Research Report

Linking Competencies in Educational Settings and Measuring Growth

Alina A. von Davier

Claus H. Carstensen

Matthias von Davier

Research & Development

June 2006
RR-06-12

Linking Competencies in Educational Settings and Measuring Growth

Alina A. von Davier
ETS, Princeton, NJ

Claus H. Carstensen
IPN, Kiel, Germany

Matthias von Davier
ETS, Princeton, NJ

June 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Measuring and linking competencies require special instruments, special data collection designs, and special statistical models. The measurement instruments are tests or tests forms, which can be used in the following situations: The same test can be given repeatedly; two or more parallel tests forms (i.e., forms intended to be similar in difficulty and content and that still need to equated) can be given at different time points; or two or more test forms that may be less parallel to various degrees can be given at different points in time. In some circumstances, the goal of the analysis is to make the scores of parallel test forms comparable across different time points and different samples of relatively similar ability (horizontal equating). In other situations, we aim at the comparability of scores of tests forms that are not parallel, and although they are intended to measure the same competencies, they are of different difficulties and are taken by samples that show large differences in ability (vertical linking). In other cases, we want to evaluate the change in competencies over time (with or without covariates) for the same individuals measured by the same instrument or by different instruments (longitudinal linking). This paper will briefly discuss a variety of techniques for relating scale scores from different data collections points and will then discuss models for measuring growth. Each of these areas is a large field in itself, and each has a potential strong impact on educational policies such as the No Child Left Behind Act (for example, the vertical linking of state or national assessments and longitudinal studies are a potential basis for informative analyses for the policy makers), on the life of students and parents (equating of achievement tests), or on the life of professionals (equating of licensure tests). In these times when more and more standardized testing is used for assessing competencies in different domains nationally and internationally, we are also discovering more challenges in ensuring that the process and the results are fair and accurate. In turn, these challenges and these new social implications open the door toward more research in support of fair assessments, both in improving upon the test construction process and in advancing the statistical methods involved.

Key words: Comparability of scores, vertical linking, horizontal linking, longitudinal linking

Acknowledgments

The authors are thankful to Dan Eignor, Wendy Yen, Anne Fitzpatrick, and Anna Kubiak for their suggestions and comments on the previous version of the manuscript. The authors also thank Kim Fryer for editorial help. The opinions and conclusions contained in this paper are those of the authors and do not necessarily reflect the position or policy of ETS, IPN, or of the colleagues who reviewed the previous version of the manuscript.

Table of Contents

Page

List of Tables

Page

1. Introduction

All linking processes have basically three components: a measurement instrument, a data collection design, and a statistical method for achieving the desired form of comparability of two or more test forms. In this paper we will differentiate between horizontal, vertical, and longitudinal linking based on the first two components; the third component will be addressed by describing general tools used within any of the linking processes. The term *linking* is used with slightly different meanings in the field of educational assessments: (a) as a general term for denoting a relationship between test forms (at the total score level, at the parameter level, etc.); (b) as a weaker form of equating; and (c) as a synonym to the item response theory (IRT) item parameters calibration. In this paper, the term *linking* is mostly used as in (a) and (c) above.

Also, in this paper we take the view that modelling change or growth over time with or without covariates relies on an implicit assumption that the competency measured is scaled across the measurement points, and therefore, we present the models for measuring growth as a higher level built on top of the linking stage.

This paper focuses on two types of research questions that arise in educational measurement: (a) how to compare the competency measured by two test forms given to different groups of examinees, and (b) how to measure change in the competency measured by the same test form or two or more test forms given to the same group or different groups of examinees at different points in time.

In order to answer each of these two questions, the outcome variable (i.e., the competency measured) must be comparable across time points (i.e., each scale point of the measure must retain an identical meaning over time). The outcome variable must also remain construct-valid for at least the entire period of observation and maybe beyond, if prediction of future success is part of the study. These two requirements speak to the necessity of establishing a common scale for the measurement instruments/tests we use, or, in other words, establishing a link between the test forms (see Harris, Hendrickson, Tong, Shin, & Shyu, 2004; Patz, Yao, Chia, Lewis, & Hoskens, 2003). This is true in educational contexts even if we use the same test form at different points in time, if the construct itself might evolve over time.

When do questions like these arise? For example, in achievement and licensure tests the meaning of the reporting scale should be preserved across administrations and the fairness for the examinees that take different test forms should be insured. In such a process, also called a

horizontal equating process, two test forms *X* and *Y* that are built to be parallel are placed on the same scale. By parallel test forms, we mean here test forms that were constructed following the same content and statistical specifications, that were intended to be similar in difficulty, and that were built with the intention of measuring the same construct. However, since the construction of parallel test forms is never perfect, test equating is used to make the scores on the two tests forms interchangeable. This form of linking of two parallel test forms is called equating and is a preliminary step in answering the first research question above. In the (horizontal) equating designs, even if the groups of examinees who take the parallel forms differ in ability, these differences are usually small. Equating is the most stringent type of horizontal linking, and it refers specifically to linking/equating scores. Other types of (weaker) horizontal linkings are concordances (where tests that are not built to be parallel are linked on relatively similar populations of test takers—see Holland, Dorans, & Petersen, in press).

In other circumstances, we need to make scores of test forms—that measure the same domain or construct but differ in difficulty—comparable across years of study, to enable measurement of growth in a particular domain. This particular type of linking is often called *vertical linking* in the field of educational measurement. In a vertical linking design, the examinees that take the two (or more) test forms are from representative samples of their cohorts, the examinees are assessed at the same point(s) in time, the samples of examinees that take different test forms differ significantly in ability, and the forms to be linked are not parallel (Harris et al., 2004; Harris & Hoover, 1987; Kolen & Brennan, 2004; Yen & Burket, 1997). Vertical linking is a preliminary requirement for answering the second research question. Many elementary and secondary test batteries report scores on a vertical scale, such as Iowa Test of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001, 2003) or ACT's Educational Planning and Assessment System (EPAS; ACT, Inc., 2000). Vertical linking is a weaker type of linking than equating.

In addition to the two practical circumstances described above, there are assessments that focus on individual development over time under a particular type of treatment or educational exposure. Such situations arise in formative assessments as well as in summative assessments. In a longitudinal linking process, the measurement instrument might be the same, might be parallel test forms, or might be forms that are less parallel across time points but are taken by the *same* individuals at different points in time.

Longitudinal linking in a sense is related to vertical linking but it requires a more restrictive design. In the longitudinal (panel) design, the same subjects are asked to respond the same or different measurement instruments/tests at different points in time. This approach enables measurement of growth for the individual examinees as a function of time and (linked) test score or as a function of other variables of interest. The difference between longitudinal linking and vertical linking resides in the difference between collecting data from a longitudinal design and collecting data from a cross-sectional design. The instrument of measurement might or might not be the same in both designs. The linking that underlies longitudinal studies in the educational context is mostly carried out in research mode, given the costs associated with testing the same subjects over time. However, there are valuable longitudinal studies in the educational field, such as the Early Childhood Longitudinal Study (ECLS) described in Rock, Pollack, and Weiss (2004). The ECLS attempts to identify different patterns of cognitive growth in kindergarten and first grade associated with selected subpopulations based upon a nationally representative longitudinal sample. Of special interest in the Rock et al. study is the estimation of the growth rates of subpopulations of children that are considered to be at risk educationally and/or economically. The authors use vertically equated measurement instruments in their longitudinal analysis. Also, it is worth noting the increase in supply and demand for formative classroom assessments (Stanford Learning First from Harcourt Assessment, Inc., 2006; *System 5*™ from ETS, 2006a), which emulate a longitudinal design, in the sense that the same examinees are tested in a classroom environment over time. The items used for these test forms sometimes come from precalibrated item banks and therefore, in these cases, no linking takes place after the tests have been administered. Even as such, we envision that these valuable longitudinal types of data will lend themselves well to studies for measuring growth. A caveat is that these data is subject to additional statistical challenges such as estimating models with small sample sizes.

It is common to use the terminology of vertical linking in the field of education, specifically when the interest is in measuring growth in a particular domain that is taught over several years in primary education (therefore, vertical). In this context, two or more measurement instruments (i.e., test forms) are used; the test forms differ since they cover the domain at different time points. In some of these circumstances, the vertical scale is used to express the growth in a particular domain as a function of the scaled scores of the subjects and the time

points. It is worth noting that it is only recently that schools and schools districts have started to consider the use of vertical scales for measuring growth in performance, although some vertical scaling procedures have been used with elementary achievement test batteries, such as ITBS (Hoover et al., 2001, 2003).

In contrast, longitudinal linking, a terminology used in many other fields (medicine, biology, economics, etc.), often uses the same measurement instrument over time or, if not, it assumes that a common scale has been created and attempts to explain the intra- and interindividual differences by the means of complex explanatory models, among which the best known are: (explanatory) IRT models, hierarchical (linear) models (HLM), and structural equation models (SEM).

In this paper we give an overview of the existing methodologies for establishing a common scale for the three types of settings—horizontal, vertical, and longitudinal—with the scores reported at the individual level. This paper also addresses briefly the linking or trend analyses conducted in large-scale assessments that use a matrix design for collecting data and report results at the group level; we will describe this data collection design and the calibration methods usually employed in these assessments because the matrix design is also used for other linking purposes.

The first part of the paper focuses on equating and vertical linking and the second part focuses on measuring growth. This paper provides a theoretical and descriptive overview and does not focus on practical examples. For details and examples of vertical linking, the reader should consult Kolen and Brennan (2004).

In the next section, we will describe the data collection designs that are usually employed for linking. Then we will give an overview of the statistical methods used to achieve score comparability.

2. Data Collection Designs

The role of the data collection design is slightly different for observed-score linking as compared to IRT linking. In the framework of observed-score linking, the role of the data collection design is to disentangle the differences in the test forms from the differences in the abilities/competencies of the examinees that take those test forms. In consequence, there are two major classes of data collection designs: those where the examinees come from the same population (equivalent groups [EG] design, single group [SG] design) and those where the

examinees are drawn from different populations and a set of items, called an anchor test, is taken by the test takers from these populations (non-equivalent groups design with anchor test [NEAT] design). In an IRT framework, which is built around the claim that there is a separation of the influence of items and persons on the item responses,[1] the main purpose of the data collection design is to place the item parameters from test forms on the same scale (which implicitly adjusts for differences in ability). In the IRT context, the role of the data collection design is more complicated because it also has to address the intrinsic indeterminacy of the IRT models and therefore, the need for scale alignment, in addition to the differences in the test forms.

In consequence, in the NEAT design and for observed-score equating, the anchor set is used to adjust for the possible differences in ability of the groups of examinees. In the IRT context, the items from the anchor test are used for aligning the scales, assuming that the IRT model fits the data from the two populations well (and of course, implicitly adjusting for the differences in ability, since the joint origin of the aligned scales is defined by the anchor test items).

These three data collection designs (EG, SG, NEAT) are the basis for building more complex designs as will be shown in this section. How these three data collection designs are used for horizontal linking and equating will be described in detail in the following section. Then, the data collection designs for vertical linking and the matrix design for large-scale educational surveys and other assessments will be described. Longitudinal studies can use any of or a combination of these designs, and the restriction that the same individuals are tested over time represents a part of the design that establishes a link across data collection points.

2.1 Equivalent Groups (EG) and Single Group (SG) Designs

Assume there are two test forms, *X* and *Y*, for which scores are to be linked or equated on a target population of examinees. In the EG design, there is only one population of test takers, and two samples are (randomly) drawn from it. Each sample of test takers takes one test form, *X* or *Y*. This design requires large samples for stable equating or linking results. In the SG design, one sample randomly drawn from a target population of examinees takes both tests forms to be equated. The sample size required for an accurate equating in an SG design is much smaller, since the correlation between the two tests is taken into account in computing the standard errors of equating. However, the underlying assumption in an SG design is that there are no order effects associated with the test taken second. The data structure for the EG and SG designs is

illustrated in Tables 1 and 2 (see also von Davier, Holland, & Thayer, 2004). These data collection designs are used in horizontal linking and in equating.

These two designs are also the basis for the other types of linking: vertical and longitudinal linking. Usually, these two designs are combined in various ways as will be described next (see also von Davier et al., 2004).

Table 1

Data Collection in an Equivalent Groups (EG) Design

Population	Sample	*X*	*Y*
P	1	√	
P	2		√

Table 2

Data Collection in a Single Group (SG) Design

Population	Sample	*X*	*Y*
P	1	√	√

2.2 Non-Equivalent Groups With Anchor Tests (NEAT)

In the NEAT design (see, for example, von Davier et al., 2004; Kolen & Brennan, 2004), there are two populations, *P* and *Q*, of test takers, and a sample of examinees from each. The sample from *P* takes test *X*, the sample from *Q* takes test *Y*, and both samples take a set of common items, the anchor test, *V*. In observed-score equating and horizontal linking, the common set of items is used to adjust for the differences in ability between the two non-equivalent groups; in an IRT context, the common set of items is used for scale-linking purposes, which is a way of accounting for differences between populations through constraints on the item parameters in the anchor test. The NEAT design is often used when only one test form can be administered at one test administration because of test security or other practical concerns. The two populations may not be equivalent (i.e., the two samples are not from a common population). The NEAT designs can have internal (items in *V* are also part of both *X* and *Y*) or external (items in *V* are neither in *X* nor in *Y*) anchor tests. The data structure for a NEAT design is described in Table 3. Usually, when the data is collected following a NEAT design for horizontal linking or equating purposes, the differences in ability between *P* and *Q* are relatively small.

From Tables 2 and 3, we see that the NEAT design contains two independent SG designs, one on population *P* and one on population *Q*.

Table 3

Data Collection in a Non-Equivalent Groups With Anchor Test (NEAT) Design

Population	Sample	*X*	*Y*	*V*
P	1	√		√
Q	2		√	√

2.3 Data Collection Designs Used in Vertical linking

In vertical linking, the designs used for data collection are combinations of the three designs described above. Usually, there are three categories of such combination designs: non-equivalent groups that have levels with (a) overlapping content, (b) no overlapping content, or (c) common items and a scaling test (Patz, 2005).

In order to better adjust for the differences in competency at the extreme levels (in an educational context that is often a lower grade, such as grade three or four, and the highest grade), a combination of the designs is used in some cases, as described in Tables 4, 5, and 6.

In Table 4, we describe a version of the NEAT design used in vertical linking. The vertical linking process will use the common sets of items that overlap on two levels (*Y, Z, W*) to place scores on the test forms *X, Y, Z, W, S* (which are constructed to measure the same construct but are of different difficulties) on the same scale; in this way, each test taker can be placed on the common scale. By doing this, the study of growth in a particular domain for each individual becomes possible (under appropriate assumptions and when the whole process is carried out carefully).

Table 5 shows a case where a group of examinees from two adjacent levels takes two test forms and a group from the next two adjacent levels takes two test forms, one of them being the same as in the previous adjacent levels. In both cases described in Tables 4 and 5, the groups of examinees, P_1 to P_4, who take different test forms, differ in ability.

The design from Table 5 is enhanced by randomly assigning the test takers from each grade to take the common set of items that will be used to link back to the previous level or to one that will be used to link forward to the next level, in which case a combination of the designs described in Tables 1, 2, and 3 is used.

Table 4

Example of Data Collection Design With Overlapping Content Used for Vertical Linking (or Common Items)

Population	Sample	*X*	*Y*	*Z*	*W*	*S*
P_1	1	√	√			
P_2	1		√	√		
P_3	1			√	√	
P_4	1				√	√

Table 5

Example of Data Collection Design With No Overlapping Content Used for Vertical Linking

Population	Sample	*X*	*Y*	*Z*	*W*	*S*
P_1 & P_2	1	√	√			
P_2 & P_3	1		√	√		
P_3 & P_4	1			√	√	
P_4 & P_5	1				√	√

In Table 6, we describe another version of the NEAT design used in the vertical linking process (this design is usually called a scaling test—see Kolen & Brennan, 2004; Patz, 2005). Here there is a set of common-items for each pair of adjacent levels to adjust for the differences in the populations from adjacent levels and will place the test forms *X*, *Y*, *Z*, and *W* (which again are constructed to measure the same construct but are of different difficulties) on the same scale. The scaling test *V* contains items that are appropriate to all levels. Obviously, there are numerous challenges associated with its construction because of the required appropriateness of the construct and difficulty across all levels.

Table 6

Example of Data Collection Design With a Scaling Test Used for Vertical Linking

Population	Sample	*X*	V_1	*Y*	V_2	*Z*	V_3	*W*	*V*
P1	1	√	√						√
P2	2		√	√	√				√
P3	1				√	√	√		√
P4	2						√	√	√

The vertical linking has its challenges at different stages of the process: From Table 6, we can see that constructing a set of common items that covers several adjacent levels is a challenge

and might adjust less appropriately for the differences in competency from the extreme levels. There are other challenges for vertical linking that are not addressed here, such as how to construct the tests, the anchors, and the scaling test; how to address the possible shifts in construct; and how to introduce new test forms.

2.4 Balanced Incomplete Block (BIB) Designs

In large-scale educational survey assessments as well as for some state assessments, balanced incomplete block (BIB) or matrix designs are used. BIB designs assume that the different test forms are assigned to the same population (not different populations as in the NEAT design) and that every student only takes a subset of the items, called blocks of items, that are combined in intrinsic ways. Such data collection designs are used in all major international survey assessments such as TIMSS, PISA, and PIRLS and are based on several assumptions. The data collected from these designs are analyzed using extensions of IRT methods that were first introduced by ETS in projects for the National Assessment of Educational Progress (NAEP; Allen, Jenkins, & Schoeps, 2004). We will discuss these procedures in detail later. Because not every test taker takes all the items, these procedures are mostly used for large-scale survey assessments, where the reported results are at the group level (as opposed to individual level, as in the cases discussed above).

However, there are assessments where individual scores are reported that also use a matrix design, as is the case of Massachusetts Comprehensive Assessment System (MCAS), which is a state assessment (Massachusetts Department of Education, 2001):

> The matrix-sampled items are used to equate test forms across MCAS administrations and to field-test items for possible future use as common items. In *School* and *District Reports,* the *Subject Area Subscore* pages include data generated from responses to both common and matrix-sampled items: this is the **only** instance in which matrix-sampled items are used to generate MCAS results. All other school and district results are generated from student responses to common items only. (p. 9)

Table 7 shows a very small example of a balanced incomplete block design using 5 blocks, B_1 to B_5, in three positions combined into five booklets.

Table 7

Simple Example of a Balanced Incomplete Block (BIB) Design

Booklet	Position		
	1	2	3
1	B_1	B_4	B_3
2	B_4	B_2	B_5
3	B_3	B_5	B_2
4	B_2	B_3	B_1
5	B_5	B_1	B_4

In the balanced incomplete block design, all blocks are present in all three booklet positions, although all blocks are not combined with all other blocks, and not every test taker sees all the items. Usually, the number of booklets (and blocks) is much larger than in the example above, and blocks may contain items from multiple subdomains, such as reading, mathematics, and science. In such cases, the number of possible combinations is very large, and the number of booklets is comparably small, so that the incomplete pairing in the BIB design becomes an issue.

In cases with multiple domains, there may be booklets that only contain mathematics items (say M_1 to M_5), only science items (say S_1 to S_5), or only reading items (R_1 to R_5). This precludes the estimation of covariances between the three domains, but enables an estimation that is more accurate within domain. These focused booklets may be paired with booklets that contain at least two different domains, say reading and mathematics or science and mathematics, so that covariances between these subject matters can be estimated. This obviously means increasing the number of different booklets.

The logistics of administering booklets constructed using the BIB design is roughly as follows: The set of booklets is spiralled throughout the sample, so that classrooms tested within the assessment will receive a (pseudo-)random selection of the booklets (a similar approach as in an EG design). In this way, when the set of booklets are given out to the examinees during test administration, clustering of booklets will be minimized and approximately the same number of booklets will be given to approximately equivalent subsamples.

Either IRT and traditional equating methods can be used for most of the equating designs, including the NEAT design. However, if the data are collected following a BIB design, then the large missing-by-design data feature can be addressed only by the IRT methods. Similarly, in vertical linking settings, where the differences between the abilities of the groups of test takers

are large, the IRT methodologies are more appropriate for linking purposes than the observed-score methods, even though there are circumstances where observed-score equating methods have been used.

In this section we have described the data collection designs used in different types of linking processes. In the next section, we will provide a description of the statistical tools used to conduct the actual linking once the data have been collected.

3. Horizontal and Vertical Linking and Scaling

The role of the data collection designs is to provide suitable raw data for statistical methods aimed at producing scale linkages across time points, populations, or multiple test forms. In this section, we provide a succinct description of the statistical methods usually employed for equating and linking.

The statistical methods used to achieve score comparability may be classified as IRT methods (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; and many others), as classical or traditional equating methods that are based on observed scores (Kolen & Brennan, 2004) or as methods based on the classical test theory, such as the true-score Levine linear equating method (Kolen & Brennan).

If IRT is used in the equating or linking of the test scores, it is necessary to first use some sort of linking procedure or calibration method to place the IRT parameter estimates on a common scale (see von Davier & von Davier, 2004). Once this is accomplished, then additional methods, such as IRT true or observed-score equating (Kolen & Brennan, 2004) can be undertaken. Then a third step is employed, which refers to placing the raw scores onto some reporting scale. However, in many settings, mainly in vertical linking and in large-scale assessments, after the calibration is done, the ability or person parameter estimates are directly placed onto the reporting scale (Yen, 1984), and the second step is skipped altogether.

In this section, we will first focus on horizontal equating and linking. We will describe the IRT linking/calibration procedures (i.e., procedures for placing the item parameters on the same scale) used for data collection designs that involve common items (for the EG and SG designs, the linking of item parameters is achieved by calibrating jointly the data from the two test forms). In addition, the IRT true score equating will be briefly described. Then, traditional observed-score equating methods will be succinctly presented.

In the second part of this section, the scaling methodology used for vertical linking will be described in addition to IRT based methods used to calibrate and link in surveys that use BIB or matrix designs.

3.1 IRT Calibration Methods for the NEAT Design

In this section, we describe the IRT calibration or IRT linking methods used in the NEAT design. More exactly, we discuss the usual IRT-linking methods: mean-sigma and mean-mean, concurrent calibration, fixed parameters calibration, the Stocking and Lord characteristic curves approach, and the Haebara characteristic curves approach (see Kolen & Brennan, 2004, chapter 6, for a detailed description of these methods). The item calibration procedures for the matrix design will be discussed in section 3.2.

Unidimensional IRT models express the probability of a response z_{ni} of a given person, n (n = 1,…, N), to a given item, i (i = 1,…, I), as a function of the person's competency or ability (which is a latent variable), denoted θ_n, and a possibly vector valued item parameter, β_i, that is,

$$P_{ni} = P(X = z_{ni}) = f(z_{ni}, \theta_n, \beta_i)$$

In the case of the well-known three-parameter logistic model (3PL) model (Lord & Novick, 1968) that is used to fit data from dichotomous items, the item parameter vector is three-dimensional. Its dimensions are the slope or discrimination that is usually denoted by a, the difficulty b, and the guessing parameter c, respectively, that is, $\beta_i^t=(a_i, b_i, c_i)$. However, most results presented here do not depend on the specific choice of the model and apply to models for both dichotomous and polytomous data.

Table 3 shows that in the NEAT design, X is not observed in population Q, and Y is not observed in population P. To overcome this feature of the NEAT design, all linking methods developed for the NEAT design (both observed-score and IRT methods) must make additional assumptions of a type that does not arise in the other linking designs. IRT-based linking makes the following assumptions:

Assumptions. The tests to be equated, X and Y, and the anchor, V, are all unidimensional (i.e., all items measure the same unidimensional construct), carefully constructed tests, in which the local independence assumption holds (Hambleton et al., 1991); the chosen IRT model fits the data well.

When conducting scale linking in the NEAT design, the parameters of the model from different test forms need to be placed on the same IRT scale. The methods available for the scaling or calibration purposes are mentioned below. Operationally, the necessary item calibrations are usually obtained using the marginal maximum likelihood (MML) estimation method for IRT models (Bock & Aitkin, 1981). In the past, the joint maximum likelihood (JML) method has been used; conditional maximum likelihood (CML) methods are used for one parameter logistic IRT models (1PL, Rasch model) or for 1PL models with fixed slopes such as the OPLM (Verhelst & Glas, 1995), and in some recent developments, empirical Bayesian estimations methods are used in the context of Markov chain Monte Carlo (MCMC) estimation (see Patz & Junker, 1999). Note that the approaches described below can be employed using any of the available statistical estimation methods: marginal maximum likelihood (which is implemented in, for example, the software package PARSCALE, Muraki & Bock, 1997), joint maximum likelihood (for example, the method used by the software package LOGIST, Wingersky, Barton, & Lord, 1982), and Bayesian inference with MCMC (which can be implemented using, for example, the software package BUGS, Spiegelhalter, Thomas, Best, & Gilks, 1995) methods.

Joint items calibration with no linking or separate calibration. Separate calibration in a NEAT design can be obtained in two ways: (a) by estimating jointly all the parameters given all the data but without restrictions on the parameters of the common items and treating the items that an examinee did not take as not administered or (b) by carrying out the estimations separately: The item and ability distribution parameters for population *P* are estimated given data, *X* and *V* in *P*, separately from the item and ability distribution parameters for population *Q* given data, *Y* and *V* in *Q*. The two methods provide similar results and accomplish no linking. The usual practice is to use the second approach and to do the calibration separately on the two populations. However, the first approach provides the opportunity for testing the hypothesis that the common items behave as common items as intended (see von Davier & von Davier, 2004, in press; or Glas, 1999).

Concurrent calibration. As an alternative, the item parameters from *X*, *V* (in both populations), and *Y* can be estimated jointly, assuming that the items in *V* are the same for both populations and coding the items that an examinee did not take as not administered, since these outcomes were unobserved and are missing by design.

Fixed parameters scale linking. In this method, common items whose parameters are known (for example, from a previous administration calibration or a separate calibration) are anchored or fixed to their known estimates during calibration of other forms. By treating the common item parameters as known and, therefore, not reestimating them, the remaining item parameters from the not common set of items in the two test forms are forced onto the same scale as the items with fixed parameters. This calibration procedure is more restrictive than concurrent calibration, and it is not appropriate for cases where the populations who take the two test forms differ significantly in ability.

Mean-mean and mean-var IRT scale linkings. If an IRT model fits the data, any linear transformation (with slope *A* and intercept *B*) of the theta scale also fits these data, provided that the item parameters are also transformed in the same way (see, for example, Kolen & Brennan, 2004, chapter 6). In the NEAT design, the most straightforward way to transform scales when the parameters were estimated separately is to use the means and standard deviations of the item parameter estimates of the common items for computing the slope and the intercept of the linear transformation. Loyd and Hoover (1980) described the mean-mean method, where the mean of the *a*-parameter/slope estimates for the common items is used to estimate the slope of the linear transformation. The mean of the *b*-parameter/difficulty estimates of the common items is then used to estimate the intercept of the linear transformation (see Kolen & Brennan). The mean-var IRT scale linkage (Marco, 1977) can obviously be implemented in the same way, with only a slight difference in the restrictions used. The means and the standard deviations of the *b*-parameters are used to estimate the slope and the intercept of the linear transformation. Both methods are seldom used in practice and tend to be inadequate if the groups taking the test forms differ in abilities.

Stocking and Lord scale linking. Characteristic curves transformation methods were proposed (Haebara, 1980; Stocking & Lord, 1983) to avoid some issues related to the mean-mean and mean-var approaches, such as the fact that various combinations of the item parameter estimates produce almost identical item characteristic curves over the range of ability at which most examinees score.

The Stocking and Lord IRT scale linking finds parameters for the linear transformation of the item parameters from the anchor set in one population (say *Q*) that matches the test characteristic function of the anchor in the reference population (say *P*).

Haebara scale linking. Haebara (1980) expressed the differences between the characteristic curves as the sum of the squared differences between the item characteristic functions for each item over the common items for examinees of a particular ability θ_n. The Haebara method is more restrictive than the Stocking and Lord method because the restrictions take place at the item level (i.e., for each item from the set of common items), while the Stocking and Lord approach poses a global restriction at the anchor test level.

von Davier and von Davier (2004) proposed a new perspective on IRT scale linking by viewing any linking function as a restriction function on the joint log-likelihood function based on all the data. Rewriting any linking as a restriction function and estimating the model parameters under this restriction implies a larger flexibility in the linking process—when dealing with vertical linking, for example. This new method can incorporate the modelling of growth, possibly expressed as a hierarchical structure on the item parameters in the anchor (discussed in more detail below). The approach presented in the paper by von Davier and von Davier (2004) may easily be extended to multidimensional IRT models, at least for simple structure multiscale IRT models (like the one used in NAEP and other large scale assessments).

3.2 IRT Calibration in the Balanced Incomplete Block Design

Operationally, matrix samples of item responses from balanced incomplete block designs are calibrated jointly using MML estimation of IRT models (Bock & Aitkin, 1981), since MML estimation does *not* require that all items be taken by all examinees. The number of items across all blocks in survey assessments reaches hundreds, and the number of examinees in nationally representative samples reaches hundreds of thousands. Due to computational constraints in the past, some assessment programs use only a subset of examinees to carry out item parameter estimation. This reduction of the sample used for estimation is no longer necessary when using MML estimation with recent computer hardware.

MML estimation of IRT models with a matrix design can be done both with Rasch type models (or 1PL model) and with two-parameter logistic (2PL) or three-parameter logistic (3PL) models, as well as mixed models for dichotomous and polytomous response data. The approach taken is a multistage estimation; Patz and Junker (1999) used the phrase *divide and conquer* for this approach. First step is estimating the structural parameters of the measurement model (item parameters), the second step is estimating the structural parameters of a population model (often

a latent regression on grouping variables) that may include a potentially large number of covariates; Then, the third step follows, which is concerned with estimating distributions, percentiles, and percentages above cut points for policy relevant subpopulations.

It is important to note that IRT models include homogeneity assumptions that may not be met by default in complex samples as well as samples from composite populations. Careful item selection, and if necessary, treatments of a small number of misfitting items such as splitting items (i.e., different items parameters have to be assumed in different groups to account for differential item functioning, DIF) or collapsing response categories are some of the tools used in quality control during item calibration.

Even though the mechanics of item calibrations have become easy with up-to-date software and hardware, the calibration of item parameters for matrix samples of item responses requires special attention with respect to response rates in balanced incomplete block designs. Rates of omitted responses across blocks, context effects, and effects based on block order are some of the threats to IRT calibration that can only be touched on in this paper. Technical reports of major assessment programs such as NAEP, PISA, and TIMSS contain chapters on quality control of item calibrations. von Davier, Sinharay, Oranje, and Beaton (in press) gave an overview of the current approach to estimating population characteristics in survey assessments. Obviously, some of the issues mentioned here represent challenges for all types of linking.

Linking across assessment cycles using data from balanced incomplete block designs is a nontrivial endeavour for several reasons. First, assessment cycles usually are 2 to 9 years apart, so that the difficulty of items may change between cycles. Second, not all blocks from previously administered assessments can be reused, since some of the items are usually released together with the reports on results of preceding cycles. In addition, assessment frameworks change and may contain additional subdomains or may require new item formats.

In the best case, only released blocks of items will be replaced with new blocks and a substantial percentage of blocks will be the same (and administered in the same form, same print, and same position) across assessment cycles. In that case, a joint calibration (that accounts for the potential nonequivalence of assessment cohorts by estimating separate ability distributions in a multiple group IRT model) of both the old and the new sample will accomplish the task, give and take a few items that need to be separated between cycles in the case that item difficulties were observed to be very different in the two cycles.

The reality of linking in large-scale and the specifics of the multiple steps of analyses and quality control are more complex than in other linking designs, especially if more than one previous cycle contains linkage blocks and several subdomains have to be linked. In less than ideal (but more realistic) cases, the assessment frameworks and the construct to be measured changes somewhat, so that it has to be established empirically whether link items still fit in the new framework, in conjunction with the newly introduced items and subdomains. After the joint calibration has been carried out, the new (in practice, slightly different) scale has to be linked back to the original scale used for the previous assessments.

3.3 Equating of Scores in the IRT Framework

If equating of scores is desired, different programs and testing companies use different techniques after the calibration process. In some programs, an interim step is performed, where additional methods, such as IRT true score or observed-score equating are used to link the raw scores on the two forms. After that, a third step is employed, which refers to placing the raw scores onto some reporting scale. In other programs, the second step is skipped altogether, and the ability scale is linked directly to the reporting scale (Yen, 1984) as briefly described below.

Raw score-to-scale score scoring table. Yen's method generates a raw score-to-scale score scoring table. Placing item parameters from a theta scale on a scale score metric is done by (a) defining appealing additive and multiplicative constants (e.g., 400, 40) at the beginning of a testing program, and (b) in future years, items are placed on the base θ–metric via linking (typically using the Stocking and Lord method), and then the (400, 40) transformation is applied (W. M. Yen, personal communication, December 14, 2005). Some testing programs do this transformation in one step, by setting the reference parameters for the Stocking and Lord method to the scale score metric.

Once item parameters are in the scale score metric, they can be used in item response pattern scoring or to generate a raw score-to-scale score scoring table, using a procedure like the one described in Yen (1984) for the 3PL model or using inverted test characteristic curves (TCCs). Additional methods such as true score equating are unnecessary.

This method is used in scoring tests of state assessments and published tests, such as TerraNova (CTB/McGraw-Hill, 2000).

IRT true score equating. In this subsection, we briefly describe the IRT true score equating that is used by certain testing companies. For example, this method is used at ETS for assessments such as the TOEFL® iBT (ETS, 2006b). We do not describe the IRT observed-score equating in this paper (a detailed description of it is given in, for example, Kolen & Brennan, 2004).

In a NEAT design equating process for tests with multiple-choice items, for example, the 3PL IRT model is separately fitted to the data from the two populations *P* and *Q* in Table 3. Then, the characteristic curve method (Stocking & Lord, 1983) is used to place the separately estimated item parameters onto a common scale. After all the item and person parameters are on the same scale, they are used to estimate the true score equating function. The number-correct true score for a given θ is obtained by summing the (conditional) probabilities over the number of items in the test. Then the true score equating method is used to obtain equivalent scores on *X* and *Y* (see von Davier & Wilson, 2005; Kolen & Brennan, 2004; Petersen, Kolen, & Hoover, 1989). According to this method, for a given true score of the new test form *X*, one finds (via iterative procedures, such as the Newton-Raphson method) the value of competency/ability θ and then computes the true score on the old form *Y*.

The IRT true-score equating requires that the tests are number-right scored, which involves an implicit assumption that there are no omits (see Kolen & Brennan, 2004). If the tests are formula-scored, then some sort of transformation is necessary; this transformation will treat the omits as wrong. The IRT true score equating introduces one more assumption: The relationship between the true scores holds also for the observed scores. This assumption has not been theoretically proved, but was confirmed in research studies (Lord & Wingersky, 1984). In IRT true score equating that uses the common 3PL IRT model, the lowest possible true score is the sum of the c_j, the so-called guessing parameters, and not 0. In this case, there is some arbitrariness of the results of the conversion of the observed scores that are outside the range of possible true scores on the new form *X* (see Kolen & Brennan for details). Finally, after the raw-to raw conversion has been obtained, the equated scores are placed on the reporting scale.

3.4. Observed-Scores Equating Methods

In contrast to IRT linking methods discussed above, the observed-score linking methods relate person measures of proficiency from two different tests, in terms of the total scores,

without specifying an explicit model for the person by item interaction. In this subsection, we briefly describe observed-score equating methods, both the traditional procedures and a newly proposed equating procedure.

The definition of score interchangeability for the observed-scores linking and equating methods is based on relationships between the moments (which leads to linear equating) or percentiles of the two score distributions (which leads to equipercentile equating) to be equated.

In addition to the two test forms to be equated, *X* and *Y*, we also have an explicit target population, *T*, on which the equating is to be done. In an EG design, the tests are given to two samples that are (assumed to be) randomly drawn from a population of examinees, *P* (in this case, we assume that, $T = P$—see Livingston, 2004, for a slightly different view and definition of a target population). The target population, *T* for the NEAT design, is assumed to be a *weighted average* of *P* and *Q*. *P* and *Q* are given weights that sum to 1; the population defined in (1) below is also called the synthetic population in the literature, to emphasize that units of it are not real test takers. This is denoted by

$$T = wP + (1 - w)Q. \tag{1}$$

The partition of *T* is determined by the weight *w* (see Angoff, 1971; von Davier et al., 2004; or Kolen & Brennan, 2004, for a discussion of the target population in the NEAT design and of the role of the weights).

Many observed-score equating methods are based on the equipercentile equating function. It is defined on the target population, *T*, as:

$$e_{Y;T}(x) = G_T^{-1}(F_T(x)) \tag{2}$$

where $F_T(x)$ and $G_T(y)$ are the cumulative distribution functions (cdfs) of *X* and *Y*, respectively, on *T*. Given that *X* and *Y* are discrete random variables, their cdfs are step functions. However, in order for this definition to make sense and to insure that the inverse equating function exists, we also assume that $F_T(x)$ and $G_T(y)$ have been made continuous or continuized so that the inverse functions exist for $F_T(x)$ and $G_T(y)$.

Several important classes of observed-score equating methods may be viewed as only differing in the way that the continuization of $F_T(x)$ and $G_T(y)$ is achieved. The traditional equipercentile equating method (also called the percentile rank method) uses linear interpolation

of the discrete distribution to make it piecewise linear and therefore, continuous. The kernel equating (KE; von Davier et al., 2004; Holland & Thayer, 1989) method uses a Gaussian kernel smoothing to approximate the discrete distribution by a continuous density function.

The equipercentile equating function leads to linear equating if o $F_T(x)$ and $G_T(y)$ have the same shape while differing only in mean and variance. The linear equating function, $\text{Lin}_{Y;T}(x)$, is defined by

$$\text{Lin}_{Y;T}(x) = \mu_{YT} + \sigma_{YT}((x - \mu_{XT})/\sigma_{XT}). \quad (3)$$

In Theorem 1.1 of von Davier et al. (2004), it is shown that any equipercentile equating function can be decomposed into the corresponding linear equating function and a nonlinear part.

The observed-score equating functions for the NEAT design, equipercentile and linear, also make assumptions in order to overcome the missing by design data, a feature of the NEAT design. The assumptions and the formulas for the classical linear and equipercentile equating are given in Kolen and Brennan (2004) and von Davier et al. (2004).

von Davier et al. (2004) viewed any observed-score test equating as having five steps or parts, each of which involves distinct ideas. They are: (a) presmoothing of the score distributions, (b) estimation of the score probabilities on the target population, (c) continuization of the discrete fitted score distributions, (d) computing the equating function, and (e) computing the standard error of equating and related accuracy measures. In some assessment programs, where the samples are very large, the practitioner might decide to skip the presmoothing step; however, von Davier et al. recommended the presmoothing even for these circumstances.

The kernel method of test equating. The KE method (von Davier et al., 2004; Holland & Thayer, 1989) is an observed-score equating procedure that promises to unify several observed-score methods of test equating into a single method while, at the same time, providing new statistical information that can be used in the practice of test equating.

KE brings together these five steps into an organized whole rather than treating them as disparate problems. KE exploits presmoothing by fitting loglinear models to score data and incorporates the error introduced by the smoothing procedure into step (e) above. KE provides new tools for comparing two or more equating functions and for rationally choosing between them based on newly introduced indices.

Kernel equating is an equipercentile equating procedure in which the score distributions to be equated are converted from discrete distributions to continuous distributions by using a normal (Gaussian) kernel as opposed to using linear interpolation as in the traditional equipercentile equating method. By varying the bandwidth values of the Gaussian kernel (see von Davier et al., 2004), KE can approximate the traditional equipercentile and linear equating methods. When optimal bandwidths are chosen, the KE will approximate the traditional equipercentile equating method. The process of choosing the optimal bandwidth is fully automatic and involves the minimization of a penalty function. When the bandwidths used are 10 times the standard deviation of the scores or larger (i.e., large bandwidths), the continuized distributions will be nearly normal, in which case the KE functions can be regarded as approximately linear. Thus, linear equating can be regarded as special case of equipercentile equating in the framework of KE. The KE framework also introduces the percent relative error (PRE) that aids the diagnosis of the equating function and introduces the standard error of the difference between two equating functions (the SEED). The SEED can help rationalize the linear/nonlinear decision.

As in the IRT true-score equating, several steps are employed in order to equate and report scores: The observed-score equating methods are used to link the observed scores on the two forms. After that a second step is employed, which refers to placing the raw scores onto some reporting scale.

The observed-score equating methods are mostly used in horizontal equating for tests where every test taker has all the items. For tests built using short blocks of items (as in the case of a BIB design) where each test taker has some of the blocks, and in most vertical linking settings, the IRT methods are preferable for linking purposes.

3.5 Vertical Scaling

As in the equating process, the first step in vertical linking is to place the competencies or abilities from several test forms collected from designs of the type described in Tables 4, 5, and 6 onto a common scale and to rank order the examinees on an interim-score scale. Then the scores from each level are linked to a reporting scale, such as a grade-equivalent scale. This procedure is called scaling. The scaling procedure can use the observed scores or can be IRT-based.

The most commonly used linking methods for creating scale scores in vertical scale settings are the Hieronymus, Thurstone, and IRT scaling procedures (see Hendrickson, Kolen, &

Tong, 2004; Tong & Harris, 2004; Yen, 1986; Yen & Burket, 1997). In all these methodologies, an interim scale is chosen (for example, in Table 6, the data from a representative sample on the scaling test *V* is used for a target scale; in a design such as in Table 5, the tests given to P_3 could be the target scale).

Hieronymus scaling. This method (Petersen et al., 1989) was developed for data collection designs with a scaling test, as shown in Table 6; the method makes use of the total number-correct score for dichotomously scored tests or the total number of points for polytomously scored items. The scaling test is constructed to be representative of content from the lowest to the highest level of testing, and it is administered to a representative sample from each testing level or grade. The true-score distributions of competencies at each level are assumed to have the same mean as the observed distributions but a particular variance, as expected following classical test theory; it is assumed that the scaling test is representative of the domain/construct to be measured at each level. To conduct Hieronymus scaling, the median number-correct score on the scaling test *V* from Table 6 for each grade level is assigned a prespecified score scale value (these values are based on various considerations that include the domain to be measured, the measurement point in time in relation to the domain, etc.—see Kolen & Brennan, 2004, p. 382). Hence, the within- and between-level variability and growth are determined on an external scaling test, which is the special set of common items described in the design from Table 6; this design is usually paired with this scaling method.

Thurstone scaling. This process (Thurstone, 1925, 1938) creates first an interim-score-scale as above and then normalizes the distributions of the variables at each grade. That is, it assumes that scores on an underlying scale are normally distributed within each group of interest and makes use of the total number-correct scores for dichotomously scored tests or the total number of points of polytomously scored items to conduct scaling. These normalized distributions are place on the final score scale using the means and variances of these normal distributions at each grade. In other words, Thurstone scaling normalizes the raw scores and linearly equates, and it is usually conducted with equivalent groups.

IRT scaling. Model-based, IRT scaling considers the person-items interactions. In theory, at least, one could conceive of an IRT scaling for all existing IRT models, including multidimensional IRT models or diagnostic models. However, in the practice of vertical scaling

of educational assessments, the models used are unidimensional models such as the Rasch and/or partial credit models (PCM) or the 3PL and/or generalized partial credit (GPC) models.

The methods for the item-parameter linking are those described before under IRT calibration (see also Harris & Hoover, 1987). The only major difference between the use of IRT in horizontal linking and the use in vertical linking is that if concurrent calibration is used for linking the item parameters onto the same scale for all the tests taken by several very different populations of test takers, the estimation method should allow for estimation of the parameters of multiple ability distributions. This estimation can be done separately by level and then linked (via characteristic curves, for example), or it can be done via a concurrent calibration with multiple populations.

4. Measuring Growth in Longitudinal Settings

There are many models available for measuring intra- and interindividual growth. The goal of this section is to review some of these models and to place them in a common background. We will first present the layout that is common to most measurement of change studies. Then we will mention some IRT extensions of the current methods for modelling change, which are the structural equation models (SEMs) and hierarchical or multilevel models (H/MLMs). In this section, we do not refer to survival analysis, and we make only a few references to causal inference analysis.

This section relies on the sections above by assuming that (a) the same measurement instrument is used on all measurement occasions and this instrument continues to measure the same construct or that (b) a common scale has been established.

Although the abundance of available methods is a good thing, measuring growth is still not an easy endeavour. In applied studies using models for measuring growth, one often finds considerable sparseness in the data across time points: There seems to be a general challenge in obtaining samples that are large enough to support an accurate estimation of the (complex) models—as the number of measurement points in time, as the data per time measurement, and as the data per subgroup of interest. Therefore, it appears that the inferential ambitions of the models usually exceed the capacity of the data to support these inferences, and, at the same time, the consequences of these inferences might be serious in the political and social framework where these types of research and analyses are required.

This section focuses on defining the type of growth mentioned above, on reviewing the assumptions required by (most) of the existing methodologies, and on reviewing several approaches for measuring growth. Also, it will briefly discuss the differences between intra-and interindividual measurements of growth. Hence, here we try to be explicit about what are the research questions behind the different types of measurement of growth models and the assumptions that underlie them, and we briefly review several approaches.

Research questions. von Davier and von Davier (in press) identified four research questions that motivate the study of change (regardless of the domain—psychology, education, or other fields):

1. What is the change that each person experiences over time? (individual change)
2. Do the rates at which each individual changes differ by values/outcomes of background variables? (interindividual systematic change)
3. Does a specific treatment have an effect on how an individual changes? (causal inferences)
4. How does a cohort change over time?

Obviously, an answer to the first question on the individual trajectories over time is a prerequisite for answers to the subsequent two questions. In other words, modelling the change that an individual person experiences with time is at the core of the study of change. However, the answer to the fourth research question may or may not rely on the individual trajectories.

Assumptions. When we talk about process assumptions, we refer to those assumptions necessary to answer the questions above. von Davier and von Davier (in press) identified three types of process assumptions: (a) assumptions about the data, (b) assumptions about the instrument/ outcome variable, and (c) assumptions about the model(s).

Assumptions about the data. The data should have appropriate features depending on which research question a study aims to answer. To answer research questions such as the first and second above, ideally data should be available longitudinally on many individuals (when time points and individuals have been sampled representatively) for at least three time points. Ideally, data should be balanced (although the H/MLM type of approaches can relax this requirement). If one wants to make causal inferences (the third research question above), then the

similarity between units/examinees across treatment and control groups should be ensured (Holland, 2005; Raudenbush, 2004).

To answer questions such as the fourth one above, the data does not necessarily need to be longitudinal. A design where random samples are independently drawn with replacements from a cohort at different time points might suffice.

Assumptions about the outcome variable. In SEM/HLM approaches the outcome variable must be a continuous variable at either the interval or the ratio level. In other approaches, however, the outcome variable does not need to be continuous. Models for categorical outcome variables can be handled by extensions of SEM/HLM models or IRT models for change (Cronbach & Furby, 1970; Embretson, 1991; Meiser, 1996; Raudenbush, 2004; Willett & Sayer, 1994; Wilson, 1989).

The outcome variable must be comparable across time points (i.e., each scale point of the measure must retain an identical meaning over time). The outcome variable must remain construct-valid for the entire period of observation. These two assumptions speak to the necessity of establishing a common scale for the instrument(s) as discussed previously in this paper.

Model assumptions. To answer the first research question, one needs a model for the individual change with change/growth parameters that aims to model the individual trajectories over time.

Ideally, this model is based on a substantive theory. In most cases, the model has to be simple, since only a few time points are available: for example, a linear or a quadratic function of time with a stochastic measurement error (and additional technical assumptions on the distribution of error terms). Usually, plots of the observed patterns of change are used as first and explorative steps.

If a study aims to answer the second (or third) type of research question(s) separately or simultaneously with the first question, an additional model is needed. A measurement of change study that tries to explain the systematic differences among the individual trajectories introduces a model for the interindividual change in the change/growth parameters from the first model.

These second order models are based on distributional assumptions about individual growth parameters and contain additional predictors of change. The second order model might try to answer both types of questions related to interindividual differences in change (the second and third research questions). The difference between the models is in how the results are

interpreted given the data at hand and given the additional requirement of similarity of units/subjects for causal inference as it was mentioned above.

A model that aims to answer only a research question of type four (i.e., to measure changes of population distributions over time) may or may not rely at all on a model of individual trajectories.

4.1 Measuring Change Using IRT

The purpose of this section is to briefly describe the approaches used for measuring inter- and intraindividual growth using IRT models. This section relays heavily on the literature survey described in von Davier and von Davier (in press).

Measuring individual growth. In this category, we place IRT models that incorporate measuring change (and the IRT calibration/linking if the test forms differ) in one step. Glück and Spiel (1997) and Meiser, Stern, and Langeheine (1998) provided comprehensive theoretical descriptions of several relevant IRT models as well as detailed comparisons of the results of the models' applications to the same data. These papers show how to apply different IRT models to data with repeated observations on the same items and the same individuals at different occasions. Both papers show how one could model change using the 1PL IRT model (Rasch model) and latent class models: mover-stayer mixed Rasch model (Glück & Spiel), linear logistic test model (LLTM) for measuring change—generalized LLTM (Fischer & Ponocny, 1994, 1995), and mixture distribution Rasch model (von Davier & Rost, 1995; Rost, 1990). In the examples provided in these papers, the samples used to illustrate the methods were small, especially in the Glück and Spiel (1997) paper. In these studies, the outcome variable is discrete. The software that was used in the analyses reported in these papers was LPMC (Fischer & Ponocny, 1994) and WinMira (von Davier, 2001). Other IRT-based approaches that could be placed in the same category are the multidimensional Rasch model (Kelderman, 1996) and the Saltus model (Wilson, 1989).

Hierarchical models for vertical scaling. In this category, we mention two approaches described by Patz et al. (2003) and von Davier and von Davier (2004) that try to address vertical linking and a more complex modelling of growth in a one step procedure. In addition to establishing the scale, the paper by Patz et al. proposed a method for modelling growth across years. The scaling and the growth modelling are both cast in a full Bayesian framework in a hierarchical model. The hierarchical multigroup IRT approach proposed by Patz et al. allows the

explicit estimation of the functional form (assumed to be quadratic) of the grade-to-grade growth patterns. The parameters of this complex model are subsequently estimated using an MCMC approach. The paper also investigates a multidimensional, multigroup IRT model that captures differences in dimensionality and scale definition across grade levels. The hierarchical approach proposed by Patz et al. is "a more general version of concurrent estimation of the unidimensional IRT model" (p. 40) and their motivation was to unify the two most commonly used linking methods for vertical equating, the concurrent calibration method and separate calibration followed by a test characteristic curves linking.

Hierarchical models with IRT measurement models. This category includes models that try to explain interindividual differences. If the measurement instruments differ, then these approaches assume that the scores have been placed onto a common scale (see Raudenbush, 2001, or Willett & Sayer, 1994). The models discussed here represent extensions of the SEM or HLM approaches. Both SEMs and HLMs can borrow from statistical methods that allow for non-normal outcome variables. The link functions used in general linear models (GLMs) and the item response functions used in IRT are obvious candidates for such an adoption of methods.

The Rasch model is the model of choice in many applications of these merged models because of its alternative view as either a logistic regression model or as a loglinear model. These alternate views facilitate the use of Rasch model in existing developments of SEM and HLM that are based on these alternative formulations. Kamata (2001) presented a model that represents an integration of the hierarchical linear model and the Rasch model. Raudenbush and Sampson (1999) presented an interesting application of such a model and showed how to estimate indirect effects of a three-level hierarchical model. The authors introduced a Rasch type measurement model (the model for the latent variables) at the first level. The next two levels are the usual ones for the (hierarchical structure of the) data. More specifically, the levels used in Raudenbush and Sampson, Raudenbush, Johnson, and Sampson (2003), and Johnson and Raudenbush (in press) are the following: The first level is a logistic regression with mixed effects, the random effects are subjects and groups of subjects (blocks), and the fixed effects are items; the second level has fixed effects for time interval of observation, additional true values of the two latent variables, and an error term; and the third level has the grand mean and an error (variance, covariance) term.

The features that make this approach appealing for analyses of data from educational and social sciences include (a) the model incorporates three levels with time being introduced on second level, (b) the Rasch type modelling of items on level one enables the use of binary observables instead of sum scores, and (c) the model can be estimated with penalized ML using Laplace approximation described in Raudenbush, Yang, and Yosef (2000), which is estimable in HLM.

One important aspect of these models is that an IRT type measurement model on level one needs to assume certain invariance properties. Again, one minimal assumption is that a common scale has been established.

Problems may arise if the scale linkage across time points relies only on subsets of items or link items phased out over a limited number of time points to be replaced by other items. In this case, the link can be assumed to be somewhat weaker, and it can be argued with some justification whether what changed was not the subject, but the subject matter (i.e., we may not be able to establish that the same latent variable was measured over time).

This is much less a problem in short-term studies (where the time frame is days or weeks rather than school years) like the one described above, but it is a serious problem in educational assessments and studies in developmental psychology, where some items actually become obsolete as children grow older or go through certain educational levels.

5. Discussion

This paper reviews the existing methodologies for equating and linking tests that measure the same construct over time. The first part of the paper describes horizontal equating, where interchangeability of the scores is desired. Then, vertical scaling is discussed as it is used in educational assessments, where measuring growth in a particular domain and comparability of scores on test forms that measure the same construct but differ in difficulty is desired. We also address the challenges of covering large content domains in educational survey assessments over many cycles and discuss some of the solutions that have been developed using extensions of IRT models for reporting subgroup distributions. In the previous section, we discussed some existing explanatory models for inter- and intraindividual growth. Each of these areas is a large field in itself, and it potentially has a strong impact on the educational policies, on the life of students and parents, or on the life of professionals.

Nowadays, when more and more standardized testing is used nationally and internationally, we are also discovering more challenges in ensuring that the process and the results are fair and accurate. For example, among the challenges and research opportunities for test linking, we easily can mention the definition of growth, the construction of the anchor sets, the choice of the reporting scale, and the characteristics of the samples used for establishing the scale, which ideally should be representative for the population of test takers. In addition, psychometricians worry about the maintenance of the scale: how to introduce new forms, how to monitor the scale over time, and how to adjust to changes in the administration mode.

In conclusion, we have noticed that many researchers and practitioners already work together in addressing these challenges, and we hope that many universities will consider implementing training curricula that prepare the future generation of psychometricians for the challenges in the field of education in the 21st century.

References

ACT, Inc. (2000). *ACT educational planning and assessment system.* Iowa City, IA: Author.

Allen, N.L., Jenkins, F., & Schoeps, T.L. (2004). *The NAEP 1997 arts technical analysis report* (ETS-NAEP 04-T01). Princeton, NJ: ETS.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508—600). Washington, DC: American Council on Education.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–445.

CTB/McGraw-Hill. (2000). TerraNova CAT, The second edition [Computer software]. Retrieved March 21, 2006, from http://www.ctb.com/mktg/terranova/tn_intro.jsp

Cronbach, L.J., & Furby, L. (1970). How we should measure "change"—Or should we? *Psychological Bulletin, 74,* 68–80

von Davier, A.A., Holland, P.W., & Thayer, D.T. (2004). *The kernel method of test equating.* New York: Springer Verlag.

von Davier A.A., & von Davier, M. (in press). *A micro-survey for measuring change.* Princeton, NJ: ETS.

von Davier, A.A., & Wilson, C. (2005). *A didactic approach to the use of IRT true-score equating* (ETS RR-05-26). Princeton, NJ: ETS.

von Davier, M. (2001). WinMira 32 Pro: A program system for analyses with a variety of discrete mixture distribution models [Computer software]. Retrieved February 16, 2006, from the Assessment System Corporation: http://www.assess.com/Software/WINMIRA.htm

von Davier, M., & von Davier A.A. (2004). *A unified approach to IRT scale inking and scale transformations* (ETS RR-04-09). Princeton, NJ: ETS.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371–379). New York: Springer Verlag.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (in press). Marginal estimation of population characteristics: Recent developments and future directions. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics. Vol. 27: Psychometrics.* Amsterdam: Elsevier.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–516.

ETS. (2006a). *System 5™ newsletter for K-12 educators.* Retrieved February 16, 2006, from http://www.ets.org/Media/Newsletters/System5Connection/2005/0512_qa.html

ETS. (2006b). *TOEFL®—Test of English as a Foreign Language™*. Retrieved February 16, 2006, from http://www.ets.org/toefl

Fischer, G.H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika, 59,* 177–192.

Fischer, G.H., & Ponocny, I. (1995). Extended rating scale and partial credit models for assessing change. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 353–370). New York: Springer Verlag.

Glas, C.A.W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64*(3), 273–294.

Glück, J., & Spiel, C. (1997). Item response-modelle für meßwiederholungsdesigns: Anwendung und grenzen verschiedener ansätze [Item response models for repeated measures designs: Application and limitations of different approaches]. *Methods of Psychological Research Online, 2*(1). Retrieved February 16, 2006, from http://www.mpr-online.de//issue2/art1/article.html

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

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Harcourt Assessment, Inc. (2006). Stanford learning first [Computer software]. Retrieved February 16, 2006, from http://harcourtassessment.com/HAIWEB/Cultures/en-us/dotCom/SLF/Stanford+Learning+First.htm

Harris, J.D., & Hoover, H.D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11,* 151-159.

Harris, J.D., Hendrickson, A.B., Tong, Y., Shin, S.-H., & Shyu, C.-Y. (2004, April). *Vertical scales and the measurement of growth.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Hendrickson, A.B., Kolen, M.J., & Tong Y. (2004, April). *Comparison of IRT vertical scaling from scaling test and common items designs.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Holland, P.W., Dorans, N.J., Petersen, N. (in press). Equating test scores. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics. Vol. 27: Psychometrics.* Amsterdam: Elsevier.

Holland, P.W. (2005). Lord's paradox. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2). New York: Wiley.

Holland, P.W., & Thayer, D.T. (1989). *The kernel method of equating score distributions* (ETS RR-89-07). Princeton, NJ: ETS.

Hoover, H.D., Dunbar, S.D., & Frisbie, D.A. (2001). *The Iowa Tests of Basic Skills. Interpretive guide for teachers and counselors. Forms A and B. Levels 9-14.* Itasca, IL: Riverside Publishing.

Hoover, H.D., Dunbar, S.D., & Frisbie, D.A. (2003). *The Iowa tests. Guide to development and research.* Itasca, IL: Riverside Publishing.

Johnson, C., & Raudenbush, S.W. (in press). A repeated measure, multilevel Rasch model with application to self-reported criminal behavior. In C.S. Bergeman & S.M. Boker (Eds.), *Quantitative methodology in aging research.* Mahwah, NJ: Erlbaum.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79–93.

Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. Applied *Psychological Measurement, 20,* 155–168.

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

Livingston, S.A. (2004). *Introduction to test equating (without IRT).* Princeton, NJ: ETS.

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

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

Lord, F.M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equating." *Applied Psychological Measurement, 8,* 452–461.

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 curves solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139–160.

Massachusetts Department of Education. (2001). *Overview of the MCAS 2001 tests.* Retrieved February 16, 2006, from http://www.doe.mass.edu/mcas/2001/overview/complete.pdf

Meiser, T. (1996). Loglinear Rasch models for the analysis of stability and change. *Psychometrika, 61,* 629–645.

Meiser, T., Stern, E., & Langeheine, R. (1998). Latent change in discrete data: Unidimensional, multidimensional, and mixture distribution Rasch models for the analysis of repeated observations. *Methods of Psychological Research Online, 3*(2), 75–93. Retrieved February 16, 2006, from http://www.mpr-online.de//issue5/art6/meiser.pdf

Muraki, E., & Bock, R.D. (1997). PARSACAL 3.0, IRT item analysis and test scoring for rating scale data [Computer software]. Chicago, IL: Scientific Software International.

Patz, R. (2005, June). *Methods and models for vertical scaling.* Paper presented at Linking and aligning scores and scales: A conference in honor of Ledyard R Tucker's approach to theory and practice, Princeton, NJ.

Patz, R., & Junker, B. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 2,* 146–178.

Patz, R., Yao, L., Chia, M., Lewis, D., & Hoskens, M. (2003, April). *Hierarchical and multidimensional models for vertical scaling.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Raudenbush, S.W. (2001). Toward a coherent framework for comparing trajectories of individual change. In L.M. Collins & A.G. Sayer (Eds.), *New methods for the analysis of change* (pp. 35–64). Washington, DC: American Psychological Association.

Raudenbush, S.W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121–129.

Raudenbush, S.W., Johnson, C., & Sampson, R.J. (2003). A multivariate, multilevel Rasch model for self-reported criminal behavior. *Sociological Methodology, 33*(1), 169–211.

Raudenbush, S.W., & Sampson, R. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29*(1), 1–41.

Raudenbush, S.W., Yang, M.L., & Yosef, M. (2000). Maximum likelihood for hierarchical models via high-order, multivariate LaPlace approximation. *Journal of Computational and Graphical Statistics 9*(1), 1–41.

Rock, D.A., Pollack, J.M., & Weiss, M. (2004). *Assessing cognitive achievement growth during the kindergarten and first grade years* (ETS RR-04-22). Princeton, NJ: ETS.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271–282.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W. (1995). BUGS Bayesian inference using Gibbs sampling, Version 0.50 (1995) [Computer software]. MRC Biostatistics Unit, Cambridge.

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

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology*, *16*(7), 433–451.

Thurstone, L.L. (1938). *Primary mental abilities* (Psychometric Monographs No. 1). Chicago: University of Chicago Press.

Tong, Y., & Harris, D.J. (2004, April). *The impact of choice of linking and scales on vertical scaling.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Verhelst, N.D., & Glas, C.A.W. (1995). The one parameter logistic model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and application.* New York: Springer Verlag.

Willett, J.B., & Sayer, A.G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116,* 363–381.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. Psychological Bulletin, 105, 276–289.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). LOGIST users guide [Computer software manual]. Princeton, NJ: ETS.

Yen, W.M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*(2), 93–111.

Yen, W.M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 299–325.

Yen, W.M., & Burket, G.R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement 34*(4), 293–313.

Notes

[1] This claim is appropriate if models such as the Rasch model (1PL) and (to some extent) the one-parameter logistic model (OPLM; Verhelst & Glas, 1995) fit the data, so that conditional maximum likelihood can be used for estimating the item parameters. Strictly speaking, more complex IRT models that require joint estimation or a distributional assumption for the latent ability variable when estimating item parameters do not share the feature of parameter separability.