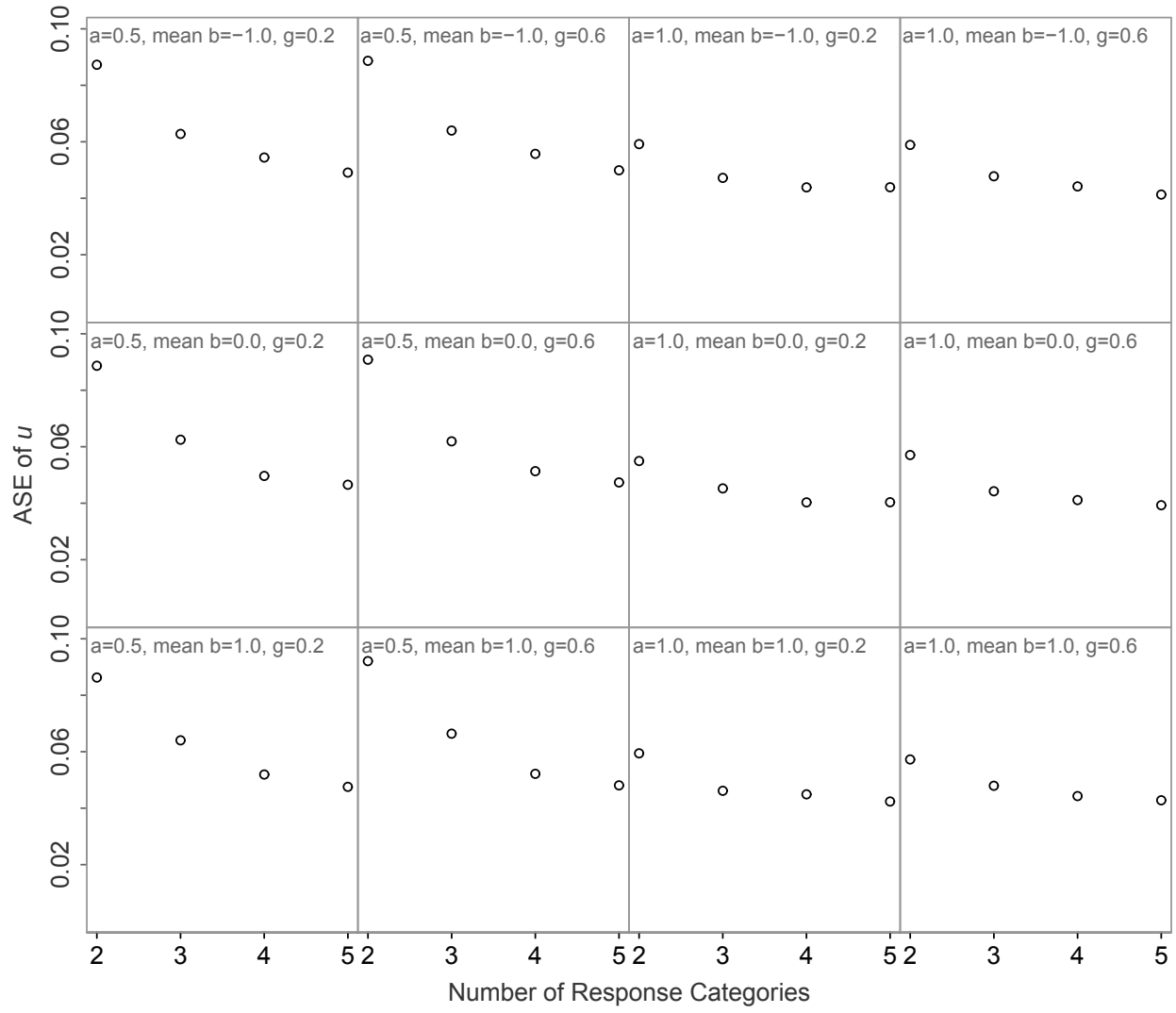

Figure C.7. ASE of u with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). A $N(0, 1)$ test-taker distribution was used for response simulation.

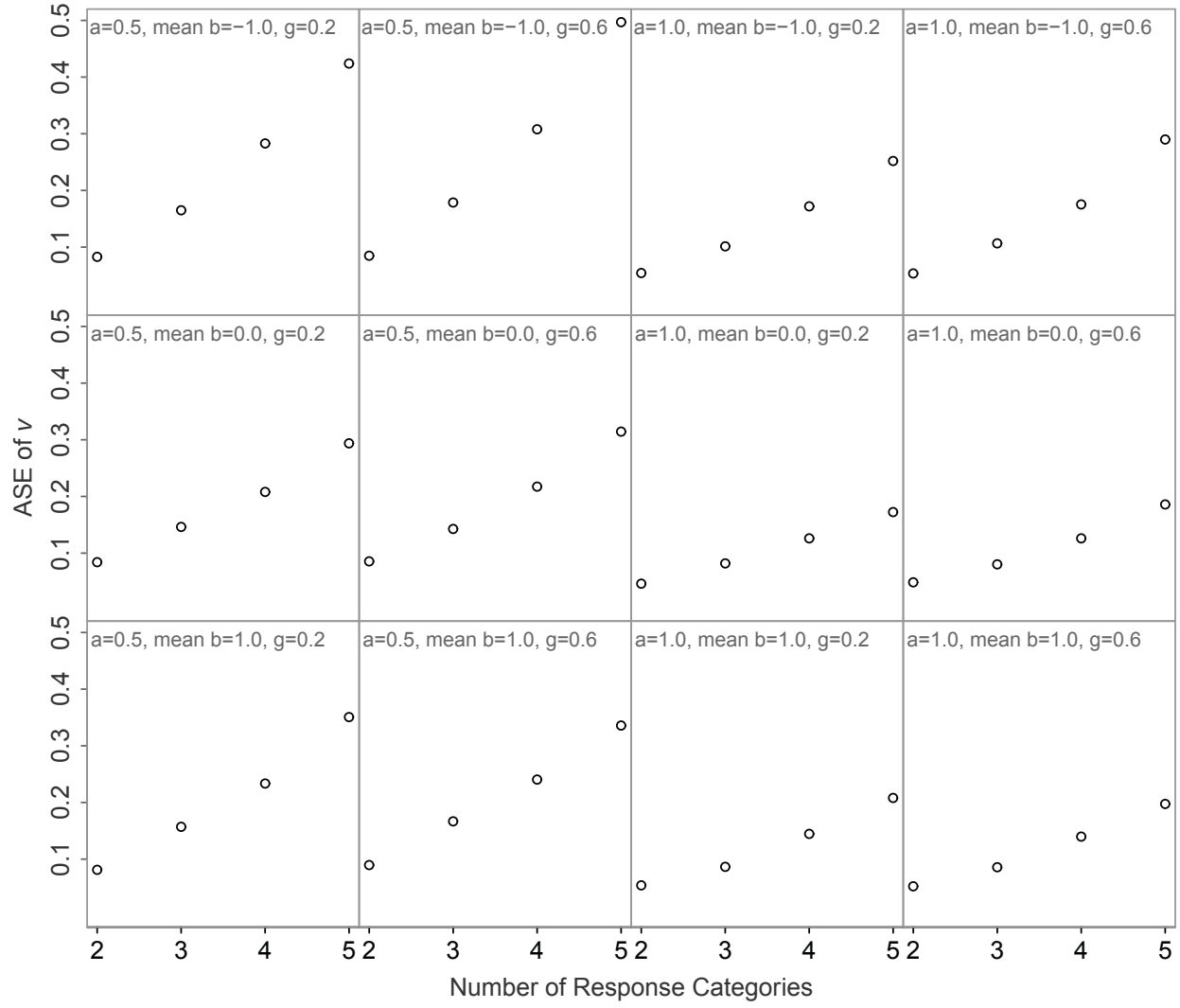

Figure C.8. ASE of v with number of response categories for 48 items with varying true discrimination (a), average of the items true b parameters, and distance between each of the items true b parameters (g). A $N(0, 1)$ test-taker distribution was used for response simulation.

Summary

If two tests comprised of different items intended to measure the same construct are given to different groups of test-takers at two different points in time, users of the test results typically still want to make comparisons about the performance of those test-takers on the tests. The validity of the comparison rests on psychometric models used to describe the behavior of the different items and the behavior of the different test-takers. With a few exceptions, the problem of linking item response model parameters from different item calibrations has been conceptualized as an instance of the problem of equating observed scores on different test forms. This thesis argues, however, that the use of item response models does not require any post hoc *observed-score equating*, but that the necessity of parameter linking is due to a fundamental problem inherent in the formal nature of these models—their general lack of *identifiability*. More specifically, item response model parameters need to be linked to adjust for the different effects of the identifiability restrictions used in the different item calibrations.

Based on this premise, the research first characterizes the formal nature of linking functions for monotone, continuous, dichotomous response models. While the general shape of the linking function for the three-parameter logistic (3PL) model is found to be that of previous methods, new definitions emerge for the slope, and consequently, intercept linking parameters. In addition, derivation of the linking function for the c parameter reveals an identity function, bringing into question response function linking methods that confound estimation error in the c parameter with linking function parameter estimates. It is shown that to minimally identify the linking function for the 3PL model, a linking element comprised of a single common item, of a pair of common items, or of a pair of common test-takers is required.

Next, closed-form asymptotic standard errors (ASE) of the linking parameters are derived for the 3PL model at the level of the linking element. From this foundation, a precision-weighted estimator for the linking function is proposed, in which the contribution of linking elements with relatively higher ASE is reduced. The new estimator exhibits the desirable feature of monotone decrease in linking error as linking elements are added to the design. In addition, it is shown that the estimator outperforms the previous mean/mean and mean/sigma linking approaches, especially when a common item exhibits relatively large parameter estimation error in one or both calibrations.

With the new estimator, as it is then possible to quantify the contribution of a single linking element to linking error, optimal linking design becomes possible. Several examples of optimal

linking design– minimizing linking error through the optimal selection of the common items to serve as linking elements– are provided. The examples show the usefulness and simplicity of implementing such procedures to improve response model parameter linking operationally.

Finally, methods are extended to well-known polytomous response models. Again, the new estimator exhibits the desirable feature of monotone decrease in linking error as linking elements are added to the design. The new estimator also enables exploration of the relationship between the number of response categories of an item and the ASEs of the estimated linking parameters. For the generalized partial credit model, analysis is presented that suggests that asymptotic standard error for the slope parameter in the linking function will hardly change with the number of response categories of an item, while the error for the intercept parameter increases with the number of categories. Empirical examples are provided that concur with the analytic results.

Samenvatting

Wanneer twee tests bestaande uit verschillende items, die beide hetzelfde construct meten, bij verschillende groepen personen worden afgenomen, zullen gebruikers niettemin de prestaties van deze personen willen vergelijken. De geldigheid van deze vergelijking is afhankelijk van psychometrische modellen die worden gebruikt om het gedrag van de verschillende items en personen te beschrijven. Met slechts een paar uitzonderingen, is het probleem van het linken van item-responsmodel parameters uit verschillende calibratiestudies opgevat als een instantie van het probleem van het equivaleren van gobserveerde scores op verschillende tests. In dit proefschrift wordt evenwel betoogd dat het gebruik van item-responsmodellen nooit post hoc equivaleren van geobserveerde scores met zich meebrengt, maar dat de noodzaak van parameter linking eigen is aan een fundamenteel probleem dat inherent is aan de aard va deze modellen- hun algemeen gebrek aan identificeerbaarheid. Meer specifiek, item-responsmodel parameters moeten gelinkt worden om te corrigeren voor de verschillende effecten van de identificeerbaarheidsrestricties die in verschillende calibratiestudies worden gebruikt. Op basis van deze premisse, karakteriseerde het onderzoek in dit proefschrift eerst de fomele aard van linkfunkties voor monotone, continue, dichotome responsmodellen. Alhoewel de algemene vorm van deze linkfunkties voor het drie-parameter logistische (3PL) model blijkt te voldoen aan de vorm die in bestaande methoden al werd aangenomen, werden in het onderzoek nieuwe definities voor de parameters in deze linkfuncties afgeleid. Daarnaast leverde de afleiding van de linkfunktie voor de gisparameters een identiteitsfunktie op, hetgeen erop wijst dat de bestaande responsfunktie methoden voor het linken van parameters schattingsfouten in de gisparameters met fouten in de linkparameters verwart. In het onderzoek werd eveneens aangetoond dat om de linkfunkties te kunnen identificeren, een ontwerp voor een linkstudie met slechts en gemeenschappelijk item, n paar gemeenschappelijke items of n paar gemeenschappelijk personen volstaat. Vervolgen werden voor deze minimale ontwerpen de schatters en hun asymptotische standaardfouten voor de linkparameters voor het 3PL model afgeleid. Voor grotere ontwerpen werd voorgesteld om deze minimale schatters te combineren, met gebruikmaking van de inversen van hun standaardfouten als precisiegewichten. In tegenstelling tot de bestaande mean/mean en mean/sigma methoden voor parameterlinken, blijkt de nieuwe schatters de gewenste eigenschap van monotoniciteit in het aantal gemeenschappelijke elementen in de linkstudie te bezitten. Bovendien levert hij gunstigere standaardfouten op, speciaal wanneer een gemeenschappelijk element in de linkstudie een relatief grote schattingsfouten in de modelparameters oplevert. Met de nieuwe schatters, waarvoor de bijdrage van ieder gemeenschappelijk element in de linkstudie expliciet is, was

het vervolgens mogelijk om het ontwerp van linkstudies te optimaliseren. Verschillende voorbeelden van optimale ontwerpen werden voorgesteld, die ieder de aanwezigheid van linkfouten minimaliseren. Alle voorbeelden lieten zien hoe nuttig en eenvoudig het is om deze optimaliseringsprocedures in operationele linkstudies toe te passen. Tenslotte werden alle methoden gegeneraliseerd voor gebruik met polytome reponsmodellen. Ook voor deze toepassing blijken de standaardfouten voor de nieuwe schatters de gewenste monotoniciteit in het aantal gemeenschappelijke elementen in het ontwerp van de linkstudie te bezitten. Vervolgens werden de resultaten gebruikt om de invloed van het aantal responsalternativen in het model op de standaardfout voor de geschatte linkparameters te exploreren. Voor het gegeneraliseerde partial credit model worden analyses gepresenteerd die suggeren dat de asymptotisch standarddfout voor de hellingsparameter in de linkfunctie nauwelijks afhangt van het aantal alternatieven, terwijl de fout voor de interceptparameter toeneemt met het aantal alternatieven. Empirische gegevens die deze analytische resultaten ondersteunden, konden worden gepresenteerd.

Curriculum Vitae

Michelle Derbenwick Barrett is the Director of Assessment Technology at Pacific Metrics Corporation, an ACT technology company, where she leads a team of research scientists and software engineers in the development of innovative assessment tools and services. Michelle started her doctoral work in 2011 at the University of Twente under the supervision of Prof. Dr. Wim J. van der Linden, while concurrently serving as the Director of Research Systems and Analysis at McGraw-Hill Education CTB. In that role, she led the development of systems to support psychometric analysis, automated test assembly, computer adaptive testing, and automated scoring. Prior to joining McGraw-Hill Education CTB in 2004, Michelle served as a Senior Consultant at the Colorado Department of Education, where she directed an Enhanced Assessment Grant for alternate assessment and coordinated Colorado's efforts on the National Assessment of Educational Progress, and as an Assessment Policy Analyst at the Colorado Commission on Higher Education. Michelle also previously taught middle- and high-school mathematics and worked as a manufacturing engineer for a surgical device company.

Michelle holds a master's degree in education from the Harvard Graduate School of Education (2001) and a bachelor's degree in mechanical engineering from Stanford University (1996). She also holds a graduate certificate in large-scale assessment from the University of Maryland, College Park (2003).