Pontificia Universidad Católica de Chile
Faculty of Mathematics
Department of Statistics

Linking measurements: A Bayesian nonparametric approach

Inés María Varas Cáceres

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
PhD IN STATISTICS

October 2019

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
DEPARTMENT OF STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Mathematics for acceptance a thesis entitled **"Linking measurements: A Bayesian nonparametric approach"** by **Inés María Varas Cáceres** in partial fulfilment of the requirements for the degree of **PhD in Statistics**.

Dated: October, 2019

Research Supervisor : _____
Jorge González
Pontificia Universidad Católica de Chile

Research Co-Supervisor : _____
Fernando Andrés Quintana
Pontificia Universidad Católica de Chile

Examing Committee : _____
Ernesto San Martín
Pontificia Universidad Católica de Chile

Examing Committee : _____
Alejandro Jara
Pontificia Universidad Católica de Chile

External Supervisor : _____
Matthew Johnson
Educational Testing Service, ETS

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Date: October, 2019

Author	:	Inés María Varas Cáceres
Title	:	Linking measurements:
		A Bayesian nonparametric approach
Department	:	Statistics
Degree	:	PhD in Statistics
Convocation	:	October
Year	:	2019

To Edmundo Alberto, Carmen Lidia
Eduardo Emilio and Matteo Alonso

Acknowledgements

Contents

CONTENTS

List of Tables

List of Figures

Introduction

The main goal of this dissertation is to propose a new method for linking measurements obtained by different instruments. The proposal is developed in agreement with the idea of comparable scores defined in the context of educational measurement, specifically on equating methods (Angoff, 1971; Kolen and Brennan, 2014; von Davier et al., 2004; González and Wiberg, 2017). The proposal extends the methods already applied in psychometrics and tackle some of its drawbacks by considering measurements as ordinal random variables. The latent representation of these variables, together with the Bayesian nonparametric approach we adopt, allow flexibility to define customised relations between specific subgroups of the population of interest. As it will be shown in the application's sections through the dissertation, an important feature of the proposal is that it could be applied in contexts broader than educational measurement such as psychology and health-related areas.

The remainder of this introduction section is organised as follows. First, the concept of linking measurements is defined. General ideas of comparable scores defined for equating methods are explained to introduce the statistical framework in which the proposal is founded. An overview of each chapter is given and some remarks concerning the structure of this dissertation are mentioned at the end of this section.

Linking measurements

A conceptual definition

An example that could summarise the main ideas and concepts related to linking measurements, as understood in this thesis, is to convert temperature measurements defined on Celsius (C°) scale to its exact equivalent measurement in Fahrenheit (F°) scale or vice-versa. Converting measurements of temperature between these scales is a straightforward process using the formulas:

$$\begin{aligned} \varphi_C(F) &= \frac{5}{9}(F-32)\,, \\ \varphi_F(C) &= \frac{9}{5}C+32 = \varphi_C^{-1}(F)\,. \end{aligned} \tag{1}$$

Why and how were these formulas developed? They came from the need of making the Celsius and Fahrenheit scales comparable. Since the freezing point is 0° on the Celsius scale and 32° on the Fahrenheit scale, we subtract 32 when converting from Fahrenheit to Celsius, and add 32 when converting from Celsius to Fahrenheit. Additionally, the boiling points for the Celsius and Fahrenheit scales are 100° and 212°, respectively. Thus, there are 100 degrees between the freezing (0°) and boiling points (100°) of water on the Celsius scale, and 180 degrees between 32° and 212° on the Fahrenheit scale. Writing these two scales as a ratio, we obtain $F^\circ/C^\circ = 9/5$. Flipping the ratio, we have $C^\circ/F^\circ = 5/9$. Then, by using the function (1), a temperature of 57 Fahrenheit degrees means the same as a temperature of $\varphi_C(57) = 15$ Celsius degrees.

Note that, in this example, the *measurement* of interest is the temperature which is quantified by two *measurement instruments*, Celsius and Fahrenheit thermometers. Even though as devices they could be very similar, each one is defined on specific *scales*; Celsius and Fahrenheit degrees scales in the example. The two temperature measurements that are obtained by different instruments are related through the *function* $\varphi(\cdot)$. The same process will be generalised throughout this dissertation. Linking measurements will be understood as "procedures, based on statistical models, in which relations among measurements obtained by different instruments, are defined". In particular, our focus is to

define and estimate a function that is capable to map measurements obtained from different instruments such that measurements represent the same relative position on the scales they are defined.

Motivation

The motivation for the development of procedures for linking measurements comes from the educational measurement context. Any testing program continually produces different editions of a test, in what follows *forms* of a test, which are defined on score scales that should be maintained over time to assure comparability. Although they intend to measure the same construct and are built to have the same test specifications, differences in the difficulty of the forms are unavoidable. Without adjustments, examinees would expect lower scores in the hardest form. An important objective for testing programs is to eliminate the effects of differences in difficulty on the reported scores. Test equating methods have been developed to report test-scores as fair as possible so that test-scores mean the same, regardless of the test form administered.

Equating methods are a specific group of *linking methods* developed in the educational measurement setting. The word "linking" refers to a general class of transformations between scores from different tests. The transformations can be obtained by a range of ways depending, in part, on some features of testing situations being the most important one the construct for which the test, as instrument, was built for. There are three main categories of linking methods: predicting, scaling (scale alignment), and equating (Dorans and Holland, 2000). Although similar statistical procedures are used on each group, equating is a statistical process that is used to adjust differences on difficulty among forms that are built to be similar in difficulty and content, allowing to obtain comparable scores, i.e., allowing scores to be used interchangeably for any purpose.

Statistical framework

In order to generally describe the statistical problem associated to the development of linking measurements, we first introduce some notation specifications. We consider a variable of interest that can be measured by using at least two different measurement instruments. Capital letters represent different instruments, e.g., A, B, C, etc. Each instrument generates measurements defined on equal or different scales, denoted by $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$, etc. Italic letters A, B, C, etc denote the random variables that represent the measurements of each instrument.

We consider that different instruments define subgroups on the population of interest. Just for exposure purposes, let us suppose there are only two measurement instruments A and B. We consider that there is a function relating the sample spaces $\mathscr{A}$ and $\mathscr{B}$,

$$\varphi_B(\cdot) \quad : \quad \mathscr{A} \longrightarrow \mathscr{B} \,,$$

satisfying that for all $a \in \mathscr{A}$, there is a value $b = \varphi(a)$ having on $\mathscr{B}$, the same relative position that a has in $\mathscr{A}$. Thus, the statistical problem is to estimate this function based on a random sample of size n_A and n_B from A and B, respectively.

The function φ_B defines a relation between the scales of the instruments A and B. This formulation has also been considered in equating methods (González and Wiberg, 2017). In the educational measurement context, it is necessary to find equivalent test-scores from different forms of a test such that they could be used interchangeably, i.e., they mean the same on each test form. In the context of this dissertation, A and B would be two different forms of a test defined on scales $\mathscr{A}$ and $\mathscr{B}$, respectively. Observed test scores a and b, from test form A and B, respectively, are considered "comparable scores" if $b \in \mathscr{B}$ satisfies the following equation:

$$F_A(a) \;=\; p \;=\; F_B(b) \qquad p \in (0,1) \,, \tag{2}$$

where F_A and F_B are the cumulative distribution functions (CDF) of the random variables A and B, respectively. From a statistical point of view, this means that a and b have the

same percentile in the probability distribution of A and B, respectively. Consequently, they have the same relative position on the scale where they are defined.

Different approaches have been considered in the literature to estimate the CDFs in (2) leading to parametric, semiparametric and nonparametric estimators of $\varphi(\cdot)$. An overview of these methods is described throughout this dissertation, discussing pros and cons of using them. In order to improve on most of the disavantages traditional equating methods exhibit, we propose a new flexible model based on a latent model for ordinal random variables.

Outline of the dissertation

The organisation of the dissertation is as follows:

- **Chapter 1:** An overview of linking methods is provided. A complete reference of theoretical statistical aspects of the linking process is given, in particular for equating methods. In addition, characteristics of the statistical models that support our approach are also covered.

- **Chapter 2:** Based on ideas discussed in Chapter 1, the statistical model assumed for the measurements of interest as well as each step of the proposed linking method are described. To evaluate the performance of the proposal, a simulation study is carried out. Additionally, our approach is applied to a real data set widely studied in the equating literature.

- **Chapter 3:** We extend the model proposed in Chapter 2 by incorporating information from covariates into the model. Results of a complete simulation study are shown. The proposal is applied to a real data set to obtain comparable measurements of different depression scales applied on the Chilean population.

- **Chapter 4:** Both overall conclusions and open questions about the topic of this dissertation are discussed. Theoretical aspects as well as generalisations of the proposal

are described as a future work.

Final considerations

The dissertation is based on manuscripts that are either submitted/accepted for publication or are still work in progress for future submission. As a consequence, there is some overlap between the chapters. Each chapter can be read as a self-contained chapter.

Two chapters correspond to the following original publications:

Chapter 2: Varas, I. M., González, J., Quintana, F.A.. A Bayesian nonparametric latent approach for score distributions in test equating. (Journal of Educational and Behavioural Statistics. Under review, invited resubmission).

In addition, this chapter is partially based on the original publication:

Varas, I., González, J., Quintana, F. A. (2019). A new equating method through latent variables. In M. Wiberg, S. A. Culpepper, R. Janssen, J. González & D. Molenaar (Eds.), Quantitative psychology. pp 343-353. Cham: Springer.

Chapter 3: Varas, I. M., González, J., Quintana, F. A.. Linking measurements: a Bayesian nonparametric approach. (Work in progess)

Chapter 1

Background Material

"The comparability of measurements made in differing circumstances by different methods and investigators is a fundamental pre-condition for all of science"
Dorans and Holland (2000).

There are several areas where it is of interest to establish either comparable measures or a relation between measurements obtained from different measurement instruments. For instance, in health-related fields, measurements can be obtained by a medical device or by a technician. In cognitive health areas, e.g., cognitive psychology and neuropsychology, it is common to use more than one cognitive screening instrument to indicate the likelihood of genuine cognitive impairment (Cullen et al., 2007; Casaletto and Heaton, 2017). In particular, "no single instrument for cognitive screening is suitable for global use" (Cullen et al., 2007). As a consequence, the development of new screening instruments for assessing cognitive function has increased over the last years (van Steenoven et al., 2014). In the setting of educational testing, different forms of a test are used to evaluate the knowledge of a student. Because of several reasons, the different forms of a test are in continuous change through the years. Advances in technology and increased knowledge of diseases' processes, have allowed improvements in measurement methods. Thus, there are several instruments measuring either the same or similar quantities.

In general, all these measures are fundamental to make important decisions at different levels, e.g., to define diagnostic and prognostic evaluation of patients, to select students for a scholarship, etc. However, measurements can be obtained from different *instruments* (medical device-technician/ several cognitive instruments/ different forms of a test), defined on equal or different *scales*. Because different instruments could lead to different results, in order to make the decisions as accurate and fair as possible, it is relevant to define equivalent measurements among the instruments, i.e., measurements having the same meaning on the scale they are defined.

From a statistical perspective, suppose there are K instruments to measure a characteristic of interest on a specified population. Let M_k be the random variable denoting the characteristic of interest measured using the instrument k. Each random variable has cumulative distribution function (CDF) F_{M_k}, for $k = 1, 2, \ldots, K$ and are defined on sample spaces $\mathscr{M}_k$ for $k = 1, 2, \ldots, K$ which, by definition, correspond to the set of all possible values taken by M_k. In this case, each sample space correspond to the scale defined by the instrument. The nature of the sample spaces $\mathscr{M}_k$ could be either subsets of $\mathbb{Z}$ -for example, the total number of correct answers from different versions of a test- or subsets of $\mathbb{R}$ -the proportion of red cells in 5ml of blood taken from two different instruments. The challenge is to define and estimate how the measurements from different instruments are related to each other. Whatever the relation between measurements is, it should consider qualities and features of the probability distribution function (or equivalently, its cumulative distribution function) assumed for the random variables M's. For instance, a possible relation could be defined in terms of equal quantities of interest such as means or variances. A more general approach could be to determine how different are their distributions. Thus, before defining the relations we are interested in, some common approaches already developed in the literature are described.

1.1 Method comparison studies

Method comparison studies (MCS, Choudhary and Nagaraja, 2017), mostly applied in medical and biomedical research areas, are designed to compare two or more competing instruments of measurement of the same quantity, having a common unit of measurement. Under these methods, none of the instruments in the study produce the true values, i.e., it is assumed that the true values remain unknown and the methods measure them with error. Mixed-effect models are a flexible framework for modelling observations in these analyses because the measurements are taken by each method on every subject on the study and there may or may not be replications. Considering only two measurement instruments, a common model relating the observed Y to the true value b is the classical linear model:

$$Y_{i,j} \quad = \quad \beta_{0,j} + \beta_{1,j} b_i + \epsilon_{i,j} \qquad i = 1, \ldots, n_j \;\; j = 1, 2 \tag{1.1}$$

where β_0 and β_1 are fixed constants specific to the measurement method and ϵ_j is the random error vector of the method j. It is assumed that the true value b has a probability distribution over the population of subjects with mean μ_b and variance σ_b^2. The error ϵ has a distribution with mean zero and variance σ_ϵ^2. Also, independent distributions are assumed for the errors and the true values.

The *precision* of a method is defined as the reciprocal of the error variance $1/\sigma_{\epsilon,j}^2$ and the *sensitivity*, the ability of a method to distinguish small changes in the true value, is defined as $\beta_{1,j}/\sigma_{\epsilon,j}$. Because of identifiability reasons, it is common to consider a reference method such that $\beta_{0,1} = 0$ and $\beta_{1,1} = 1$.

The main goal of of MCS is to determine whether the instruments are similar and have sufficient agreement. Similarities are evaluated in terms of the marginal distributions of the methods. For instance, it is desirable that the methods have similar precision and sensitivities. To evaluate differences in these quantities, both the precision ratio (λ) and the squared sensitivity ratio (γ^2), are commonly used, which are defined as

$$\lambda \quad = \quad \frac{\sigma_{\epsilon,1}^2}{\sigma_{\epsilon,2}^2} \qquad \gamma^2 = \frac{\beta_1^2/\sigma_{\epsilon,2}^2}{1/\sigma_{\epsilon,1}^2} = \beta_1 \lambda^2 \,,$$

If these two quantities are close to 1, then the methods are considered to be similar. Complementary, measurements of agreement consider examinations of features of the joint distributions of the measurements such as the concordance correlation coefficient (CCC, Lin, 1989) which measures how tightly concentrated the bivariate distribution (Y_1, Y_2) is around a straight line. Also, the coverage probability (CP, Lin et al., 2002) which computes the proportion of the population of the random variable $D = Y_1 - Y_2$ contained within the margins $\pm\delta$, for a small positive margin δ, i.e., $CP(\delta) = P(|D| \leq \delta)$. High agreement is considered when large values between $(0, 1)$ are obtained for δ. Finally, if the measurements agree well enough, the preferred one is the instrument that is cheaper, faster, less invasive or the easiest to use. If measurements do not agree enough, it is analysed why and how they differ.

The MCS define relations among measurements from different instruments in terms of how different they are with respect to some moments of the distributions of the measurements (marginal and joint distributions). Characteristics of these methods are restrictive to be applied in the setting of linking measurement as it has been defined. In contexts where it is a need to link measurements, as those described at the beginning of this chapter, the sample units (usually people) are not necessarily measured by all the available instruments. In addition, it is possible that the instruments are not defined on the same scale, for instance, two versions of a test. In addition, we believe that not only moments of the distributions should be taken into account to develop relations among measurements but also the whole distribution of the random variables defined by the instruments must be considered. Finally, the objective of both MCS studies and linking methods is the main issue that differenciate them. The former chooses between methods while the latter is devoted to find equivalent measurements among the methods.

1.2 Equating methods

In test theory, it is assumed that instruments (test forms) measure a specific unobserved construct. “Theoretical constructs are often related to the behavioural domain through ob-

servable variables by considering the latter as measures or indicants of the former" (Lord and Novick, 1968). Through a number of items, tests are built to measure a construct of interest. After applying the test to randomly sampled examinees from a population, data could be the pattern of answers on each item, or aggregated data (for instance, adding the number of correct answers). In the former case, all the analyses consider item response theory (IRT) based methods (Lord, 1980) by modelling the probability that a person answer an item correctly in terms of person's ability and difficulty of the items. The approach we consider in this dissertation is based on cases where items' information is aggregated across the test takers, namely observed scores. Then, the test-score distributions of the different forms are the parameters of interest.

Observed score data are considered realisations of a random variable that represents the score of an examinee belonging to a certain group in a population of interest. Score linking methods are used to describe the transformation (the *link*) from a score on one test to a score on other test. Holland and Dorans (2006) classifies the different types of links into three categories: *predicting*, *scale aligning* and *equating*. Predicting is the oldest form of score linking. The goal is to minimise errors of prediction of a score on a test in terms of other variables which could possibly include information from other tests. Discussions of these linking methods can be found in Kelley (1927). In addition, Holland and Hoskens (2003) proposed to model the score in a test (the dependent variable) based on information of other predictor variables such as other tests. The goal of scale aligning is to transform the scores from two different tests onto a common scale. Some subcategories of these methods are *calibration* (Holland and Dorans, 2006) and *concordance* (Pommerich and Dorans, 2004). A goal that distinguishes Equating methods from other forms of linking is the purpose to develop a link between test-scores such that the scores from each test form can be used as if they had come from the same test. As a consequence, both tests involved in equating methods and the method used for linking the scores must satisfy strong requirements. These requirements were discussed in Angoff (1971); Lord (1980); Petersen et al. (1989) and later, some discussions about them can be found in Dorans and Holland (2000); Kolen and Brennan (2014); von Davier et al. (2004). A brief summary of the re-

quirements is described now: (i) *equal construct*: tests should measure the same construct; (ii) *equal reliability*: tests should have the same level of reliability; (iii) *symmetry requirement*: if $\varphi(\cdot)$ is a link function relating test-scores from scale $\mathscr{X}$ to test-scores defined in $\mathscr{Y}$, then $\varphi(\cdot)^{-1}$ should link test-scores from $\mathscr{Y}$ to test-scores defined on $\mathscr{X}$; (iv) *equity*: it should be a matter of indifference to an examinee which test form he/she takes; (v) *group invariance*: the link function should be the same regardless of the choice of population or subpopulations from which it is derived. Discussion about these conditions can be found in van der Linden (2013).

A characteristic part of testing programs is that tests are used in one or more administrations. As a consequence, because of several reasons, there is no single version of a test but alternate forms. However, these alternate forms of a test are built to be *parallel* tests (Lord, 1964), i.e., the forms should have the same test specifications, such as similar structure, item types and formats. Because the process of test construction is not perfect, the difficulty of the form tests will not be the same. When test-scores are used to make important decisions, it is necessary to compensate for the form-to-form variation in test difficulty. Equating methods have been defined as statistical models and methods used to make test-scores comparable among two or more forms of a test which intent to measure the same attribute in order to eliminate differences in difficulty of the tests. (Holland and Rubin, 1982; von Davier et al., 2004; Dorans et al., 2007; von Davier, 2011; Kolen and Brennan, 2014). Comparable test-scores means that they can be used interchangeably, i.e., equated scores from different forms could be treated as if they came from the same test (Kolen and Brennan, 2014; González and Wiberg, 2017).

Test-score differences are not exclusively due to differences in the difficulty of the tests forms. An additional challenge is to avoid the confounding of differences in form difficulty with the differences in the abilities of the group of examinees. These differences are disentangled considering specific data collection designs (Kolen and Brennan, 2014). In fact, considering only two forms of a test, A and B, there are several ways to collect score data: (a) *single group design* (SG): a unique sample group of examinees from the same population is used and all examinees take both test forms in the same order; (b) *equivalent*

group design (EG): two independent samples from the same population of interest are considered and each group take only one test form; (c) *counterbalanced design* (CB): two independent samples from the sampling population are used and both samples take both tests but in different order. Note that all the sample designs described so far consider samples from the same population. However, there are situations where it is possible to take samples from different populations. In the sampling design (d) *non equivalent group with anchor test design* (NEAT): two samples are taken from two different populations. Each group take only one test form and a common anchor test form is administered to both samples. The anchor test is a shorten version of the test and is used to measure and control for differences in ability of the examinees. When there are two samples from different populations but there is no available anchor test, the (e) *non equivalent group with covariates design* (NEC) uses relevant covariates to account for differences in the groups of examinees (Wiberg and Bränberg, 2015).

From a statistical perspective, the goal of equating methods could be restated as the aim to obtain the equivalent test-scores of $x \in \mathscr{M}_1$, from one test form M_1, into $\mathscr{M}_2$, the scale defined for the test form M_2. The *equating transformation* (González and Wiberg, 2017) is defined as a function between the sample spaces $\mathscr{M}_1$ and $\mathscr{M}_2$. Assuming distribution functions F_1 and F_2 for the random variables M_1 and M_2, the development of an expression for the equating function is related to the comparison of any two samples or distribution functions (e.g., Wilk and Gnanadesikan, 1968).

1.2.1 Equipercentile function

Considering two different test forms, M_1 and M_2, Braun and Holland (1982) stated that $\varphi(\cdot)$ equates M_1 and M_2 if, for $x \in \mathscr{M}_1$ and $y \in \mathscr{M}_2$

$$F_{M_2}(y) = F_{\varphi(x)}(y) \, ,$$

i.e., the equated score $\varphi(x)$ and y are comparable scores as in (2). Using this definition, an explicit form of the function $\varphi(\cdot)$ could be obtained. In fact, if F_{M_2} and F_{M_1} are continuous

cumulative distribution functions, then:

$$y = \varphi(x) = F_{M_2}^{-1}\left(F_{M_1}(x)\right) . \tag{1.2}$$

This function is known as the *equipercentile function* (Angoff, 1971). The main idea of this function is to obtain equivalent measures based on the percentiles of the distributions of the test forms. A graphical representation of the equipercentile function is shown in Figure 1.1. Note from the figure that both x and y represent the same percentile p of the distributions F_{M_1} and F_{M_2}, even when they could be defined on different scales. A relevant conclusion obtained from this quotation is that cumulative distribution functions are informative with respect to the relative position of the values the random variables take.

1.2.2 Estimation methods

The development of (1.2) was based on the assumption that the test-scores CDFs, F_{M_1} and F_{M_2}, are continuous functions. However, although various types of scoring techniques could be used, test scores are mostly discrete, generally the sum scores (e.g., the total number of correct answers), so that $\mathscr{M}_1$ and $\mathscr{M}_2$ are subsets of the integer numbers. Consequently, any possible distribution function assumed for the test-scores will lead on a step function. In general, when random variables are defined on a countable set, i.e, a discrete support, there would be a theoretical problem with the definition (1.2). This is because the CDF of these random variables is a non-decreasing step function with steps at each possible value of the support. Even though the inverse CDF of the discrete random variable M_2 is well defined as the quantile function $Q_{M_2}(p)$,

$$Q_{M_2}(p) = F_{M_2}^{-1}(p) = \inf\{y \in \mathscr{M}_2 : F_{M_2}(y) \geq p\} ,$$

it is almost impossible to find a value $y = F_{M_2}^{-1}(p)$ in $\mathscr{M}_2$ such that exactly $p = F_{M_1}(x)$ for any $x \in \mathscr{M}_1$.

The discreteness problem of the CDF previously described is not only theoretical but also practical. A natural estimator of the equipercentile function is based on the empirical

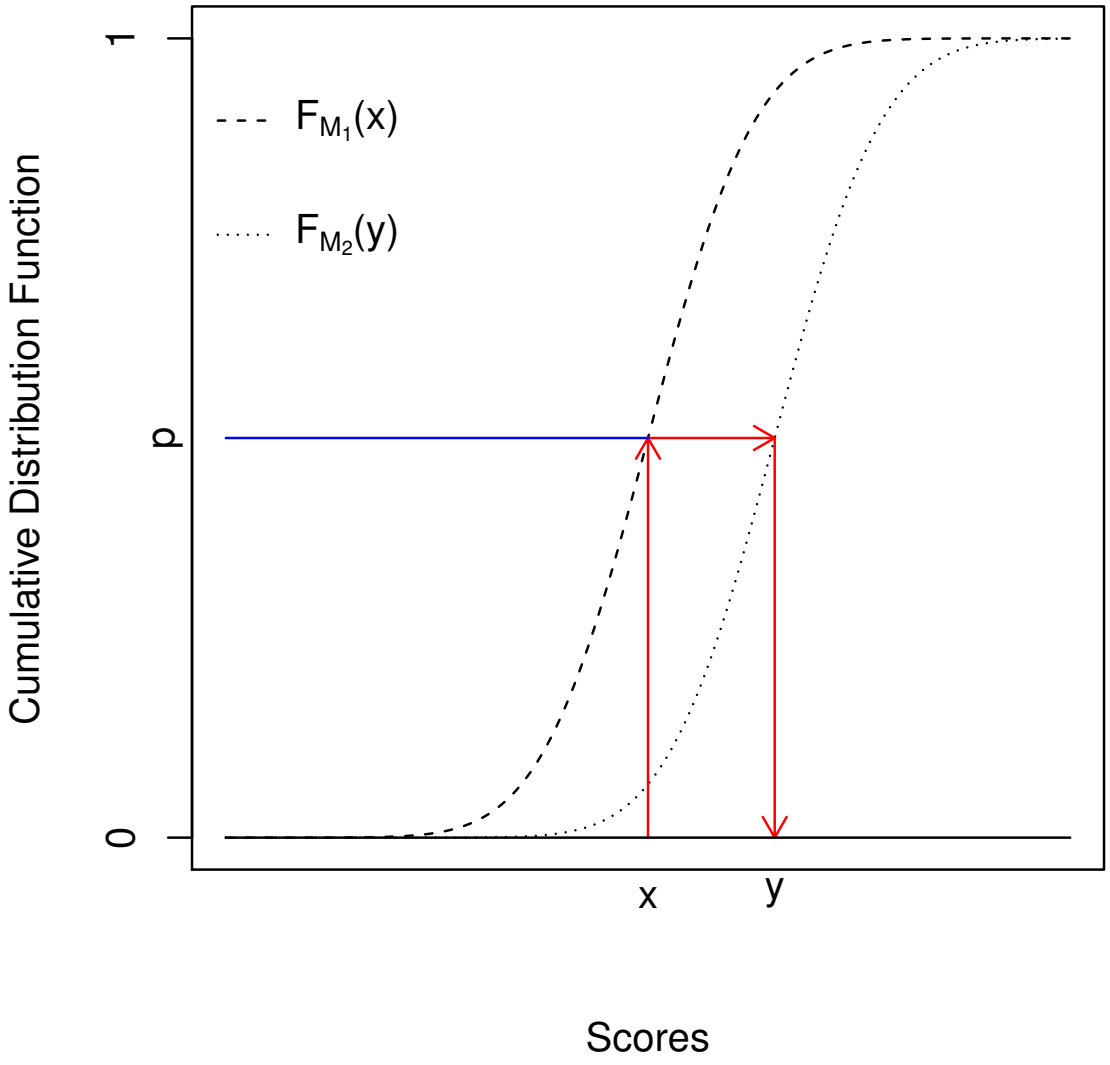


Figure (1.1) Graphical representation of the equipercentile equating function.

cumulative distribution function (ECDF) of M_1 and M_2. By definition, the ECDF of a random variable M_1 based on a sample $M_{1,1}, \ldots, M_{1,n}$ is defined as:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{l=1}^{n} \mathbb{1}(M_{1,l} \leq x) \, ,$$

where $\mathbb{1}(A)$ represents the indicator function on the set A. However, this function is also a non-decreasing step function with steps at the sample observations. This feature of the ECDF is common for all random variables, i.e., it holds when the the variable is either continuous or discrete.

In educational and psychological measurement, it is a tradition to view discrete scores as being continuous by using percentiles and percentile ranks (Holland and Thayer, 1989). The elementary definition of percentile ranks is to consider that examinees with a discrete score x are uniformly distributed in the interval $(x - 0.5; x + 0.5)$ such that the percentile rank of x, $P_R(x)$, is given by:

$$P_R(x) = 100 * \left[F(x-1) + \frac{f(x)}{2} \right]$$

where $F(x-1)$ represents the proportion of test takers scoring at most $x - 1$ and $f(x)$ the proportion of examinees scoring x (Kolen and Brennan, 2014). This definition of percentile rank can be formulated as a kernel smoothing process. A convolution between the discrete distribution function of the test-scores and a continuous uniform random variable $U \sim \mathcal{U}(-0.5, 0.5)$ lead to a continuous random variable. The cumulative distribution function of this new variable is the percentile rank function. Generally, common practice estimates of $\varphi(\cdot)$ are based on continuous approximations of the measures' distributions F_{M_1} and F_{M_2}. In the context of equating methods, this procedure is called the *continuization step*. Linear interpolation (Angoff, 1971; Braun and Holland, 1982), kernel smoothing techniques (Holland and Thayer, 1989; von Davier et al., 2004) and continuized log-linear methods are typically used as continuization methods.

As a result of applying a method to continuize discrete distributions, either parametric, semiparametric and nonparametric estimations of $\varphi(\cdot)$ can be obtained (González and von

Davier, 2013). A parametric estimator of this function is obtained under certain assumptions (Braun and Holland, 1982; von Davier et al., 2004). By considering a location-scale family of distributions for M_1 and M_2, indexed by location-scale parameters (μ_1, σ_1) and (μ_2, σ_2), respectively, if H is a distribution function such that

$$F_{M_1}(x) = H\left(\frac{x-\mu_1}{\sigma_1}\right) \quad \text{and} \quad F_{M_2}(y) = H\left(\frac{y-\mu_2}{\sigma_2}\right) ,$$

then, the equipercentile function takes the following form:

$$\varphi_{M_2}(x;\theta) \;=\; \mu_2 + \frac{\sigma_2}{\sigma_1}(x-\mu_1) .$$

After estimating the vector of parameters $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2)$, the estimation of the parametric version of $\varphi(\cdot)$ is obtained. Note that the previous result is valid only if the distribution function of both measures are from the same location-scale family of distributions. A semiparametric estimator of the equipercentile function was proposed by von Davier et al. (2004), the Kernel equating function. In this approach, the probabilities $p_x = P(M_1 = x)$ and $p_y = P(M_2 = y)$ are considered to be the parameters of a multinomial distribution and the distribution functions F_{M_1} and F_{M_2} are estimated using nonparametric kernel smoothing techniques. This procedure result on the following equating transformation:

$$\varphi_{M_2}(x;\theta) \;=\; F^{-1}_{M_2,h_2}\left(F_{M_1,h_1}(x,\boldsymbol{p}^x);\boldsymbol{p}^y\right) ,$$

where $\boldsymbol{p}^x = (p_x)_{x\in\mathscr{M}_1}$ and $\boldsymbol{p}^y = (p_y)_{y\in\mathscr{M}_2}$ are the vector of probabilities and h_1, h_2 are the bandwidth parameters controlling the degree of smoothness for the kernel estimates. Extensions of kernel equating can be found in van der Linden (2011); Wiberg et al. (2014) and the references therein.

1.3 Bayesian nonparametric models

Bayesian nonparametric (BNP) models are probability models defined on infinite dimensional probability spaces, including priors on random probability functions, random mean

functions and more (Mitra and Müller, 2015). In the same way a prior distribution is the key element in Bayesian parametric models, the key element in BNP is the random probability measure (RPM) which is a prior distribution over a collection of distributions. The Dirichlet process (DP) prior (Ferguson, 1973) is the most commonly used RPM.

A random distribution function G is said to be a DP with parameters M and G_0, denoted by $DP(M, G_0(\boldsymbol{\eta}))$, if it can be written as

$$G(\cdot) \quad = \quad \sum_{\ell=1}^{\infty} p_\ell \delta_{\theta_\ell}(\cdot) \; , \tag{1.3}$$

where δ_{θ_ℓ} denotes a point mass function at the atom θ_ℓ, $\theta_\ell \overset{iid}{\sim} G_0(\boldsymbol{\eta})$, $\boldsymbol{\eta}$ is a vector of hyperparameters that defines the base measure G_0, and the weights p_ℓ are obtained by the following recursive expression:

$$p_1 = v_1 \; , \qquad p_\ell = v_\ell \prod_{j<\ell} (1 - v_j) \qquad \ell > 1 \; , \tag{1.4}$$

where $v_\ell \overset{iid}{\sim} Beta(1, M)$. This last decomposition is called the *stick-breaking* representation of a DP prior (Sethuraman, 1994). The DP prior is characterised by the base measure G_0 that generates the locations of the atoms θ_ℓ and the total mass parameter M that determines the distribution of the fractions v_h. The resulting random probability function (1.3) is almost surely discrete (Blackwell and MacQueen, 1973). Properties of the DP prior and alternative constructions can be found in Ghosal (2010), Lijoi and Prünster (2010) and Barrientos et al. (2017).

In many applications where continuous random variables are involved, the discrete nature of the DP random measure makes them inadequate to be considered in the modelling process. In order to avoid this problem, the DP mixture (DPM) model (Ferguson, 1983; Lo, 1984) considers a convolution with a continuous kernel. Let $Z \sim F(\cdot)$, then a DPM model is defined as

$$F(z) \quad = \quad \int h(z \mid \theta) G(d\theta) \qquad G \sim DP(M, G_0(\boldsymbol{\eta})) \; ,$$

where for every $\theta \in \Theta$, $h(z \mid \theta)$ is a continuous density function, G is a DP defined on $\Theta \subset \mathbb{R}^p$ and $\boldsymbol{\eta}$ is a vector of hyperparameters that defines the base measure G_0. A random

distribution from this model is denoted by $F \sim DPM(M, G_0, h)$. Breaking the mixture, this model can be written in a hierarchical representation as follows,

$$\begin{aligned} z_i \mid \theta_i &\sim h(z_i \mid \theta_i) \\ \theta_i &\sim G \end{aligned}$$

where $G \sim DP(M, G_0(\boldsymbol{\eta}))$. The DPM model is one of the most common BNP prior used for random distributions. Inácio de Carvalho et al. (2015) use it as a prior for the distribution of test outcomes to develop inference on ROC curves, Daniels and Linero (2015) consider it for longitudinal outcomes under different missing patterns, among others applications.

All these models can also be considered when the interest is to model covariate dependent random probability measures $\mathscr{G} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathscr{X}\}$, where $\mathscr{X}$ denotes the space of the covariates. Such is the case, for instance, in a regression problem. Considering information of a response variable y and a set of covariates $\mathbf{x}$ from n unit samples, the regression problem could be defined as

$$y_i \mid \mathbf{x}_i = x_i, \mathscr{G} \sim G_{\mathbf{x}} \qquad i = 1, \ldots, n\,.$$

If $\mathscr{G}$ is not restricted to be indexed by a finite dimensional parameter vector, from a Bayesian perspective, a prior model needs to be defined for $\mathscr{G}$. The most popular models in the literature for this purpose is the dependent Dirichlet Process (DDP) (MacEachern, 1999, 2000). The point masses in the representation (1.3) are still independent across ℓ however both the weights $p_\ell(\mathbf{x})$ and the atoms $\theta_\ell(\mathbf{x})$ are modified such that :

$$G_{\mathbf{x}}(\cdot) = \sum_{\ell=1}^{\infty} p_\ell(\mathbf{x}) \delta_{\theta(\mathbf{x})_\ell}(\cdot)\,, \tag{1.5}$$

where $p_\ell(\mathbf{x}) = v_\ell(\mathbf{x}) \prod_{j<\ell} (1 - v_j(\mathbf{x}))$ and $\theta(\mathbf{x})$ and $v_\ell(\mathbf{x})$ are stochastic processes. Simplified versions of (1.5) are obtained by considering either covariate-dependent atoms or weights, i.e.,

$$G_{\mathbf{x}}(\cdot) = \sum_{\ell=1}^{\infty} p_\ell \delta_{\theta(\mathbf{x})_\ell}(\cdot) \quad \text{or} \quad G_{\mathbf{x}}(\cdot) = \sum_{\ell=1}^{\infty} p_\ell(\mathbf{x}) \delta_{\theta_\ell}(\cdot)\,,$$

respectively. For simplicity, in this work we consider dependency only in the atoms of the DP model.

In the same way that the DP model is combined with a continuous kernel to obtain a continuous density functions, the DDP is also considered to that extent leading to covariate dependent density distributions of the form:

$$F_x(y) \;\; = \;\; \int h(y \mid \theta) G_x(d\theta) \; ,$$

with a DDP prior on $\{G_x,\ x \in \mathscr{X}\}$. This idea has been used by De Iorio et al. (2004) to define an ANOVA-DDP type model. Similar approaches have been used in spatial modelling (Gelfand et al., 2005), survival analysis (De Iorio et al., 2009), functional data (Dunson and Herring, 2006) and classification (De la Cruz et al., 2007). Dunson et al. (2007). Dependence in the weights of the DP representation have been considered in Duan et al. (2007); Rodríguez and Dunson (2011). Müller et al. (2004) incorporate dependency by means of weighted mixtures of independent random measures. Hjort et al. (2010) and Barrientos et al. (2012) summarise more details about Bayesian nonparametric statistics, DDP models and its variations as well as their applications.

In order to estimate these models, its hierarchical representation allows straightforward posterior inference with Markov Chain Monte Carlo (MCMC) simulation. There are two strategies considered for computations of standard DP models. One approach is to employ a truncation of the stick-breaking representation (Ishwaran and James, 2001). Another approach is to use a marginal Gibbs sampling where the mixing distributions are integrated out from the model (MacEachern and Müller, 1998; Neal, 2000).

1.4 Our proposal

The aim of this dissertation is to propose a new method for linking measurements obtained from different instruments $\mathrm{M}_1, \ldots, \mathrm{M}_K$ such that measurements obtained from the linking procedure represent the same in the scales they are defined. Throughout this chapter, a general description of statistical procedures related to the concept of linking measure-

ments have been described. In what follows, an overview of the proposal developed in this dissertation is explained. Even though it is defined in agreement with the philosophical ideas of equating methods, we explain how our approach extends those methods to a general context for linking measurements.

Considering there are K measurement instruments to measure a characteristic of interest, we assume there is a function that maps measurements obtained from the different instruments. In other words, we suppose there is a function $\varphi_{M_j}(\cdot)$ mapping measurements from one scale into another such that, for $y \in \mathscr{M}_j$ and $x \in \mathscr{M}_i$, $y = \varphi_{M_j}(x)$ has the same relative position in the scale $\mathscr{M}_j$ that x has in the scale $\mathscr{M}_i$, for some $i \neq j$ and $i, j = 1, \ldots, K$. Without loss of generality, hereafter we consider $i = 1$ and $j = 2$ to simplify notation. Mathematically, the function $\varphi_{M_2}(\cdot)$ defines a relation between the sample spaces of the random variables M_1 and M_2, i.e.,

$$\begin{aligned} \varphi_{M_2}(\cdot) \; : \; \mathscr{M}_1 &\longrightarrow \mathscr{M}_2 \\ x &\longrightarrow \varphi_{M_2}(x) \, . \end{aligned} \tag{1.6}$$

Note that, because it is not common to have available information about the relation between measurements from different instruments, i.e., information about the structural form of $\varphi_{M_2}(\cdot)$, the statistical problem is nonparametric by default.

In order to characterise the function $\varphi_{M_2}(\cdot)$, it is important to establish, mathematically, the idea of having measurements "meaning the same" on its scales. Because measurements could not necessarily be defined on the same scale, our proposal is based on the idea of finding which measurements represent the same relative position in the scale the measurement instruments are defined. Under this perspective, note that the main feature of the equipercentile function in (1.2) is that test-scores x and $y = \varphi(x)$ "represent the same percentile on the scales they are defined", i.e., by using the equipercentile function, measurements represent the same characteristic on the scale they are defined. As a consequence, a natural way to define the function (1.6) is as the equipercentile function (1.2).

Given that a formal definition for the function $\varphi_{M_2}(\cdot)$ has been established, the next step in the process is to estimate it. Different methods applied in psychometrics have

been already described. A critical aspect when developing procedures to estimate the equipercentile function is that all the traditional methods rely on a continuized version of the CDFs considered for the test-score distributions (see Section 1.2.2). From our point of view, the continuization step is not completely suitable for linking measurements specially when they are defined on discrete scales. i.e., when the sample spaces $\mathscr{M}_1$ and $\mathscr{M}_2$ are subsets of the integer numbers. In fact, if a continuization method is applied, then the scale of the equated measurements is not longer discrete. This result in equated measurements not properly defined on its scales and possible out of the limits of the range defined by its scale. A common approach for practitioners, for example when applying these methods in health-related areas, is to use a rounded version of the equated measurements. However, because instruments are used to classify subjects on range of measurements defined on discrete values, the rounded approach can lead to a erroneous classification process.

Our proposal define a general framework for linking measurements by tackling some drawbacks of equating methods. We consider that measurements defined on discrete scales are ordinal random variables. This assumption is founded in the fact that, with discrete measurements, sample spaces define an order relation between measurements. Thus, taking advantages of the latent representation of ordinal variables, an estimation of the linking function (1.6) in the latent setting is obtained after estimating the latent CDF's of each set of measurements. This estimation is used to obtain a set of continuous linked measurements for each discrete measurement. A procedure to obtain a discrete estimated linked measurement from this set is proposed. Based on ideas of Kottas et al. (2005), we assume a Bayesian nonparametric model for the latent variables based on mixtures induced by a Dirichlet process. Additionally, we extend the proposal to a covariate-dependent model for the latent variables based on mixture models induced by a dependent Dirichlet process which allows smoothly changes of the latent distributions of the measurements as well as the linking functions. An interesting point of the proposal is that it could be applied not only for linking measurements defined on discrete scales but also for measurements defined on continuous scales.

The proposal developed in this dissertation is unique in at least two perspectives. First,

the discreteness problem of the CDF's and ECDF's for discrete random variables is avoided because the linking function is estimated in the continuous latent setting. Second, measurements obtained from the linking procedure we propose are properly defined on its scales. The Bayesian nonparametric model assumed for the latent variables allows flexibility to estimate different relations between measurements. Additionally, considering covariates into the model result in customised relations between specific measurements of interest. All these advantages are discussed on each chapter and also practically illustrated in the applications' sections.

Chapter 2

A latent approach for test equating

"All models are wrong, but some are useful"
Box (1976).

2.1 Introduction

The comparability of test scores is an important issue in the field of educational measurement. Test scores are used to make relevant decisions in various settings, so it is crucial to report scores in a fair and precise way. Mainly due to security reasons, it is common for measurement programs to produce different forms of a test that are intended to measure the same attribute. Equating methods have been developed to achieve the goal of having comparable scores from different test forms. The main idea behind equating is to allow for the ability of *treating equated scores as if they came from the same test*.

Let $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ be the scores obtained on test forms X and Y by n_X and n_Y randomly sampled examinees, respectively. The scores random variables X and Y are defined on sample spaces $\mathscr{X}$ and $\mathscr{Y}$, respectively, which can be seen as the score scales in this context. The statistical problem in test equating is to establish the relationship between scores from two different scales. This goal is achieved using what is called an

equating transformation, φ, which is a mapping between these two sample spaces i.e., $\varphi : \mathscr{X} \rightarrow \mathscr{Y}$ (González and Wiberg, 2017). Thus, an equated score on the scale $\mathscr{Y}$ is the result of applying an equating transformation to a score $x \in \mathscr{X}$. The most popular equating transformation was defined by Braun and Holland (1982) as

$$\varphi_Y(x) \quad = \quad F_Y^{-1}\left(F_X(x)\right) \ , \tag{2.1}$$

where F_Y and F_X are the cumulative distribution functions (cdf) of Y and X respectively. Although various types of scoring techniques could be used, test scores are mostly considered to be sum scores (i.e., the obtained total number of correct answers), so that $\mathscr{X}$ and $\mathscr{Y}$ are subsets of the integer numbers. This fact causes a problem in (2.1) because it is almost impossible to find a value $y = F_Y^{-1}(p)$ in the scale of test Y such that $p = F_X(x)$ for any x score in the scale of test X. Different methods have been proposed in the equating literature to tackle this problem, all of them based on the *continuization* of the originally discrete score distributions F_Y and F_X. Continuization methods include the use of linear interpolation (Angoff, 1971; Braun and Holland, 1982) and kernel smoothing techniques (von Davier et al., 2004). A common feature of all equating methods based on the continuization of F_X and F_Y is that equated scores are not integer numbers anymore and thus are not defined on the original score scale.

Our proposal solves the equating problem while preserving the discrete nature of the score data. To this effect, building on the Bayesian nonparametric model for multivariate ordinal data developed by Kottas et al. (2005), we propose a continuous latent variable formulation of score distributions. A key feature of the proposed method is that scores are equated at a latent and continuous scale, thus avoiding the need to resort to approximations. Specifically, we consider an equipercentile-like equating method that has as final outcome, a discrete equated score for each possible value on the scale score. To the best of our knowledge, such an approach has not been used in previously studied equating methods. We discuss properties of the model and compare its performance with some of the traditional approaches.

The rest of this chapter is organized as follows. In section 2.2 we give a background of both the latent representation for ordinal variables and Bayesian nonparametric models.

In the same section we introduce the proposed Latent equating method (LE). In section 2.3 we illustrate the performance of the proposed method by a simulation study and an application to a real data set. Final conclusions and a discussion of some ideas for further work are presented in section 2.4.

2.2 Latent modeling approach

The equating method proposed in this chapter is based on the latent representation of ordinal random variables. The latent modeling approach for ordinal random variables as well as some relevant aspects of Bayesian nonparametric models are discussed in this section. The proposed model for score distributions and the latent equating method are shown at the end of this section.

2.2.1 Ordinal random variables

Ordinal categorical variables arise frequently in different contexts, e.g., studies on the quality of a service (with categories fair, good, very good), extent of agreement (strongly disagree, disagree, neutral, agree, strongly agree) and the level of education (high school, undergraduate, graduate), ordered item response data,among others. A common approach for the analysis of ordinal data is to assume that measurements are the observable indicator of some underlying continuous latent variable (see McCullagh, 1980; McCullagh and Nelder, 1989; Albert and Chib, 1993). The ordinal and the latent variables are related through a set of thresholds values that partition the support of the latent variable into disjoint intervals, each one corresponding to one of the observed levels of the ordinal variable. Let $Y_i, i = 1, \ldots, n$ be independent and identically distributed ordinal random variables, where Y_i takes one of the $C + 1$ ordered category values $\omega_0, \ldots, \omega_C$. Let $Z_i, i = 1, \ldots, n$ be a random sample from a continuous latent variable Z with cdf F_Z. The latent modeling approach establishes the following relation between these variables:

$$Y_i = \omega_k \quad \Leftrightarrow \quad Z_i \in (\gamma_k, \gamma_{k+1}], \tag{2.2}$$

where $\gamma_0, \ldots, \gamma_{C+1}$ are the thresholds such that $-\infty = \gamma_0 < \gamma_1 < \ldots < \gamma_C < \gamma_{C+1} = +\infty$. The probability distribution of Y_i is thus specified given the probability distribution of Z_i. In fact,

$$\mathbb{P}(Y_i = \omega_k) \quad = \quad \mathbb{P}(\gamma_k < Z_i \leq \gamma_{k+1}) = F_Z(\gamma_{k+1}) - F_Z(\gamma_k) \qquad k = 0, \ldots, C.$$

Parametric models are a common choice for F_Z. One example is the normal distribution, which results in the probit model, and another one is the logistic distribution, leading to the logit model (see McCullagh, 1980; Haber, 1985; Winship and Mare, 1984). In these models, the thresholds are unknown parameters. From a frequentist viewpoint, there is no closed form for the maximum likelihood estimators for both probit and logit models. Fisher scoring and modified versions of weighted least squares algorithms are typically used to estimate these models. As the number of categories increases, the estimation of these models becomes difficult (McCullagh and Nelder, 1989) and so these models are useful for random variables with few categories. In addition, the assumption of normality on F_Z can be very restrictive in cases where there is a large proportion of observations falling in the extreme levels of the ordinal scale and few observations falling in the middle of the scale. More flexible models have been proposed in the literature. For instance, a parametric Bayesian approach was proposed in Zeger and Karim (1991) where, using a Gibbs sampler algorithm, the relation between Bayesian regression and random-effect models was studied. Albert and Chib (1993) used a data augmentation approach to develop a MCMC method based on mixtures models for the distribution of Z. Even though these alternative models are more flexible, nonstandard inferential techniques (Johnson and Albert, 1999) and reparametrizations (Chen and Dey, 2000) are needed in the estimation process.

2.2.2 Bayesian nonparametric models

Bayesian nonparametric (BNP) models are probability models defined on infinite dimensional probability spaces (Mitra and Müller, 2015). In the same way a prior distribution is the key element in Bayesian parametric models, the key element in BNP is the random

probability measure (RPM) which is a prior distribution over a collection of distributions. The Dirichlet process (DP) prior (Ferguson, 1973) is the most commonly used RPM because of its computational simplicity. It is said that G is a DP with parameters M and G_0, denoted by $G \sim DP(M, G_0)$, if for every partition of the sample space $A_1, \ldots, A_k$, the vector of random probabilities $(G(A_1), \ldots, G(A_k))$ follows a Dirichlet distribution $Dir(MG_0(A_1), \ldots, MG_0(A_k))$. From this fact, it follows that for any measurable subset A of the sample space

$$G(A) \quad \sim \quad Beta(MG_0(A), M(1 - G_0(A))) \, ,$$

where $E[G(A)] = G_0(A)$ and $Var(G(A)) = \frac{G_0(A)(1-G_0(A))}{M+1}$. The parameter M is known as the total mass parameter which controls the uncertainty of $G(A)$, i.e., if it is small, the variance of $G(A)$ increases. The density G_0 is known as the base measure which specifies the mean of $G(A)$.

A random distribution function G from a $DP(M, G_0)$ can be alternatively written as

$$G(\cdot) \quad = \quad \sum_{\ell=1}^{\infty} p_\ell \delta_{\theta_\ell}(\cdot) \, , \tag{2.3}$$

where δ_{θ_ℓ} denotes a point mass function at θ_ℓ, $\theta_\ell \overset{iid}{\sim} G_0$ and the weights p_ℓ are obtained by the following recursive expression:

$$p_1 = v_1 \, , \qquad p_\ell = v_\ell \prod_{j<\ell} (1 - v_j) \qquad \ell > 1 \, ,$$

where $v_\ell \overset{iid}{\sim} Beta(1, M)$. This last decomposition is called the *stick-breaking* representation of a DP prior (Sethuraman, 1994). The resulting random probability function is almost surely discrete (Blackwell and MacQueen, 1973). Properties of the DP prior and alternative constructions can be found in Ghosal (2010) and Lijoi and Prünster (2010), respectively. Other alternative random probability measures have been defined in the literature. For instance, Ishwaran and James (2001) proposed generalizations based on the stick-breaking definition. One of them is the finite DP which is obtained by truncating (2.3) after a level of truncation of N terms with $v_N = 1$ and $p_N = 1 - \sum_{i<N} p_i$. Posterior computations under this prior model are not difficult by using a block Gibbs sampler algorithm (Ishwaran and James, 2001).

Both density and distribution estimation problems are examples where Bayesian non-parametric models have been applied. However, because DP prior models produce discrete distributions with probability one, they are not suitable for modeling continuous outcomes. The DP mixture model (DPM) (Ferguson, 1983; Lo, 1984) has thus been proposed as a simple extension of DP models that solves this problem. The DPM is a mixture of a continuous density with a DP prior. Let $Z \sim F(\cdot)$, then a DPM model is defined as

$$F(z) \quad = \quad \int p(z \mid \theta) G(d\theta) \qquad G \sim DP(M, G_0(\boldsymbol{\eta})) \; ,$$

where for every $\theta \in \Theta$, $p(z \mid \theta)$ is a continuous density function, G is a DP defined on $\Theta \subset \mathbb{R}^p$ and $\boldsymbol{\eta}$ is a vector of hyperparameters that defines the base measure G_0.

The latent formulation for ordinal random variables described in section 2.2.1 has been useful in density estimation (see Shah and Madden, 2004; Ghosh et al., 2018). Denote by $Y_1, \ldots, Y_n$ a random sample from the ordinal random variable Y. Let $Z_1, \ldots, Z_n$ be the latent variable associated to $Y_i, i = 1, \ldots, n$ from the relation (2.2). In the context of modeling multivariate ordinal data without covariates, Kottas et al. (2005) assumed a DPM for the latent variable Z. In the univariate context, this model is stated as:

$$f_G(z) \quad = \quad \int_{\Theta} N(z \mid \mu, \sigma^2) G(d\theta) \; ,$$

where $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times \mathbb{R}^+$. Breaking the mixture, this model can be written as follows:

$$\begin{aligned} Z_1, \ldots, Z_n \mid \theta_1, \ldots \theta_n &\overset{ind}{\sim} N(\mu_i, \sigma_i^2) \\ \theta_1, \ldots, \theta_n \mid G &\overset{iid}{\sim} G \\ G &\sim DP(M, G_0(\boldsymbol{\eta})) \; , \end{aligned}$$

where G follows a Dirichlet process with a base measure $G_0(\boldsymbol{\eta})$ specified as a distribution function over $\mathbb{R} \times \mathbb{R}^+$.

Assuming a DPM model for the latent variable representation of ordinal random variables has several advantages over models mentioned in section 2.2.1. A DPM model can be used to approximate any distribution for continuous outcomes, which given the latent

continuous representation introduced earlier, implies the ability to approximate any probability distribution of ordinal variables. Moreover, as argued by Kottas et al. (2005), the accuracy of the approximation is not based on random thresholds and so without loss of generality one may assume fixed thresholds. From a practical point of view, this fact facilitates the estimation process because problems to estimate/update the thresholds are avoided.

Bayesian nonparametric models have also been proposed for psychometric modelling (Karabatsos and Walker, 2009b), in particular in the context of test equating. As was pointed out by Karabatsos and Walker (2011), traditional equating methods are all based on parametric assumptions to build a continuous version of the cdfs of X and Y. Instead, Karabatsos and Walker (2009a) proposed a Bayesian nonparametric model using Bernstein polynomials process priors for (F_X, F_Y) that account for dependence between the score distributions. As an extension, González et al. (2015a,b) developed a Bayesian nonparametric model for test equating which allows the use of covariates based on Bernstein polynomials models. Despite the fact that these proposals use more flexible models for the score distributions functions, none of these two approaches produce equated scores that are properly defined on the original discrete scale score. To deal with this problem, we develop an alternative equating method that allows us to obtain equated scores directly on the ordinal scale. The proposed method is based on a Bayesian nonparametric model for the latent variable associated to the ordinal score random variables. In what follows we discuss the proposed estimation method for the score distributions using density estimation procedures under the latent formulation strategy of ordinal outcomes.

2.2.3 Bayesian nonparametric latent approach for test equating

Let X and Y be two test forms. Following the notation described in González and Wiberg (2017), X and Y are the random variables representing the score under test X and Y, respectively. The score scales for these variables are $\mathscr{X} = \{\delta_0, \ldots, \delta_{C_X}\}$ and $\mathscr{Y} = \{\omega_0, \ldots, \omega_{C_Y}\}$, respectively. Because both $\mathscr{X}$ and $\mathscr{Y}$ define an order relation between scores, we consider X and Y as ordinal random variables. We develop the latent equating

method under the assumption that $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ are random samples from X and Y, taken under an equivalent group equating design. For more details about equating designs see von Davier et al. (2004, Chapter 2), Kolen and Brennan (2014, Section 1.2) and González and Wiberg (2017, Chapter 1).

Assuming that both X and Y are ordinal random variables, we define a model based on the latent representation described in section 2.2.1. Let

$$Z_{X,i} \overset{iid}{\sim} F_{Z_X} \qquad i = 1, \ldots, n_X \tag{2.4}$$

$$Z_{Y,j} \overset{iid}{\sim} F_{Z_Y} \qquad j = 1, \ldots, n_Y \ , \tag{2.5}$$

be the continuous latent variable associated to each sample score X_i and Y_j such that

$$X_i = \delta_h \overset{iid}{\sim} F_{Z_X}(\gamma_{X,h+1}) - F_{Z_X}(\gamma_{X,h}) \qquad i = 1, \ldots, n_X \tag{2.6}$$

$$Y_j = \omega_k \overset{iid}{\sim} F_{Z_Y}(\gamma_{Y,k+1}) - F_{Z_Y}(\gamma_{Y,k}) \qquad j = 1, \ldots, n_Y \ , \tag{2.7}$$

where $\gamma_{X,0}, \ldots, \gamma_{X,C_X+1}$ and $\gamma_{Y,0}, \ldots, \gamma_{Y,C_Y+1}$ are the thresholds of Z_X and Z_Y in the latent representation (2.2). Note that the relations (2.6) and (2.7) can be summarized as a multinomial distribution. In fact,

$$X_i \mid Z_{X,i} \sim Mult(1, C_X + 1, p_{Z_X,\boldsymbol{\gamma}}) \qquad i = 1, \ldots, n_X \tag{2.8}$$

$$Y_j \mid Z_{Y,j} \sim Mult(1, C_Y + 1, p_{Z_Y,\boldsymbol{\gamma}}) \qquad j = 1, \ldots, n_Y \ , \tag{2.9}$$

where $p_{Z_X,\boldsymbol{\gamma}} = (p_{X,\gamma_0}, \ldots, p_{X,\gamma_{C_X}})$, $p_{Z_Y,\boldsymbol{\gamma}} = (p_{Y,\gamma_0}, \ldots, p_{Y,\gamma_{C_Y}})$ are obtained as

$$p_{Z_X,\gamma_h} = F_{Z_X}(\gamma_{h+1}) - F_{Z_X}(\gamma_h) \qquad h = 0, \ldots, C_X$$

$$p_{Z_Y,\gamma_k} = F_{Z_Y}(\gamma_{k+1}) - F_{Z_Y}(\gamma_k) \qquad k = 0, \ldots, C_Y \ ,$$

where F_{Z_X} and F_{Z_Y} are defined in (2.4)-(2.5) and $Mult(a, b, p)$ denotes a multinomial distribution where a is the number of trials, b the number of categories and p is the vector of classification probabilities.

Before defining the equating method, we describe the proposed model for the scores distributions. Following Kottas et al. (2005) we propose a DPM model for the latent variables. The DP prior used here is a finite DP prior (see section 2.2.2) for both F_{Z_X} and

F_{Z_Y}. In both cases the truncation level of the finite DP is the number of possible scores on each test, i.e., $C_X + 1$ and $C_Y + 1$, respectively. We define the model only for X scores but a similar formulation can be made for Y.

Let us denote $\boldsymbol{X} = (X_1, \ldots, X_{n_X})$, $\boldsymbol{Z} = (Z_1, \ldots, Z_{n_X})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{C_X+1})$ where $\theta_j = (\mu_j, \sigma_j^2)$ are the mixing parameters in the DP definition (2.3). To express the hierarchical model for the data using a finite DP prior we introduce configuration variables $\boldsymbol{K} = (K_1, \ldots, K_{n_X})$ such that $K_i = \ell$ if and only if $\theta_{K_i} = (\mu_\ell, \sigma_\ell^2)$, for $\ell = 1, \ldots, C_X + 1$ and $i = 1, \ldots, n_X$. Then, the model for scores $\boldsymbol{X}$ becomes:

$$
\begin{aligned}
X_1, \ldots, X_{n_X} \mid Z_{X,1}, \ldots, Z_{X,n_X} &\overset{ind}{\sim} Mult(1, C_X + 1, p_{Z_X,\gamma}) && (2.10)\\
Z_{X,1} \ldots, Z_{X,n_X} \mid \theta_1, \ldots, \theta_N, K_1, \ldots, K_{n_X} &\overset{ind}{\sim} N(z_i \mid \mu_{K_i}, 1/\sigma_{K_i}^2) && (2.11)\\
K_1, \ldots, K_{n_X} \mid \boldsymbol{p} &\overset{iid}{\sim} \sum_{\ell=1}^{C_X+1} p_\ell \delta_\ell(K_i) && (2.12)\\
\theta_\ell \mid \lambda, \tau, \beta &\sim N(\mu_\ell \mid \lambda, \tau/\sigma_\ell^2) G(\sigma_\ell^2 \mid \alpha_0, \beta) && (2.13)\\
\lambda &\sim N(q_0, Q_0) && (2.14)\\
\tau &\sim IG(w_0, W_0) && (2.15)\\
\beta &\sim G(c_0, C_0) && (2.16)\\
M &\sim G(a_0, b_0)\ , && (2.17)
\end{aligned}
$$

where $G(\cdot \mid a, b)$ represents a Gamma distribution with parameter a and b, and the prior density for $\boldsymbol{p}$ is a special case of the generalised Dirichlet distribution. A similar and independent model is considered for modelling the scores Y.

The model (2.10)-(2.17) is completed by specifying the hyperparameters of λ, τ, β and M. These values were fixed in a similar way as was done by Kottas et al. (2005). Hyperparameters were fixed to have finite first and second prior moments for all the parameters involved in the model. In addition, to have flexible distribution over the mixing parameters of the DP, we considered the following configuration of hyperparameters: $\alpha_0 = 3$, $q_0 = 0$, $Q_0 = 49$, $w_0 = 3$, $W_0 = 49$, $c_0 = 20$, $C_0 = 2$, $a_0 = 4$ and $b_0 = 2$.

2.2.4 Latent equating method: a discrete equating method

At this point, we have only defined the model for the scores distributions of X and Y. In this section we define all the steps related to the novel equating method we propose, which we refer to as *Latent equating method*.

To explore the posterior distribution of the proposed model, a blocked Gibbs sampler algorithm (Ishwaran and James, 2001) was implemented. A description of the algorithm can be found in Appendix G. Posterior samples of all the parameters in the model were obtained after T iterations of the algorithm. In particular, we obtained samples from the cumulative posterior predictive distribution of Z_X given a new discrete score δ_h, i.e.,

$$F_{Z_X}^{(t)}(z) \;=\; \int_{-\infty}^{z} \sum_{k=1}^{C_X+1} p_k^{(t)} N(s \mid \mu_k^{(t)}, 1/\sigma_k^{2(t)}) ds \;, \tag{2.18}$$

where $\{(\mu_k^{(t)}, \sigma_k^{2(t)}), t = 1, \ldots, T\}$ are the sampled parameters from the posterior distribution. This is also made for scores $\boldsymbol{Y}$.

Using both $\{F_{Z_X}^{(t)}, t = 1, \ldots, T\}$ and $\{F_{Z_Y}^{(t)}, t = 1, \ldots, T\}$, posterior samples of the equipercentile equating function $\{\varphi_{Z_Y}^{(t)}(\cdot), t = 1, \ldots, T\}$ were obtained according to the definition in (2.1), i.e.,

$$\varphi_{Z_Y}^{(t)}(\cdot) = F_{Z_Y}^{-1(t)}(F_{Z_X}^{(t)}(\cdot)) \;. \tag{2.19}$$

Note at this point that the samples of the equipercentile function are random continuous functions. Our proposal takes advantages of this fact in the following way. Let $\delta_h^\star$ denote the score δ_h rescaled into the support of the latent variable Z_X. The value $\delta_h^\star$ is evaluated on each posterior sample of the equipercentile function (2.19). As a consequence, each score δ_h in the original scale $\mathscr{X}$ has associated a set of T random continuous equated values, i.e.,

$$Z_{Y,h}^\star = \{\varphi_{Z_Y}^{(t)}(\delta_h^\star), t = 1, \ldots, T\} \;. \tag{2.20}$$

The equated discrete score for δ_h will be ω_k, for some $k \in \{0, \ldots, C_Y\}$, if the interval $(\gamma_{Y,k}; \gamma_{Y,k+1}]$ (by means of equation (2.2)) has the highest probability on the distribution

of values (2.20). Mathematically, if $\varphi_Y(\delta_h)$ is the discrete equated score of δ_h in the scale $\mathscr{Y}$, then:

$$\varphi_Y(\delta_h) = \omega_k \quad \Leftrightarrow \quad k = \underset{k \in \{0,\ldots,C_Y\}}{\operatorname{argmax}} \; P\left(Z^{\star}_{Y,h} \in (\gamma_{Y,k}; \gamma_{Y,k+1}]\right) \; . \tag{2.21}$$

Once the equating function $\varphi_Y(\delta)$ has been estimated and equated values obtained for all $\delta \in \mathscr{X}$, it is needed to quantify the error associated to each of these values. In the equating setting, the uncertainty on equated values is measured by what is called the standard error of equating (SEE). Different ways to compute the SEE have been discussed in the equating literature (see Lord, 1982; von Davier et al., 2004). The estimation of the Bayesian non-parametric model behind the proposed method is based on MCMC (see Appendix G). In addition, given that the discrete estimated equated scores are obtained using the set (2.20) for each possible score scale, we propose to compute the standard error of equating of $\hat{\varphi}_Y(\delta_h)$ as the standard deviation of this set, i.e.,

$$SEE(\hat{\varphi}_Y(\delta_h)) \;\; = \;\; \sqrt{Var(Z^{\star}_{Y,h})} \qquad h = 0, \ldots, C_X \; .$$

To conclude this section, we remark that the proposed model guarantees the symmetry property of equating functions because the equipercentile transformation is computed based on continuous distribution functions.

2.3 Illustrations

In this section the performance of our proposal is illustrated in a simulation study and using a real data set. In the simulation study, discrete equated scores obtained from the latent equating method were compared with two traditional equating methods (equipercentile equating and Gaussian kernel equating). Because equated scores obtained under these methods are actually continuous scores, we use a discrete version of them in order to make a fair comparison with our proposal. The application is made using scores from two forms of a mathematics test presented in von Davier et al. (2004).

The proposed method was developed using both Fortran and the R software (R Core Team, 2018) without considering presmoothing methods. Traditional equating methods

were implemented using the SNSequate `R` library (González, 2014). In the simulation study, several data sets were generated considering different simulation scenarios. In this section we describe how discrete test scores were simulated and how true discrete equated scores were obtained.

The latent equating method proposed in this work considers scores sampled from an equivalent group design. In both, the simulation study and in the application to a real data set, we considered that both tests X and Y have the same number of items i.e., $C_X = C_Y = C$. Under this condition, the vector of thresholds which define the latent representations (2.6) and (2.7) are given by $\boldsymbol{\gamma}_X = (\gamma_{X,0}, \gamma_{X,1}, \ldots, \gamma_{X,C}, \gamma_{X,C+1})$ and $\boldsymbol{\gamma}_Y = (\gamma_{Y,0}, \gamma_{Y,1}, \ldots, \gamma_{Y,C}, \gamma_{Y,C+1})$. Because the DPM model assumed for the latent variables Z_X and Z_Y (see section 2.2.2) has the advantage of using fixed values for the vector of thresholds and both tests have the same number of items, there is no loss of generality considering the same vector of thresholds for both latent variables, i.e., $\boldsymbol{\gamma}_X = \boldsymbol{\gamma}_Y = (\gamma_0, \gamma_1, \ldots, \gamma_C, \gamma_{C+1})$.

In all cases we estimated equating functions to transform scores from $\mathscr{X}$ to $\mathscr{Y}$ scale. On each simulated scenario we considered continuous latent variables with support on $\mathbb{R}$. Using the relation (2.2) discrete scores were simulated considering thresholds values fixed to equidistant values between $\gamma_0 = -10$ and $\gamma_{C+1} = 10$. A *true* version of the equipercentile function $\varphi_{Z_Y}(\cdot) = F^{-1}_{Z_Y}(F_{Z_X}(\cdot))$ was possible to compute because both Z_X and Z_Y were continuous random variables. True discrete equated scores were obtained as the discrete score associated to the interval where the result of $\varphi_{Z_Y}(\gamma^\star_h)$ lied using the relation (2.2). Here $\gamma^\star_h$ is the midpoint of the interval $(\gamma_h; \gamma_{h+1}]$ for $h = 0, \ldots, C$.

2.3.1 Simulation study

We investigate the performance of the latent equating method using simulated data under the previous considerations. Several criterion were considered for generating different scenarios. The latent distributions of Z_X and Z_Y were assumed symmetric (S), left-asymmetric (LA) and right-asymmetric (RA). Symmetric and asymmetric distributions were simulated from normal and skew-normal distributions, respectively. Two general

scenarios were considered. The Scenario I considers that both latent variables have similar shape, location and scale parameters. The Scenario II considers that both latent variables have different shapes, locations and scales parameters. On each scenario several schemes were considered which differ in the shape assumed for both F_{Z_X} and F_{Z_Y}.

Table 2.1 summarizes all the schemes considered on both scenarios for the shape of the latent variables Z_X and Z_Y. Even Scenario II considers different shape, location and scale parameters, in the scheme 4 of Scenario II both pdfs were symmetric distributions (S) but with different mean and scale parameters.

Table (2.1) Description of the shape considered for the latent distributions F_{Z_X} and F_{Z_Y}: S (symmetric distribution), RA (right-asymmetric) and LA (left-asymmetric).

Scenario I			Scenario II		
Name	F_{Z_X}	F_{Z_Y}	Name	F_{Z_X}	F_{Z_Y}
Scheme 1	S	S	Scheme 4	S	S
Scheme 2	RA	RA	Scheme 5	LA	RA
Scheme 3	LA	LA	Scheme 6	RA	LA

For each scheme, three combinations of sample sizes were considered for the pair $n = (n_X, n_Y)$: $n_1 = (80, 100)$, $n_2 = (500, 500)$ and $n_3 = (1500, 1450)$. For each sample size, 100 replicates were simulated. Illustrations for each of these schemes showing true continuous latent distribution for f_{Z_X} and f_{Z_Y}, true continuous equating function and the true discrete scores obtained for each possible score of the scale for each scheme, are shown in the Appendix A and B. Depending on the scheme, different equipercentile functions are obtained at the same time that different relations between scale scores in $\mathscr{X}$ and their true discrete equated score are developed.

Results of the latent equating method were obtained using three Markov chains generated starting from different initial values. After completing a total number of 60000 iterations and a burn-in period of 30000 iterations, each chain was subsampled every 25 iterations. Combining these chains, the result was a chain of length 3600. The convergence of the

chains was analyzed by using the statistic $\hat{R}$ (Brooks and Gelman, 1998; Gelman and Rubin, 1992) and the effective sample size (Kass et al., 1998). Results, not shown here, suggested convergence of the chains.

We divide the analysis of the proposal's performance in two parts. We assess first the fitting of the equipercentile function based on the Bayesian nonparametric model for the score distributions. Secondly, we evaluate the latent equating method proposed in this chapter in terms of the discrete estimated equated scores. The latter analysis is made in both the whole range of the scale and on each possible value of the scale score.

The fitting of the proposed Bayesian nonparametric model for the score distributions allows us to estimate the equipercentile function in the latent setting. To evaluate this estimation, the criterion used is the expected value of the L_2 norm between the estimation and the real equipercentile function, with respect to the sampling distribution. For more details of the definition of the statistic and how Monte Carlo method was applied, see Appendix H. Note that both real and estimated equipercentile functions are evaluated in the real line. For this reason we did not compute the L_2 norm of the difference between these two functions using traditional equating methods.

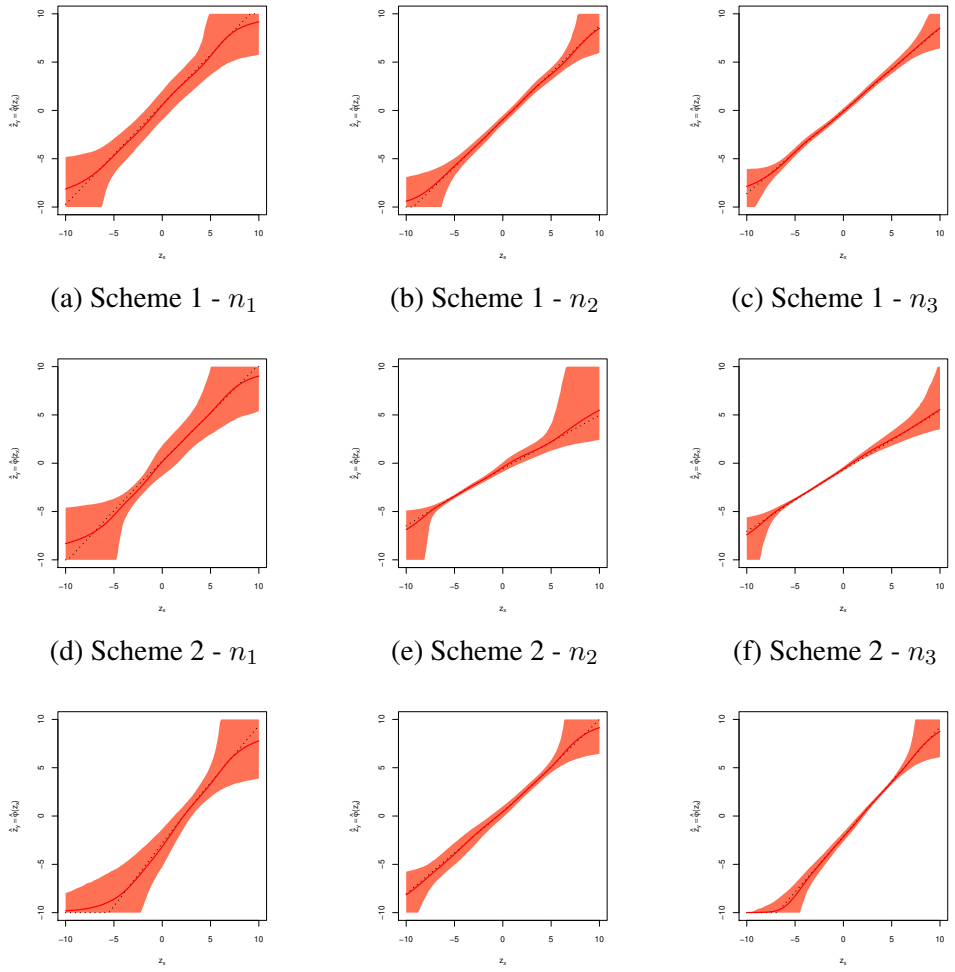
Table 2.2 shows the results for the six simulated schemes. The L_2 norm of the difference between real and estimated equipercentile function decreases as sample size increases. In the Scenario II, where the latent distributions are different in shape, almost all of these values are greater than in the Scenario I.

To illustrate the results of the estimation of the equipercentile function in the latent setting, a random sample of the replicates of each scheme was selected. The estimation of the equipercentile function with the 95% point-wise HPD interval on each sample is shown in Figures 2.1 and 2.2. The true equipercentile function is well estimated in almost all the simulated schemes. All credible intervals contain the true equating function. For small sample sizes (n_1) there is more variability in the estimation (thicker HPD intervals) than for higher sample sizes (n_2 and n_3). In the two scenarios, there is higher variability at the beginning and at the end of the continuous scale, for all the schemes. A possible explanation is that under the simulated truth, scores are concentrated in the middle portion of the

Table (2.2) Simulated data: Estimated L_2 norm of the difference between true continuous equipercentile function and its estimation from the proposed method under different simulation schemes and sample sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$, $n_3 = (1500, 1450)$.

Scenario I				Scenario II			
Scheme	n_1	n_2	n_3	Scheme	n_1	n_2	n_3
Scheme 1	6.902	3.744	2.825	Scheme 4	6.347	3.348	2.636
Scheme 2	6.611	5.314	4.176	Scheme 5	13.709	9.280	7.978
Scheme 3	7.927	3.957	2.720	Scheme 6	13.341	9.353	7.431

scale leaving little mass in the extremes. Thus there is more uncertainty in the estimation in the extremes of the scale. Latent models that allow more mass in the extremes of the scale are evaluated later.



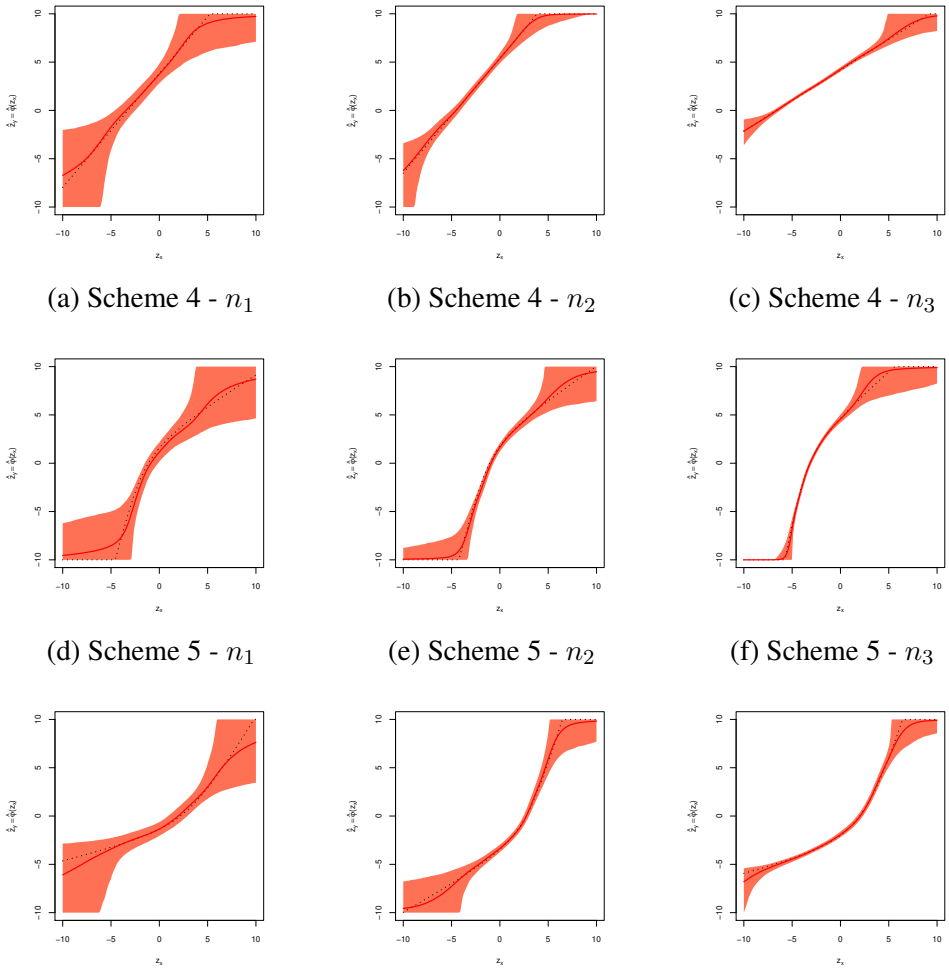
(a) Scheme 1 - n_1 (b) Scheme 1 - n_2 (c) Scheme 1 - n_3

(d) Scheme 2 - n_1 (e) Scheme 2 - n_2 (f) Scheme 2 - n_3

(g) Scheme 3 - n_1 (h) Scheme 3 - n_2 (i) Scheme 3 - n_3

Figure (2.1) Scenario I: True (dashed line) equipercentile function and its estimation (red line) for all samples sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$,$n_3 = (1500, 1450)$ on each scheme. The point-wise 95% HPD interval is displayed as the colored area.

After the estimation of the equipercentile function, the proposed approach defines a method to obtain discrete equated scores from this function (see section 2.2.3). To summarize the information about the method performance in the whole scale score, we examined the statistic Ψ_2 defined as the L_2 distance between true discrete equated scores and its esti-



(a) Scheme 4 - n_1 (b) Scheme 4 - n_2 (c) Scheme 4 - n_3

(d) Scheme 5 - n_1 (e) Scheme 5 - n_2 (f) Scheme 5 - n_3

(g) Scheme 6 - n_1 (h) Scheme 6 - n_2 (i) Scheme 6 - n_3

Figure (2.2) Scenario II: True (dashed line) equipercentile function and its estimation (red line) for all samples sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$,$n_3 = (1500, 1450)$ on each scheme. The point-wise 95% HPD interval is displayed as the colored area.

mators. For a formal definition of this statistic and how it was computed, see Appendix I. The statistic Ψ_2 was also computed under estimations obtained from two traditional equating methods: equipercentile equating (EQ) and Gaussian kernel equating (KE). As was

previously mentioned, these methods result on continuous equated values (i.e., not defined in the original discrete scale). In order to properly compare these methods with discrete scores obtained under the latent equating method, we used the largest integer number not greater than the corresponding continuous equated score obtained from traditional equating methods in the evaluation of Ψ_2.

A summary of the computation of the statistic Ψ_2 for all the schemes is displayed in Table 2.3. Lowest values of this statistic are found for the latent equating method in contrast to values obtained for traditional equating methods. On each simulated scheme, the proposed method performs better than evaluated competitors. In addition, as sample size increases, our method works better in all the schemes considered.

It can be seen in Figures 2 and 3 (see Appendix A and B), that each scheme represents different relations between scores and true discrete equated scores. The consequence of the lowest values of Ψ_2 for the latent equating method is that, at least in the whole scale score, the method (in mean) is closer to the true discrete equated scores than discrete version estimations from traditional equating methods.

To further analyze the performance of the latent equating method, we also studied the behavior among each possible value of the scale score. To achieve this objective, if δ_k is a score on the scale $\mathscr{X}$, for $k = 0, \ldots, C$, we computed the expected value of the difference between the true discrete equated score associated to the score δ_k an its estimated discrete equated score under the proposed method. This expectation was approximated by using the 100 replicates for each scheme. To compare the proposed method with traditional equating methods, we also evaluate this quantity using both equipercentile equating and Gaussian kernel equating. Discrete equated scores estimated on each value of the scale from these methods were obtained as the largest integer number not greater than the corresponding continuous equated score for each method. Figures 5 and 6 (see Appendix D and E) show the results of this expected value for the three equating methods on all the schemes and for all sample sizes. In the three schemes considered in Scenario I, the expected value under the latent equating method are quite similar to the values under equipercentile equating for low scores of the scale. In contrast, for higher scores of the scale, the latent equating

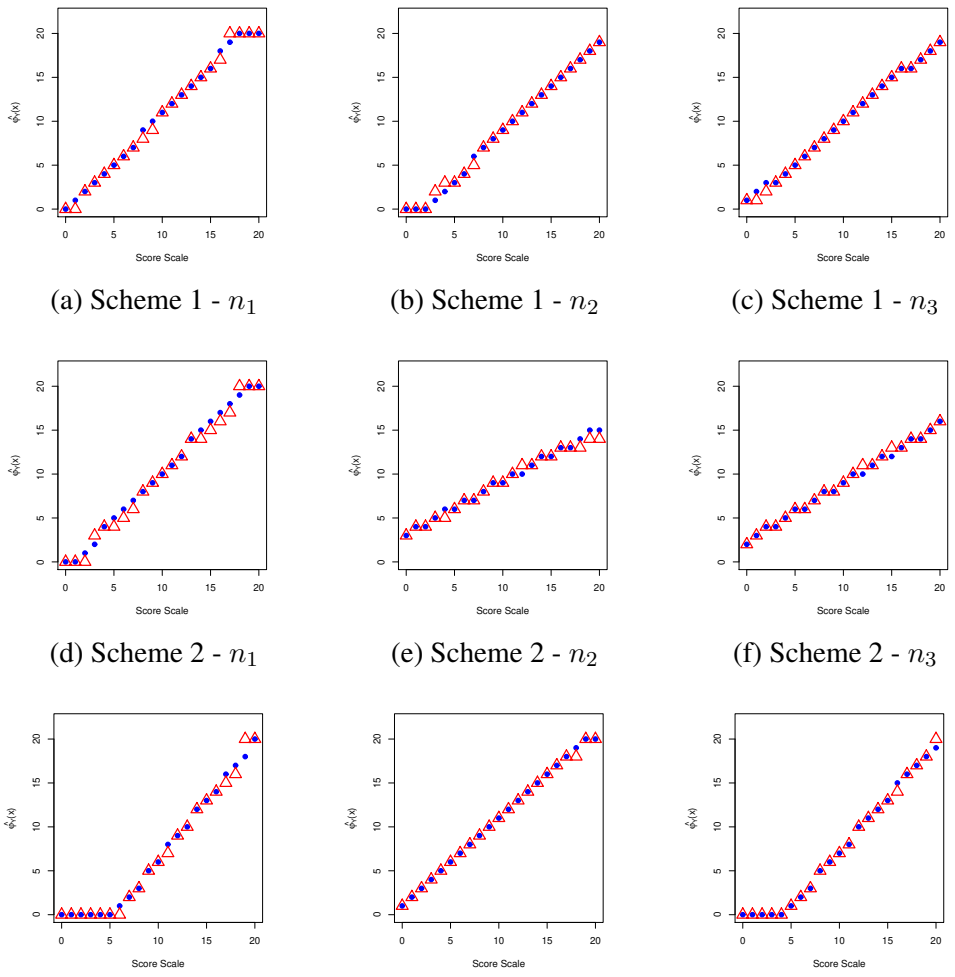
Table (2.3) Simulated data: Estimated values of Ψ_2, the L_2 norm of the difference between the vector of true discrete scores in the whole scale and its estimation from the latent equating method (LE) and discrete version of the equipercentile equating (EQ) and the Gaussian kernel equating (KE). All simulated schemes are evaluated for the sample sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$, $n_3 = (1500, 1450)$.

	Scenario I								
	Scheme 1			Scheme 2			Scheme 3		
Method	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3
LE	4.401	2.322	1.758	4.569	3.665	2.747	4.980	2.563	1.718
EQ	5.451	3.804	3.543	5.324	5.533	5.012	6.308	4.011	3.773
KE	5.664	6.260	6.815	5.935	6.706	6.190	6.000	5.157	6.140
	Scenario II								
	Scheme 4			Scheme 5			Scheme 6		
Method	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3
LE	3.897	2.031	1.513	5.939	4.070	3.384	7.352	4.258	3.205
EQ	5.324	3.477	2.780	6.220	4.554	4.600	9.817	7.387	6.334
KE	8.260	7.980	8.836	8.773	9.321	9.944	12.586	13.166	12.968

method values are lower than those from both traditional equating methods. Moreover, in most of the scale, the Gaussian kernel equating method shows the highest values. In schemes of Scenario II, the results of the equipercentile equating are lower than the proposed method for lower scores of the scale in schemes 5 and 6 for the smaller sample size. Nonetheless, in the rest of the scale, the latent equating method has lower values. Similar to Scenario I, the Gaussian kernel method shows the worst performance in almost all the scale score.

To illustrate the results obtained from the latent equating method proposed in this chapter, Figures 2.3 and 2.4 show the estimated discrete equated scores obtained for the random samples selected in the Figures 2.1 and 2.2 respectively. In both scenarios the method can recover real equated scores for almost all scores on the scale. As sample sizes increase

the precision of the estimation is better. However, there is less precision in the estimation when small sample sizes are considered.

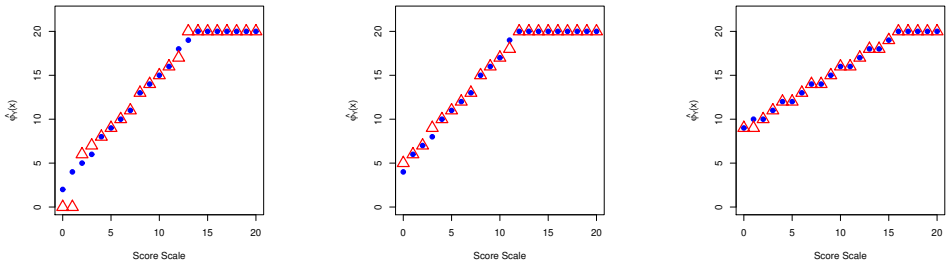


(a) Scheme 1 - n_1 (b) Scheme 1 - n_2 (c) Scheme 1 - n_3

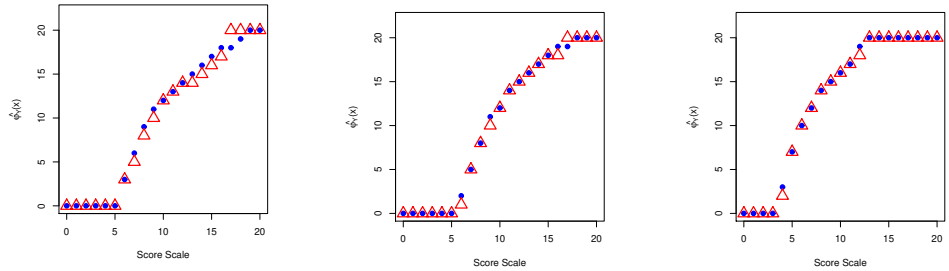
(d) Scheme 2 - n_1 (e) Scheme 2 - n_2 (f) Scheme 2 - n_3

(g) Scheme 3 - n_1 (h) Scheme 3 - n_2 (i) Scheme 3 - n_3

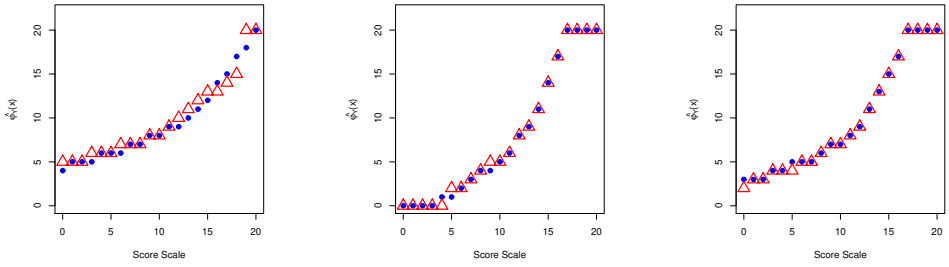
Figure (2.3) Scenario I: True discrete equipercentile scores (blue dot) and its estimation using the latent equating method (red triangle) for all samples sizes on each scheme.



(a) Scheme 4 - n_1 (b) Scheme 4 - n_2 (c) Scheme 4 - n_3



(d) Scheme 5 - n_1 (e) Scheme 5 - n_2 (f) Scheme 5 - n_3

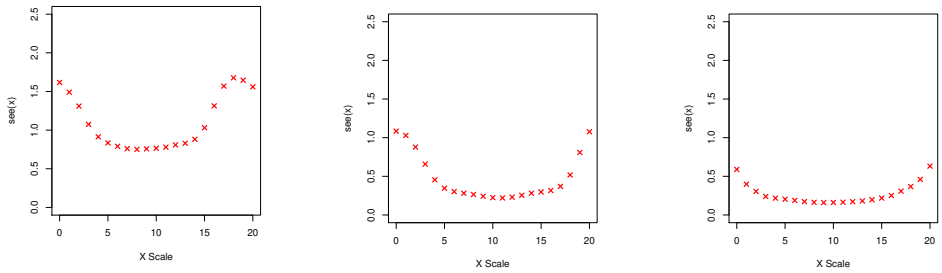


(g) Scheme 6 - n_1 (h) Scheme 6 - n_2 (i) Scheme 6 - n_3

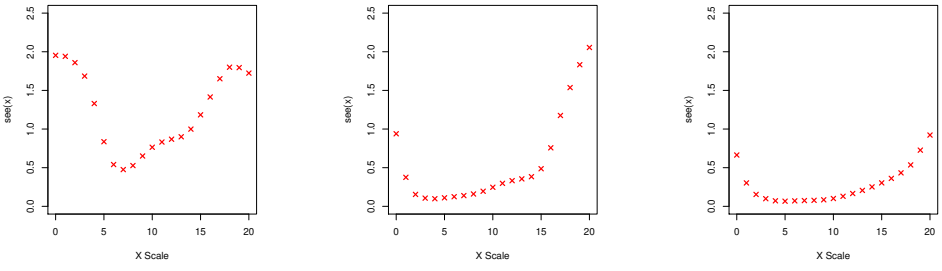
Figure (2.4) Scenario II: True discrete equipercentile scores (blue dot) and its estimation using the latent equating method (red triangle) for all samples sizes on each scheme.

In fact, the standard error of equating for each of the cases considered before (see Figures 2.5 and 2.6) show that there is more variability in the estimation for lower and higher scores of the scale when sample sizes n_1 and n_2 are considered in both scenarios. Nevertheless for the latter sample size, the SEE's are lower. Lower standard errors are found for almost all scores when high samples sizes are involved in both scenarios, with the highest SEE's

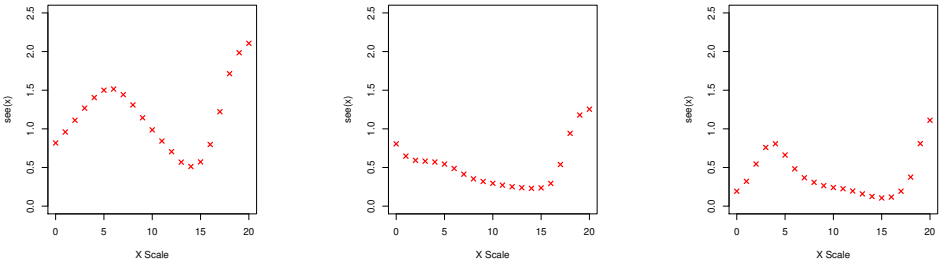
for higher scores.



(a) Scheme 1 - n_1 (b) Scheme 1 - n_2 (c) Scheme 1 - n_3



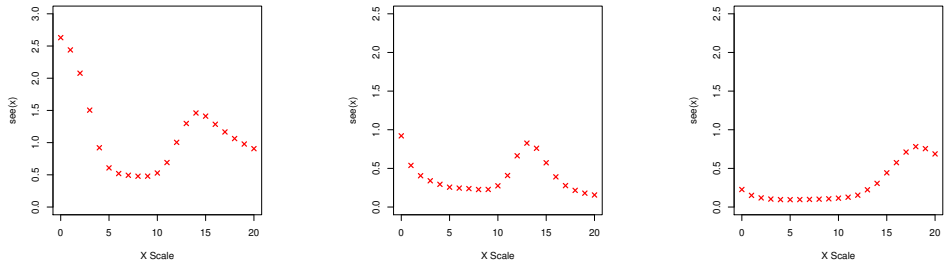
(d) Scheme 2 - n_1 (e) Scheme 2 - n_2 (f) Scheme 2 - n_3



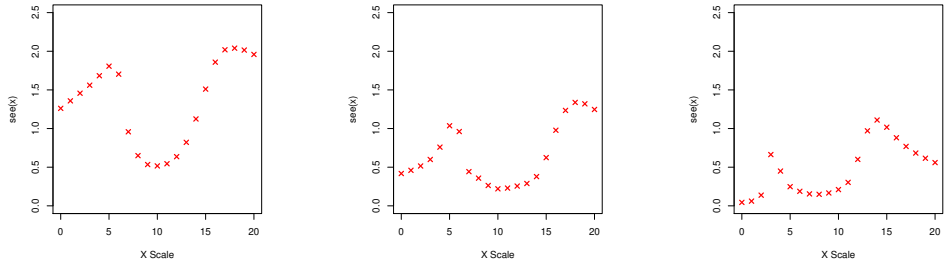
(g) Scheme 3 - n_1 (h) Scheme 3 - n_2 (i) Scheme 3 - n_3

Figure (2.5) Scenario I: Standard error of equating for samples sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$, $n_3 = (1500, 1450)$ on each scheme.

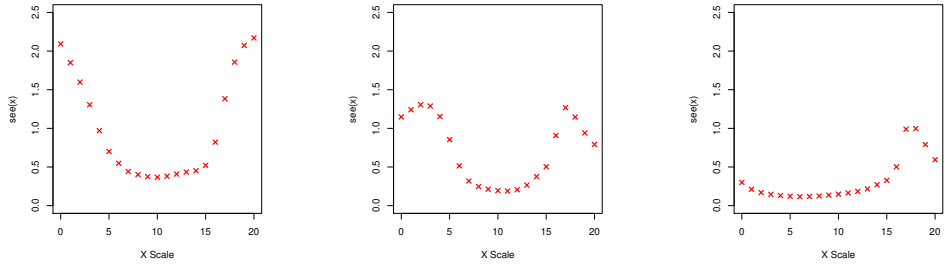
To better examine the extra variability shown in the estimation of the equipercentile function at the extremes of the scale, we carried out an additional simulation study, now considering score distributions with more mass at the extremes. In this new simulation scenario we considered bimodal latent distributions for both F_{Z_X} and F_{Z_Y} (a mixture of



(a) Scheme 4 - n_1 (b) Scheme 4 - n_2 (c) Scheme 4 - n_3



(d) Scheme 5 - n_1 (e) Scheme 5 - n_2 (f) Scheme 5 - n_3



(g) Scheme 6 - n_1 (h) Scheme 6 - n_2 (i) Scheme 6 - n_3

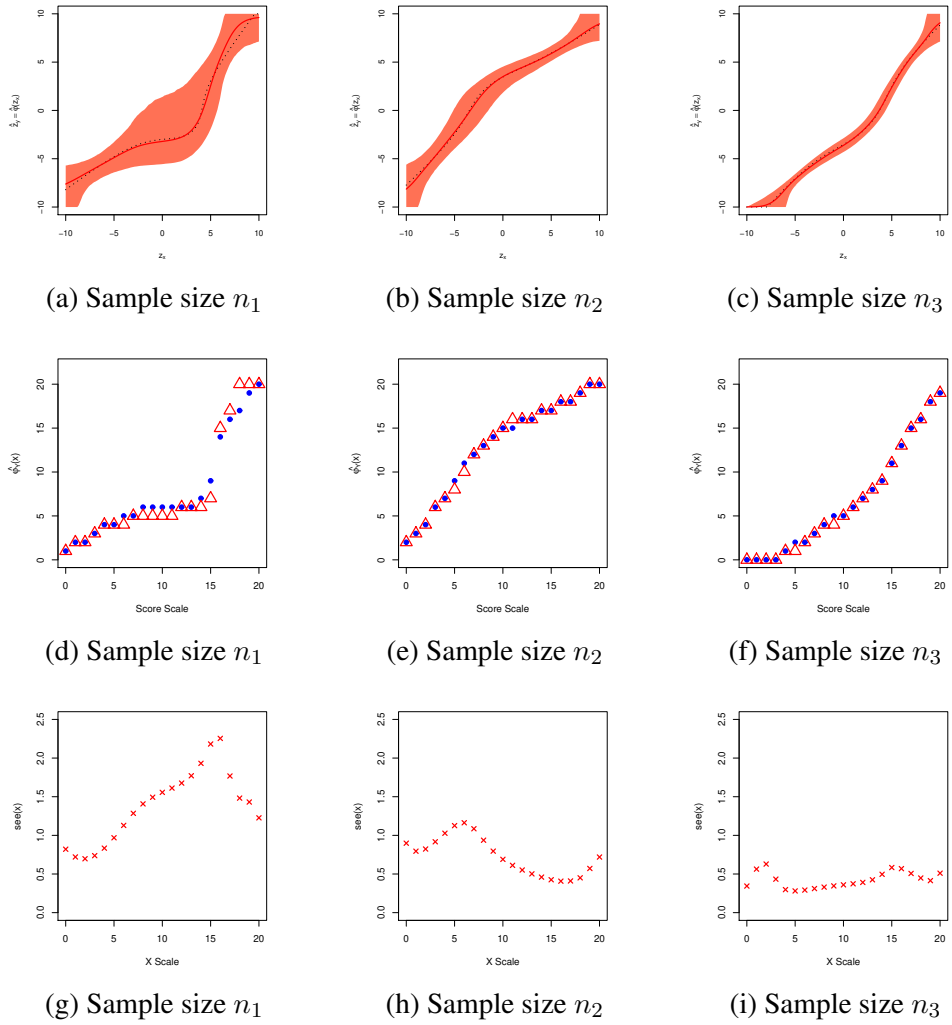
Figure (2.6) Scenario II: Standard error of equating for samples sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$, $n_3 = (1500, 1450)$ on each scheme.

normal distributions). Using these distributions, frequencies for low and high scores can be observed. The latent pdfs considered, the real equipercentile function and the true discrete equated scores of a random replicate for each sample size, are shown in Appendix C.

The analyses of the results under this new scenario follow those made for scenarios I and II. The estimated latent equipercentile function was evaluated using the L_2 distance

between the real and the estimated equipercentile function (see Appendix H). The results for $(n_1; n_2; n_3)$ are L_2=$(8.425; 6.580; 1.867)$. These values are similar to those found in Scenario I and Scenario II (see Table 2.2). The performance of the estimation is shown in Figure 2.7. The estimation of the equipercentile function has lower variability in the extremes of the scale in comparison to the results found for Scenario I and Scenario II. In fact, for almost all the cases, the width of the HPD intervals remain constant along the scale. The estimation of the equated scores is more accurate than in Scenario I and Scenario II with lower SEE for almost all the schemes considered in both Scenarios (see Figure 2.7). The differences between real discrete equated scores and a discrete version of equated scores obtained from traditional equating methods is shown in Appendix F. For small sample sizes all methods show a higher variability with respect to results from the two previous scenarios. In contrast, as sample sizes increase, the latent equating method has the lowest values for almost all the scores as was found in both scenarios I and II (see Appendix D and E).

We also computed the statistic Ψ_2 to evaluate the accuracy of estimated equated scores of the latent equating method under bimodal latent distributions. Results are compared to those coming from a discrete version of traditional equating methods. Table 2.4 summarizes the results. In contrast to discretized versions of equated values obtained from equipercentile equating and Gaussian kernel equating, the proposal has the lowest values for all sample sizes. Some differences are found when small samples sizes are considered. Considering this new simulation scenario we can conclude that the model performs remarkably well in several situations including symmetric, asymmetric, and bimodal score distributions.



(a) Sample size n_1 (b) Sample size n_2 (c) Sample size n_3

(d) Sample size n_1 (e) Sample size n_2 (f) Sample size n_3

(g) Sample size n_1 (h) Sample size n_2 (i) Sample size n_3

Figure (2.7) Bimodal latent distributions: Considering sample sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$, $n_3 = (1500, 1450)$, in the first row the true (dashed line) equipercentile function and its estimation (red line) are shown. The point-wise 95% HPD interval is displayed as the colored area. In the second row are exhibited the estimated discrete equated scores. In the last row the estimated SEE for each scale score are exposed.

Table (2.4) Estimated values of Ψ_2 for the latent equating method (LE) and discrete version of the equipercentile equating (EQ) and the Gaussian kernel equating (KE) for sample sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$, $n_3 = (1500, 1450)$ considering bimodal latent distributions.

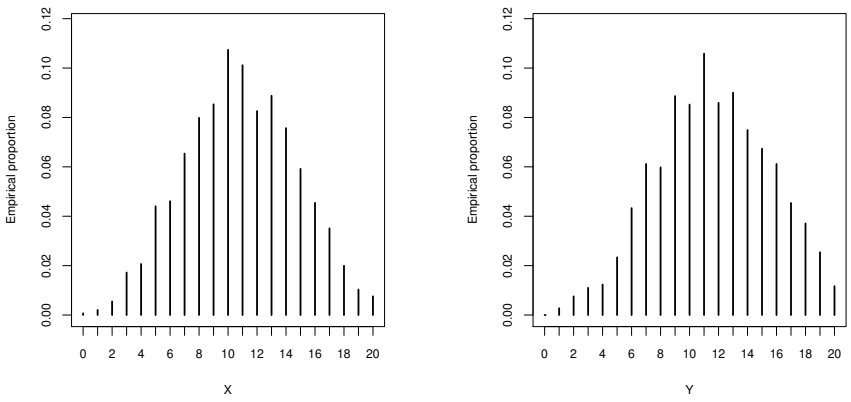
Method	n_1	n_2	n_3
LE	6.652	3.704	1.875
EQ	6.589	4.565	3.800
KE	6.721	6.191	5.623

2.3.2 Application

We consider a data set described and analyzed by Holland and Thayer (1989) and von Davier et al. (2004). The data set contains raw sample frequencies of number-right scores for two parallel 20-items mathematics tests, named X and Y, given to two samples from a national population of examinees. The number of test takers for each test is $n_X = 1453$ and $n_Y = 1455$. The empirical proportion of the observed scores for both tests are summarized in Figure 2.8. Test Y has higher frequencies for high scores scale than test X.

To equate scores X to scores Y, we apply the latent equating method proposed and also we compare the results obtained from traditional equating methods. We assumed the same model described in section 2.2.3, considering thresholds values described at the beginning of this section. The hyperparameters for the priors distributions involved in the model are: $\alpha_0 = 3$, $q_0 = 0$, $Q_0 = 49$, $w_0 = 3$, $W_0 = 49$, $c_0 = 20$, $C_0 = 10$, $a_0 = 4$ and $b_0 = 2$. Results of applying the latent equating method in the latent setting are summarized in Figure 2.9. The equipercentile function is estimated with high precision because of the thinning confidence bands, with high variability at the beginning of the real scale.

Without applying a presmoothing method, discrete scores estimated by the method and traditional equating methods (equipercentile equating and kernel equating) are displayed in Figure 2.10. Estimated discrete equated scores obtained from the three methods are quite similar for lower score of the scale. In the middle of the range of the scale, the



(a) Observed test scores X (b) Observed test scores Y

Figure (2.8) Application: Empirical proportion of two parallel mathematics test X and Y (von Davier et al., 2004)

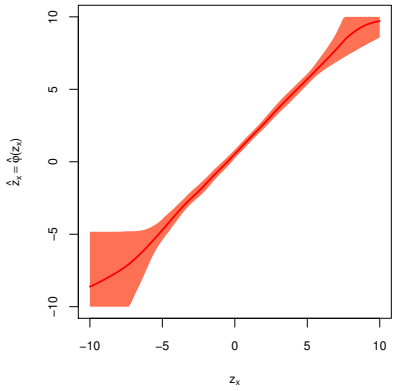
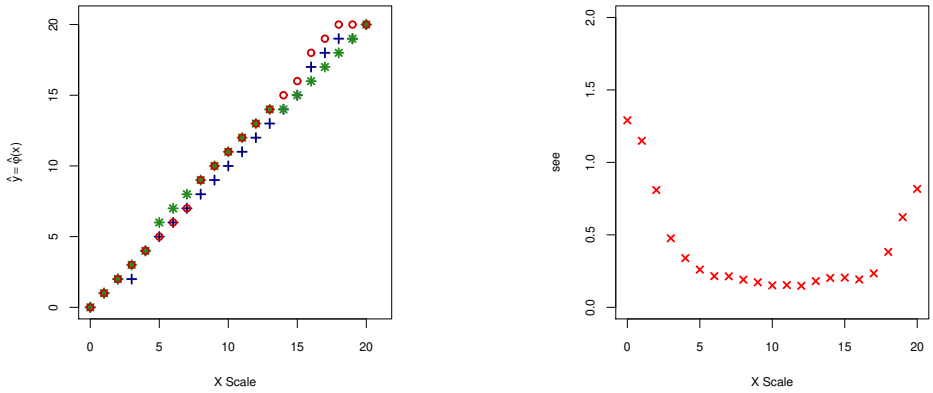


Figure (2.9) Estimated latent equipercentile function (continuous red line). The point-wise 95% HPD interval is displayed as the colored area

estimations from our method are more similar to discrete version of the equipercentile method than the Gaussian kernel method. For high scores, all the methods make different estimations. Only in the case of the highest score 20 all methods estimate the same equated score. With respect to the SEE, low variability in the estimation can be found for scores in

the range from 5 to 17. Higher variability is found at the beginning and at the end of the scale.



(a) Estimated discrete equated scores (b) Estimated SEE

Figure (2.10) Application: (a) Discrete equated scores estimated under the latent equating method (red circle), equipercentile equating (green asterisk) and Gaussian kernel equating (blue +). (b) Estimated SEE from the latent equating method.

2.4 Conclusions and discussion

Different parametric, semiparametric and nonparametric models have been proposed to estimate the equating transformation (González and von Davier, 2013). In all these methods, the equating transformation gives as a result continuous equated scores disregarding the fact that scores are actually defined on a discrete scale. In this chapter we introduced an equating method that produces equated scores that are properly defined on the original discrete scale. The continuization step commonly used in traditional equating methods is avoided under the proposed Latent equating method by considering scores as ordinal random variables.We apply ideas of Kottas et al. (2005), assuming a Bayesian nonparametric model for the latent representation of ordinal variables, which we use as a basis for developing the proposed method.

Results based on a simulation study have shown that the proposed method estimates accurately the equipercentile function in the latent setting, but with some variability in the extremes of the real scale. In comparison with discrete versions of equated values obtained by traditional equating methods, our approach has better performance considering the whole range of the scale. In almost all the simulated scenarios considered, the proposed method accurately estimated the true discrete scores on each possible test score. Considering the latent equating method applied to the data set which illustrates the equivalent group equating design (von Davier et al., 2004), our results show that the method estimates the latent equipercentile function accurately. The discrete equated scores estimated from our method and those from discretized versions of traditional equating methods differ in several parts of the scores scale. Using the latent score approach has an impact on estimated discrete equated scores in contrast to discrete version of traditional equating methods. The method also has low estimated standard error of equating for nearly all the range of scores.

Although other approaches based on a Bayesian nonparametric models have been proposed (Karabatsos and Walker, 2009a; González et al., 2015b), we take advantage of the ideas in Kottas et al. (2005) to obtain equated scores that are defined in the original scale of the tests: the latent equating method equates scores defined on a discrete scale into scores defined in a discrete scale. This idea, to the best of our knowledge, has not been developed before in the field of equating methods. In fact, the same approach can be applied in situations where it is of paramount importance to obtain equated scores in a discrete scale.

Karabatsos and Walker (2009a) discussed some drawbacks of traditional equating methods which motivated the proposed method. Their first comment is that traditional equating methods make some assumptions about the cdf's of X and Y which do not guarantee that the equated scores belong in the scale they are defined. Also, they mention that those assumptions are not made on well-founded reasons. We not only strongly agree with the authors that equated scores must belong in the original scale but also, with the proposed method, we want to add that equated scores must be discrete when they are defined on discrete scales. As a consequence, if score scales were subsets of the integer numbers, equating methods should be developed such that estimated equated scores are discrete too.

The proposed approach can be extended in different ways. The DPM model can be replaced by alternative models leading to estimation of continuous probability distributions, such as Polya trees processes (Mauldin et al., 1992; Lavine, 1992) and mixture of Pólya tree processes (Hanson and Johnson, 2002). Other extensions of the proposed model could consider covariate-dependent Bayesian nonparametric models for the latent variables (MacEachern, 1999, 2000; De Iorio et al., 2004) such that the shape of the latent scores distribution may change as a function of the covariates and, as a consequence, the form of the equating transformation can change according to covariate values.

The proposed equating method was developed for samples from an equivalent group equating design. Extending the approach to other equating designs is a topic for future research.

Chapter 3

A covariate-dependent Bayesian nonparametric approach for linking measurements

"Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise."
Tukey (1962).

3.1 Introduction

In all scientific research areas decisions are taken based on several measurements collected either by an experiment or in the setting of observational studies. Measurements are obtained by means of "measurement methods" which could be an instrument, an essay, a medical device, a clinical observer or even a (self-reported) test (Choudhary and Nagaraja, 2017). These last instruments are common in areas where cognitive variables are involved. In general, "no single instrument for cognitive screening is suitable for global use" (Cullen et al., 2007), so the development of new screening instruments for assess-

ing cognitive function has increased during the last years (van Steenoven et al., 2014). In clinical psychology as well as in cognitive psychology and neuropsychology, depression is considered a disabling and often a chronic mental disorder. Accurate measurement of depression is essential for diagnosis and treatment planning. To this purpose, it is popular the use of self-administered scales that measure and categorize symptoms of depression (Titov et al., 2011). In contemporary clinical practice and research, many different instruments are used to measure depression severity (Fried, 2016). In fact, no fewer than 280 depression scales have been developed in the last century, many of which are still in use (Santor et al., 2009).

The amount of measurement instruments used to measure a variable of interest might not be considered as a problem. Instead, the turning point of this phenomena is how to relate measurements obtained from different instruments. Because different instruments could lead to different results and so the conclusions obtained from them, if the outcome scores of different health assessment instruments are not properly linked, inferences based on them could have serious consequences such as wrong classification of cognitive impairment, misperceptions about the efficacy of a treatment, among others (Dorans, 2007). Consequently, the development of statistical procedures to establish how measurements coming from different instruments are related is relevant not only for using all the available information in the making decision process, but also to facilitate interpretations and communication among researchers in order to define comparable measurements. This last term means that measurements obtained from different instruments, which measure the same construct, can be used interchangeably. The objective of interchangeably measurements is one of the goals of method comparison studies (MCS, Choudhary and Nagaraja, 2017) where two competing methods/instruments of measurement, e.g., two procedures for measuring glucose concentration in a blood sample, are compared by the evaluation of the extend of agreement and the evaluation of similarities among them. However, in the context of cognitive scales and in particular for depression scales, these methods are not completely suitable for several reasons. Most MCS consider methods/instruments having a common (and generally continuous) unit of measurement whereas cognitive measure-

ments, including depression scales, are mostly defined on different scales which usually are either subsets of integer numbers or ordinal scales. In addition, MCS are more commonly used when measurements are obtained from the same samples, i.e., related samples, which is not the common case in the comparability of cognitive scales where samples are not commonly dependent.

The goal to obtain interchangeable measurements has also been developed in the context of educational measurement by equating and linking methods. Scores obtained in different forms of a test administered to the same or different groups of examinees are used to make important decisions, such as determining academic admissions or to whom scholarships should be granted. Although test forms are built to measure the same construct, the difficulty is implicitly not the same among all the forms. As a consequence, it is important for the decision making process to report scores as fair and accurate as possible to finally treat scores as if they come from the same test (Holland and Rubin, 1982; von Davier et al., 2004; Dorans et al., 2007; Kolen and Brennan, 2014; González and Wiberg, 2017). Such purpose is achieved by estimating the *equating transformation*, a function which maps the scores on the scale of one test form, X, into their equivalents on the scale of another, Y (González and Wiberg, 2017). The equipercentile equating transformation (Braun and Holland, 1982) is the most popular equating function defined as:

$$\varphi_Y(x) \quad = \quad F_Y^{-1}\left(F_X(x)\right) \ , \tag{3.1}$$

where F_Y and F_X are the cumulative distribution functions (cdf) of test score Y and test score X, respectively. Test scores are mostly considered to be sum scores (i.e., the total number of correct answers), such that the scale where they are defined are subsets of the integer numbers. The discreteness of the test scales generates a problem in (3.1) because it is almost impossible to find a value $y = F_Y^{-1}(p)$ in the scale of test Y such that $p = F_X(x)$ for any x score in the scale of test X. As it was mentioned before, cognitive instruments, for example depression scales, are mostly characterised by discrete or ordinal scales. This fact makes current equating methods not suitable at all for comparing cognitive instruments.

Different methods have been proposed in the equating literature to tackle the problem of discrete scores, all of them based on the *continuization* of the originally discrete score dis-

tributions F_Y and F_X. Continuization methods include the use of linear interpolation (Angoff, 1971; Braun and Holland, 1982) and kernel smoothing techniques (von Davier et al., 2004). However, a common feature of all equating methods based on the continuization of F_X and F_Y is that equated scores are not integer numbers anymore and thus are not defined on the original scale score. Moreover, as mentioned in Karabatsos and Walker (2011), the estimations of the equipercentile function from traditional equating methods can equate scores that fall outside the original range of the scores. These are not big problems in educational setting as raw scores are usually rescaled using certain (arbitrary) scale (Kolen and Brennan, 2014). However, an important consequence of this issue when comparing cognitive scales is that equated measurements do not meet the need of a discrete measure if reported as unrounded, not integer values. Notwithstanding, as cognitive measurements are used to define levels of cognitive impairment, rounding equated measurements could result, for instance on misclassification of cognitive impairment. An additional issue of traditional equating methods is that all of them assume that scores distributions of tests measuring the *same construct* are independent. We strongly agree with the discussion in Karabatsos and Walker (2011) stating that the equal construct requirement of tests involved in a equating procedure (see Section 1.2) implies that scores distributions should be related.

We propose a model-based procedure for linking measurements obtained from different instruments that can be applied to measurements defined on either continuous or discrete scales. It is shown explicitly how we tackle the problem of preserving the discreteness of the measurements while comparing discrete measurements and also how measurements' distributions are related to each other. The motivation comes from datasets of two self-administered instruments used to measure symptoms of depression applied in two independent samples of the Chilean population. The main objective is to develop equivalent measurements between these two instruments such that they can be used interchangeably. As a consequence, the use of different instruments for researchers as well as for practitioners would not be an obstacle to characterise depression symptoms in the Chilean population. In addition, because the prevalence of depression is higher for Chilean females

than for Chilean males (Chilean Health Ministry, 2016-2017), it is of interest to evaluate if this fact can lead to different equivalent measurements for males and females. If it is so, a relevant aspect is to know how they differ. By considering measurements from different instruments as ordinal random variables, a procedure to link them is developed based on the ideas of equating methods. However, the proposal avoids the use of continuous versions of the discrete distributions of measurements as well as rounded methods mentioned before. The key element of the proposal is the continuous latent representation of ordinal random variables. It allows to estimate a latent equipercentile function of the form (3.1) from the estimation of the continuous latent distributions of the measurements. Then, because of the one-to-one relation between the latent variable and the ordinal measurements, a procedure is defined to recover measurements in the original discrete ordinal scale. Moreover, the model proposed for the latent distributions includes additional information of the sample units based on a covariate-dependent Bayesian nonparametric model. Thus, a customised equipercentile function can be estimated for subgroups of interest (González et al., 2015b). Even though covariate-dependent and Bayesian nonparametric models have been used to define equating methods (see Karabatsos and Walker, 2009a; González et al., 2015b), a different approach of the proposal developed in this chapter is that linked measurements are properly defined on the scales defined by the instruments. The proposal extends the model defined by Varas et al. (2019) which, to the best of our knowledge, is the first work proposed in this direction.

The rest of the chapter is organised as follows. In Section 3.2 the two depression scales are described. In Section 3.3 some common approaches to model ordinal variables and important features of covariate-dependent Bayesian nonparametric are mentioned. Both the model considered for the measurements' distributions and the linking method proposed are explained in detail in Section 3.4. The performance of the procedure is illustrated based on a simulated study in Section 3.5. In Section 3.6 the proposal is applied to establish comparable measurements between two depression instruments applied on the Chilean population. The chapter concludes in Section 3.7 with a discussion and final remarks.

3.2 Description of the data set

Depression indexes in Chile are among the highest in the world. The National Health Survey (2016-2017) applied by the Chilean Health Ministry revealed that 15.8% of the Chilean population older than 18 years have reported depression symptoms within a period of one year. Nevertheless, only 6.2% of the population has been diagnosed with this pathology. Several instruments in the form of self administrated tests have been developed to evaluate symptoms of depression. Two of these instruments are the Beck Depression Inventory (BDI, Beck et al., 1961) and the Outcome Questionnaire (OQ-45.2, Lambert et al., 1996), which have been validated to be used in the Chilean population (see, Valdés et al., 2017; Beck et al., 2002, respectively). The BDI is a 21-item self-report inventory that assesses symptoms of depression underlying one factor. Each item is rated from 0 to 3 according to severity of difficulty experienced. Total scores range from 0 to 63 and are categorised into four levels of severity: minimal depression (total score, 0-13); mild depression (total score, 14-19); moderate depression (total score, 20-28); and severe depression (total score $\geq$ 29; Beck et al., 1996). The OQ-45.2 measures progress in psychological functioning during treatment on three dimensions: subjective discomfort, interpersonal relationships, and social role performance (Lambert et al., 1983). These dimensions are intended to monitor an overall performance of the patient, but are not intended as a diagnostic tool. It is of interest to combine information already available from these two instruments in terms of finding equivalent scores between them. The National Health survey (2016-2017) revealed higher prevalence of depression in women (10.6%) than in men (2.1%). This fact could reflect different features of depression in these two groups and, as a consequence, the relation between the instruments could be different in these groups. The development of equivalent scores will help both researchers and practitioners to use scores interchangeably for describing symptoms and features of depression in the Chilean population. Moreover, if differences in the linked measurements of males and females are found, the monitoring process will be improved.

3.3 Statistical background

A natural way for modelling ordinal variables is to conceive them as representing a discretised version of an underlying latent continuous random variable. In particular, the commonly used ordinal probit model results when a normal distribution is assumed for the latent variable, and if a logistic distribution is considered, the logit model follows (see McCullagh, 1980; McCullagh and Nelder, 1989; Albert and Chib, 1993). In general, if Y is an ordinal random variable defined on the sample space $\mathscr{Y} = \{0, 1, \ldots, C_Y\}$ and complementary information from covariates is available, $\mathbf{x}$, the latent representation establishes that the probability distribution of Y is given by

$$Y \mid Z_Y, \mathbf{x} \quad \sim \quad Mult(1, C_Y + 1, \boldsymbol{p}_{Z_Y,\boldsymbol{\gamma}}(\mathbf{x})) \,, \tag{3.2}$$

where Z_Y is a continuous random variable with distribution function F_{Z_Y}, $Mult(a, b, \boldsymbol{p})$ denotes a multinomial distribution where a is the number of trials, b the number of categories and $\boldsymbol{p}$ is the vector of classification probabilities. In this case, the vector of probabilities is defined as $\boldsymbol{p}_{Z_Y,\boldsymbol{\gamma}}(\mathbf{x}) = (p_{Y,\gamma_0}(\mathbf{x}), \ldots, p_{Y,\gamma_{C_Y}}(\mathbf{x}))$ where

$$p_{Z_Y,\gamma_k}(\mathbf{x}) \quad = \quad F_{Z_Y}(\gamma_{k+1} - \mathbf{x}^t\boldsymbol{\beta}) - F_{Z_Y}(\gamma_k - \mathbf{x}^t\boldsymbol{\beta}) \qquad k = 0, \ldots, C_Y \,, \tag{3.3}$$

where $\boldsymbol{\beta}$ is a vector of parameters, $\mathbf{x}^t\boldsymbol{\beta}$ is the mean of Z_Y and $-\infty = \gamma_0 < \gamma_1 < \ldots < \gamma_{C_Y+1} = \infty$ is a set of thresholds that defines the level of Y. The assumption of normality on the latent variable Z is restrictive, especially for data that containing a large proportion of observations at the extreme levels of the scale, and relatively few observations at moderate levels. As a consequence of the normal distribution shape, there are certain limitations on the effect that each covariate can have on the probability response curve (Boes and Winkelmann, 2006). General surveys of the parametric as well as the semi- and nonparametric literature are given, for example, in Barnhart and Sampson (1994), Clogg and Shihadeh (1994), Winship and Mare (1984), Bellemare et al. (2002), and Stewart (2004).

Bayesian nonparametric models (BNP) have also been proposed to model ordinal data due to the flexibility they provide compared to traditional parametric alternatives. To

model multivariate ordinal data, Kottas et al. (2005) formulated a DP mixture of multivariate normal distributions for the latent distributions. In the absence of covariates, this model is sufficiently flexible to uncover essentially any pattern in a contingency table while using fixed cut-offs. This represents a significant advantage relative to the parametric models mentioned earlier, since the estimation of threshold requires nonstandard inferential techniques, such as hybrid Markov chain Monte Carlo (MCMC) samplers (Johnson and Albert, 1999) and reparameterizations to achieve transformed thresholds that do not have an order restriction (Chen and Dey, 2000).

The most popular Bayesian nonparametric probability model is the Dirichlet Process (DP) (Ferguson, 1973). If G follows a DP prior with precision parameter M and base measure $G_0(\cdot \mid \boldsymbol{\eta})$, denoted by $G \sim DP(M, G_0(\boldsymbol{\eta}))$, the stick-breaking representation of G (Sethuraman, 1994) is given by:

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}(\cdot), \qquad \theta_h \overset{iid}{\sim} G_0(\cdot \mid \boldsymbol{\eta}) \tag{3.4}$$

$$w_h = V_h \prod_{j<h}(1 - V_j) \qquad V_h \mid M \overset{iid}{\sim} Beta(1, M)\ , \tag{3.5}$$

where δ_{θ_h} denotes a point mass function at θ_h and $\boldsymbol{\eta}$ is a vector of hyperparameters that defines the base measure G_0. The dependent Dirichlet Process (DDP) (MacEachern, 1999, 2000) is an approach to define a prior model for the uncountable set of random measurements indexed by covariates, $\mathscr{G} = \{G_{\mathbf{x}},\ \mathbf{x} \in \mathscr{X}\}$, where $\mathscr{X}$ is the space of the covariates. It is said that $\mathscr{G}$ is a varying location DDP (Müller et al., 2015) if, for every $\mathbf{x} \in \mathscr{X}$,

$$G_{\mathbf{x}}(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\theta_h(\mathbf{x})}(\cdot)\ ,$$

where the atoms $\theta_h(\mathbf{x})$ in (3.4) are mutually independent realisations of a stochastic process indexed by $\mathbf{x}$, for $h = 1, 2, \ldots$. This idea has been applied by De Iorio et al. (2004) to define an ANOVA-DDP type model. Similar approaches have been used in spatial modeling (Gelfand et al., 2005), survival analysis (De Iorio et al., 2009), functional data (Dunson and Herring, 2006) and classification (De la Cruz et al., 2007). Dunson et al.

(2007) and Duan et al. (2007) have introduced covariate dependence in the weights of the DP representation (named as the varying weight DDP, Müller et al., 2015). The varying weight and location DDP is obtained when both the weights and the atoms in the DP representation vary across $\mathbf{x}$ (MacEachern, 2000; Griffin and Steel, 2006). Müller et al. (2004) incorporating dependency by means of weighted mixtures of independent random measurements.

Note that the almost sure discreteness of the Dirichlet process (Blackwell and MacQueen, 1973) makes it an inappropriate model for a continuous variable Z. An additional convolution is introduced as a standard procedure for overcoming this difficulty resulting on the DP mixtures models (Antoniak, 1974):

$$h(z) \quad = \quad \int k(z \mid \theta) G(d\theta) \qquad \text{with } \; G \sim DP(M, G_0(\boldsymbol{\eta})) \; ,$$

where for every θ, $k(\cdot \mid \theta)$ is a probability density function and $\theta \in \Theta \subseteq \mathbb{R}^p$. In this context, the DDP can also be used as a mixing distribution in the mixture model such that it is possible to define a covariate-dependent mixture model as

$$h_{\mathbf{x}}(z) \quad = \quad \int k(z \mid \theta) G_{\mathbf{x}}(d\theta) \; , \tag{3.6}$$

where the mixing distribution belongs to the set of dependent probability measurements $\{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, for which it is assumed a $DDP(M, G_0(\boldsymbol{\eta}))$ prior model. Theoretical properties of DDP models and its variations can be found in Lo (1984), Hjort et al. (2010), Ghosal and Van der Vaart (2007), Barrientos et al. (2017) among others.

Motivated by the limitations of parametric and semi-parametric models for the latent variable Z, we extend the DPM model proposed by Kottas et al. (2005) by proposing a covariate-dependent Bayesian nonparametric model for the distribution of the latent variables Z. Because most of the available information in the context of cognitive test are categorical variables, the ANOVA dependent approach of De Iorio et al. (2004) is a natural way to incorporate covariates into the model.

3.4 Proposed method for linking measurements

Let S_i be an ordinal random variable denoting the measure obtained from individual i, for $i = 1, 2, \ldots, n$. Additional information of categorical variables is considered for each sample unit summarised in a $(p-1)$-dimensional vector $\mathbf{x}_i^T$ which includes the categorical variable v_i indicating the instrument used for taking the measure of the individual i. The ordinal assumption of S_i allows to describe its probability distribution in terms of a continuous latent variable Z_i (see Section 3.2). Then, the proposed model is defined as follows:

$$\begin{aligned} S_i \mid Z_i, \mathbf{x}_i &\overset{ind}{\sim} Mult(1, C_{v_i} + 1, \boldsymbol{p}_{Z_i,\gamma}) \\ Z_i \mid \mathbf{x}_i, G_{\mathbf{x}} &\overset{ind}{\sim} H_{\mathbf{x}_i}(z \mid G_{\mathbf{x}}) \, . \end{aligned} \tag{3.7}$$

where the vector of probabilities $\boldsymbol{p}_{Z_i,\gamma}$ is defined as in (3.3). Following ideas of De Iorio et al. (2004), a continuous covariate-dependent model for Z_i of the form (3.6) is proposed which describes dependence across random distributions in an analysis of variance (ANOVA)-type fashion. In particular, it is considered an ANOVA dependent Dirichlet process for the set $\{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, i.e., Z_i has density function:

$$h_{\mathbf{x}}(z \mid G_{\mathbf{x}}) = \int N(z \mid \widetilde{x}_i \boldsymbol{\beta}, \sigma^2) G_{\mathbf{x}}(d\boldsymbol{\theta}) \tag{3.8}$$

where $\widetilde{x}_i = (1, \mathbf{x}_i^T)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2) \in \boldsymbol{\Theta} = \mathbb{R}^p \times \mathbb{R}^+$ and $\{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\} \sim ANOVA - DDP(M, G_0(\boldsymbol{\eta}))$. For identifiability restrictions, it is considered that the effect associated to the first category of each categorical covariate is zero. Note that (3.8) defines a probability model in such a way that marginally, each random measure $h_{\mathbf{x}}(\cdot \mid G_{\mathbf{x}})$ follows a DPM and the dependent DP is used to define the dependence across the related random measures. By introducing latent variables, the mixture can be written as a hierarchical model. In fact,

$$\begin{aligned} z_i \mid \mathbf{x}_i, \boldsymbol{\theta}_i &\overset{ind}{\sim} N(z_i \mid \widetilde{x}_i \boldsymbol{\beta}_i, \sigma_i^2) \\ \boldsymbol{\theta}_i \mid F &\overset{ind}{\sim} F \\ F \mid M, G_0 &\sim DP(M, G_0(\boldsymbol{\eta})) \, , \end{aligned}$$

where $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, \sigma_i^2)$. The base measure $G_0(\cdot \mid \boldsymbol{\eta})$ is considered as

$$G_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}) = N_q(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) IG(\sigma^2 \mid \tau_1/2, \tau_2/2) \ .$$

The Bayesian formulation of the model is completed by specifying priors for the vector of hyperparameters $\boldsymbol{\eta} = (\mu_\beta, \Sigma_\beta, \tau_2)$. By simplicity, we consider conjugated distributions:

$$\begin{aligned} \boldsymbol{\mu}_\beta &\sim N_q(\boldsymbol{\mu}_0, \boldsymbol{S}_0) \\ \boldsymbol{\Sigma}_\beta &\sim IW_q(\nu_0, \boldsymbol{\Psi}_0) \\ \tau_2 &\sim G(\tau_{s_1}/2, \tau_{s_2}/2) \ , \end{aligned}$$

where $IW_q(a, A)$ is a q-dimensional inverse Wishart distribution with a degrees of freedom and a scale matrix A. Additionally, a prior over the precision parameter M of the DP is considered as $M \sim G(a_0, b_0)$. The hyperparameters are fixed as $\mu_0 = 0_q$, $\boldsymbol{S}_0 = \mathbf{I}_q$, $\nu_0 = 5$, $\boldsymbol{\Psi}_0 = \mathbf{I}_q$, $\tau_1 = 6$, $\tau_{s_1} = 6$ and $\tau_{s_2} = 2$, where $\mathbf{I}_p$ represents the identity matrix of order p.

3.4.1 Linking measurements

The product of the number of levels of the categorical covariates, including the number of instruments (categories of the variable v), define all the K possible subgroups found in the population of interest, denoted by $\mathbf{x}_k$, for $k = 1, \ldots, K$. The objective is to obtain comparable measurements between the subgroups defined by $\mathbf{x}_k$ and $\mathbf{x}_q$, where $v_k \neq v_q$, $k, q \in \{1, 2, \ldots, K\}$, i.e., the instruments are different. It is important to highlight at this point that this cannot be done if covariates are omitted. If $\mathscr{M}_k$ is the scale defined by $\mathbf{x}_k$, for $k = 1, 2, \ldots, K$, following ideas of equating methods (see Section 3.1), the proposal defines a method to estimate the function:

$$\begin{aligned} \varphi_{M_q}(\cdot) \ : \ & \mathscr{M}_k \longrightarrow \mathscr{M}_q \\ & s \longrightarrow \varphi_{M_q}(s) \ , \end{aligned}$$

such that that for every $s \in \mathscr{M}_k$, $\varphi_{M_q}(s) \in \mathscr{M}_q$. To this aim, the equipercentile function is estimated in the latent setting after estimating the distribution functions of Z_k and Z_q, the

latent variables associated to the measurements defined by $\mathbf{x}_k$ and $\mathbf{x}_q$, respectively. Note at this point that the ANOVA-DDP modelling of dependence between $G_{\mathbf{x}_k}$ and $G_{\mathbf{x}_q}$ induces a modelling of dependence between the latent cdf's $H_{\mathbf{x}_k}$ and $H_{\mathbf{x}_q}$. Then, it introduces dependence between the measurement's distributions, an issue not considered by traditional equating methods with the exception of Karabatsos and Walker (2009a) who proposed a bivariate Bayesian nonparametric modelling approach for the score distributions.

The proposed method of linking measurements is defined on three steps. Although the method can be applied to any kind of scales, details of the steps are first described for discrete scales.

Step 1: Posterior inferences over latent cdf's

Posterior inferences are made over the collection of covariate-dependent distributions $\{G_{\mathbf{x}} : \mathbf{x} \in \mathscr{X}\}$ after implementing a blocked Gibbs sampler algorithm (Ishwaran and James, 2001). The computational implementation was based on a finite dimensional approximation of the corresponding covariate-dependent stick-braking process assumed for $\{G_{\mathbf{x}} : \mathbf{x} \in \mathscr{X}\}$ in (3.8). As a consequence, samples from the cumulative posterior predictive of Z_l given a new measure $s' \in \mathscr{M}_l$ are obtained, i.e., $\{H^{(t)}_{Z_l}, t = 1, \ldots, T\}$, for $l = k, q$, where T is the posterior sample size of the MCMC algorithm, such that

$$H^{(t)}_{Z_l}(z) \;=\; \sum_{h=1}^{N} w_h^{(t)} \Phi\left(z \mid \widetilde{x}_i \boldsymbol{\beta}_h^{(t)}, \sigma_h^{2(t)}\right) \qquad l = k, q\,.$$

Step 2: Posterior inferences over the latent equipercentile function

From every sample of the posterior distributions in Step 1, $\{(H^{(t)}_{Z_k}, H^{(t)}_{Z_q}), t = 1, \ldots, T\}$, samples of the the posterior distribution for the equipercentile function in the latent setting are obtained as $\{\varphi^{(t)}_{Z_q}(\cdot), t = 1, \ldots, T\}$, according to the definition in (3.1), i.e.,

$$\varphi^{(t)}_{Z_q}(\cdot) = H^{-1(t)}_{Z_q}\left(H^{(t)}_{Z_k}(\cdot)\right)\,, \tag{3.9}$$

where the inverse of the function $H^{(t)}_Z$ is computed numerically for all $t = 1, \ldots, T$.

Step 3: Discrete measurements

Let $s^\star$ denote the score $s \in \mathscr{M}_k$ re-scaled into the support of the latent variable Z_q. The

value $s^\star$ is evaluated on each posterior sample of the latent equipercentile function (3.9). As a consequence, each score $s \in \mathscr{M}_k$ has associated a set of T continuous random *linked measurements*, i.e.,

$$Z_q^\star(s) \quad = \quad \{\varphi_{Z_q}^{(t)}(s^\star), t = 1, \ldots, T\} \, . \tag{3.10}$$

Note that, for all $s \in \mathscr{M}_k$, the set (3.10) is random given the randomness feature of the latent equiperentile functions $\varphi_{Z_q}(\cdot)$. Let $-\infty < \gamma_{0,l} < \gamma_{1,l} < \ldots < \gamma_{C_l,l} < \gamma_{C_{l+1},l} = \infty$ define the set of thresholds for the latent variable Z_l in the representation (3.3) where C_l corresponds to the number of items for the instrument l, for $l = k, q$. Then, the equivalent measure for s will be s_ϵ, for some $s_\epsilon \in \mathscr{M}_q$ if the interval $(\gamma_{\epsilon,q}; \gamma_{\epsilon+1,q}]$ (see Section 3.3) is the one that has the highest probability on the distribution of values (3.10). Mathematically, if $\varphi_{M_q}(s)$ is the equivalent score of s in the scale $\mathscr{M}_q$, then:

$$\varphi_{M_q}(s) = s_\epsilon \quad \Leftrightarrow \quad \epsilon = \underset{\epsilon \in \{0,\ldots,C_q\}}{\operatorname{argmax}} \; P\left(Z_q^\star(s) \in (\gamma_{\epsilon,q}; \gamma_{\epsilon+1,q}]\right) \, . \tag{3.11}$$

Then, for all score on the scale defined by $\mathscr{M}_k$, the proposed method provides as a result an equivalent measurement properly defined on the scale $\mathscr{M}_q$. We emphasise that by *properly* we mean that the linked measurement is not only discrete (or continuous) if $\mathscr{M}_k$ and $\mathscr{M}_q$ are so, but also linked measurements belong to the range defined for the scales.

In order to quantify the variability of the estimation, taking advantages of the random feature of the equipercentile function in the latent setting (see Step 2), we propose to define the standard error of the estimation as the standard deviation of the set (3.10), such that:

$$s.e.\left(\varphi_{M_q}(s)\right) \quad = \quad \sqrt{V(Z_q^\star(s))} \, , \qquad \text{for all } s \in \mathscr{M}_k \, .$$

Even though both the model and the steps of the method have been defined for discrete scales, it is straightforward to define them when continuous measurements are involved in the linking method. In fact, the discrete feature of the measurements is defined by the multinomial distribution in the model (3.7). In case of $\mathscr{M}_k$ and $\mathscr{M}_q$ define continuous scales, the ANOVA-DDP model (3.8) is considered for the measurements' distributions

such that the continuous density of the measurements from each instrument are estimated in Step 1. Then, the latent equipercentile function is estimated as described in Step 2. Finally, the estimated continuous linked measurement is defined as the mode of the set (3.10), for all $s \in \mathscr{M}_k$. The standard error of each estimated measurement is obtained as described in Step 3.

3.5 Simulation study

The main features of the proposed method are illustrated in a simulation study. Each of the steps defined previously are illustrated based on simulations considering only discrete scales for the measurements.

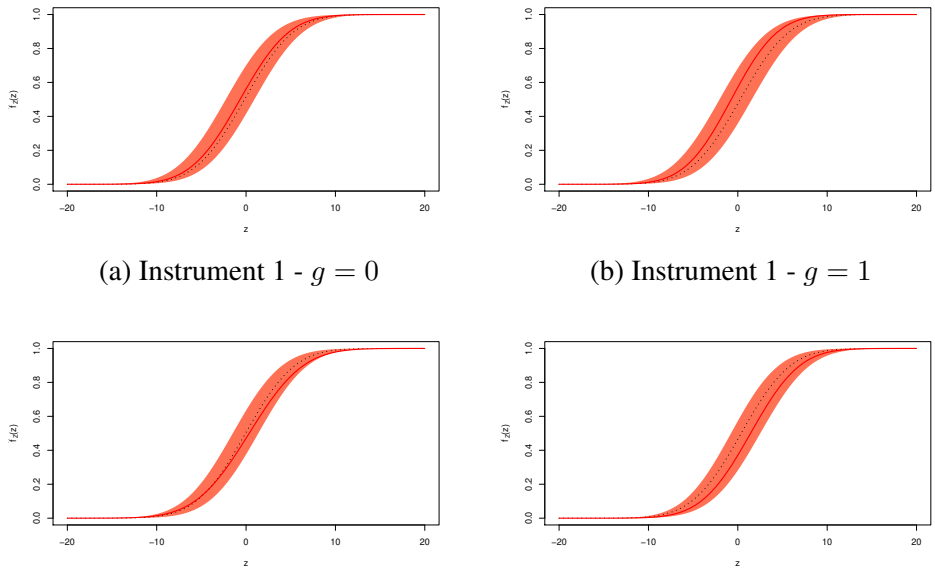
Scores were simulated from the representation of ordinal random variables (3.2) by considering the latent variable Z be a mixture of two normal distributions. An ANOVA structure for relating the probability vector (3.3) with the vector of covariates $\mathbf{x}_i = (v_i, g_i)$ was considered. The covariate v_i represents the instrument used for measuring sample unit i and g_i represents an observable categorical variable. In this simulation study, we consider the gender (male/female) and for simplicity, only two instruments are considered. Then, the total number of groups defined by the combination of levels of the covariates is 4 (2 instruments times 2 gender levels).

Two scenarios are evaluated in the simulation study. In the first one, no differences by instruments and gender are found for the measurement distributions. In the second one, differences in the measurement distributions are due to both test and gender. For each scenario, two samples sizes were considered, $n_1 = 600$ and $n_2 = 2000$ with equal number of observations for each group. Similar proportion of observations for each gender category were considered which were simulated from a discrete uniform distribution. Additionally, the instruments are defined on scales $\mathscr{M}_k$ for $k = 1, 2, 3, 4$, respectively such that all scales are equal to $\mathscr{M} = \{0, 1, \ldots, 9\}$. A number of 100 different datasets were simulated to obtain the results shown in this section. These datasets are used to evaluate the estimation of the latent equipercentile function as well as the discrete linked scores along all the scale

defined by the instruments. The performance of the proposed method is compared with results from discretised versions of equated scores obtained using two traditional equating methods, Gaussian kernel equating (KE, von Davier et al., 2004) and Equipercentile equating (EQ).

In order to exhibit each step of the proposed method, Figure 3.1 shows results of the Step 1 for a randomly selected sample of the 100 datasets generated under Scheme 1 and a sample size $n_1 = 600$. The estimation of the equipercentile function is close to the true function covered by the 95% HPD intervals in all the groups.

To illustrate the second step of the method to link Instrument 1 and Gender 0 to Instrument 2 and Gender 0, the estimation of the latent equipercentile function is shown in Figure 3.2 for the same sample described before. Note that the latent equipercentile



(a) Instrument 1 - $g = 0$ (b) Instrument 1 - $g = 1$

(c) Instrument 2 - $g = 0$ (d) Instrument 1 - $g = 1$

Figure (3.1) Step 1: True (dashed line) latent cumulative distribution function and its estimation (red line) for all groups under sample size $n_1 = 600$ on Scheme 1. The pointwise 95% HPD interval is displayed as the colored area.

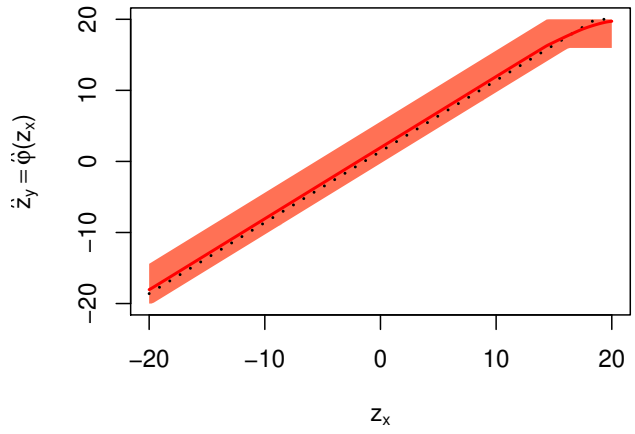
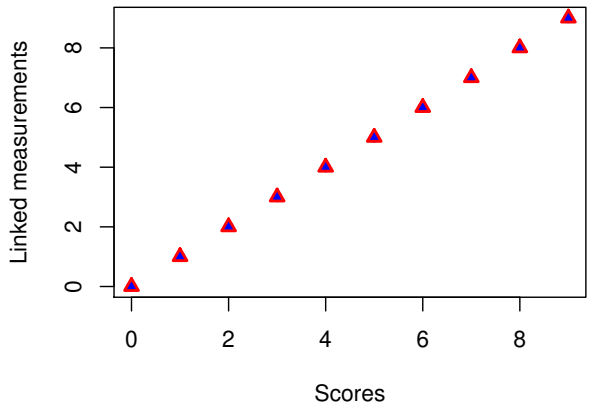


Figure (3.2) Step 2: True (dashed line) equipercentile function and its estimation (red line) for sample size $n_1 = 600$ on Scheme 1 for linking measurements from Instrument 1 and Gender 0 to Instrument 2 and Gender 0. The point-wise 95% HPD interval is displayed as the coloured area.

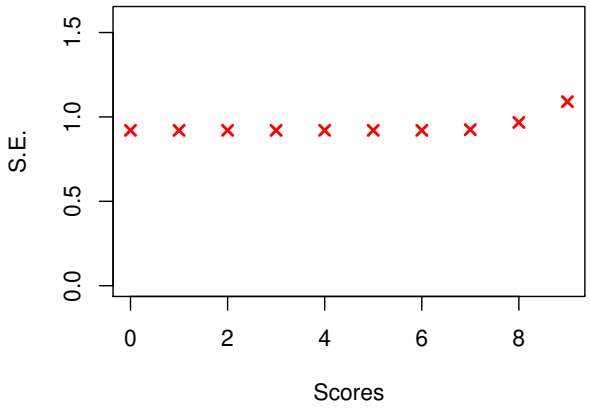
function is covered by the credible intervals along all the range of the latent variable.

In the simulation study, true discrete linked measurements were defined as the evaluation of the true latent equipercentile function on the middle point of the interval $(\gamma_j; \gamma_{j+1}]$ for all $j = 0, \ldots, C$, where C denotes the number of elements in the corresponding scale. The step 3 is illustrated in Figure 3.3 where the discrete linked measurements obtained from the proposal are shown along all the scale. True linked measurements, blue dots, are overlapped with red triangles in most of the cases, showing a good performance of the proposal. In addition, the standard errors are quite similar along all the scale.

The result described so far correspond to a random sample chosen from the 100 datasets simulated under Scheme 1 and sample size $n_1 = 600$. However, results are similar for other schemes. An illustration for other scheme simulations are found in Appendix J. In all cases, the true equipercentile function is covered by credible intervals. The estimated linked measurements are equal to the true discrete measurements in almost all the cases.



(a) I1-$g = 0$ to I2-$g = 0$



(b) I1-$g = 0$ to I2-$g = 0$

Figure (3.3) Step 3: Linked measurements from Instrument 1 and Gender 0 (I1-$g = 0$) to Instrument 2 and Gender 0 (I2-$g = 0$). (a) True (blue dots) discrete linked measurements and its estimation (red triangles) for sample size $n_1 = 600$ on Scheme 1. (b) Standard errors for the linked measurements.

Finally, along all the scale, the standard errors are similar in almost all the schemes.

In order to evaluate the performance of the method, several approaches are considered. To evaluate the accuracy of the estimation of the latent equipercentile function obtained in Step 2, the L_2-norm of the difference between the estimation and the true latent equipercentile functions is evaluated. Results for both sample sizes within both schemes are given in Table 3.1. In all the cases evaluated, as sample size increases the values are lower. Higher values are found in Scheme 2 where there are found differences on the measurement distributions on both gender and the instruments.

The log-pseudo marginal likelihood (LPML, Geisser and Eddy, 1979) is one of the criteria used in the Bayesian framework to compare two models based on their prediction performance. LPML is easy to compute and has been also widely adopted for model selection. Gelfand and Dey (1994) established asymptotic properties of the pseudo marginal likelihood and showed that it can be computed as $LPML = \sum_{i=1}^{n} \log(CPO_i)$ where the CPO_i is the predictive density calculated at the observed measurement s_i given all data except the i-th measurement ($\boldsymbol{s}_{(-i)}$), denoted by $p(s_i \mid \boldsymbol{s}_{(-i)})$. This quantity can be computed easily as the harmonic mean over MCMC scans of the likelihood factor evaluated at

Table (3.1) Simulated data: Estimated L_2-norm of the difference between true continuous equipercentile function and its estimation from the proposed method under two simulation schemes and sample sizes $n_1 = 600$ and $n_2 = 2000$.

		Sample sizes	
Linking	Scheme	n_1	n_2
I1G0-I2G0	Scheme 1	3.03	2.43
	Scheme 2	3.17	1.96
I1G1-I2G1	Scheme 1	3.11	2.55
	Scheme 2	2.97	2.20

imputed likelihood level parameters, i.e.,

$$CPO_i \quad = \quad \frac{1}{T}\sum_{t=1}^{T}\frac{1}{f(s_i \mid \boldsymbol{\theta}^{(t)})} \ ,$$

where, in the proposed model, $f(\cdot \mid \boldsymbol{\theta}^{(t)})$ represents the probability mass function given by the multinomial distribution (3.2) and $\{\boldsymbol{\theta}^{(t)}, t = 1, \ldots, T\}$ is the set of posterior samples of the vector of parameters of the model obtained from the MCMC procedure. Given two competing models, the preferred one is which maximises the LPML. Two model formulations are evaluated in this simulation study. The covariate vector in the first case (Case 1) considers both the test and the gender. In the Case 2, only the test is used in the covariate vector. Table 3.2 shows a summary of the results for the two samples sizes within both schemes. In Scheme 1 both Case 1 and Case 2 show similar values so the performance of the models with and without the gender as a covariate is the same. In contrast, higher values are obtained for Case 2 in Scheme 2. This information allows to conclude that the model taking into account the information of the test and the gender as covariates, is preferred.

Because traditional equating methods do not define a formal sampling model, the computation of the LPML is not possible for traditional equating methods. However, in order to contrast the performance of the proposed method with traditional equating methods, the

Table (3.2) Simulated data: LPML for models of both cases within each scheme and sample sizes ($n_1 = 600$ and $n_2 = 2000$).

		Sample size			
		n_1		n_2	
Linking	Scheme	Case 1	Case 2	Case 1	Case 2
I1G0-I2G0	Scheme 1	10110.08	10123.15	21272.54	22.037.54
	Scheme 2	13231.45	11014.32	28884.54	25994.32
I1G1-I2G1	Scheme 1	11243.22	11455.15	23875.43	23127.94
	Scheme 2	15768.38	10904.72	31034.28	26767.92

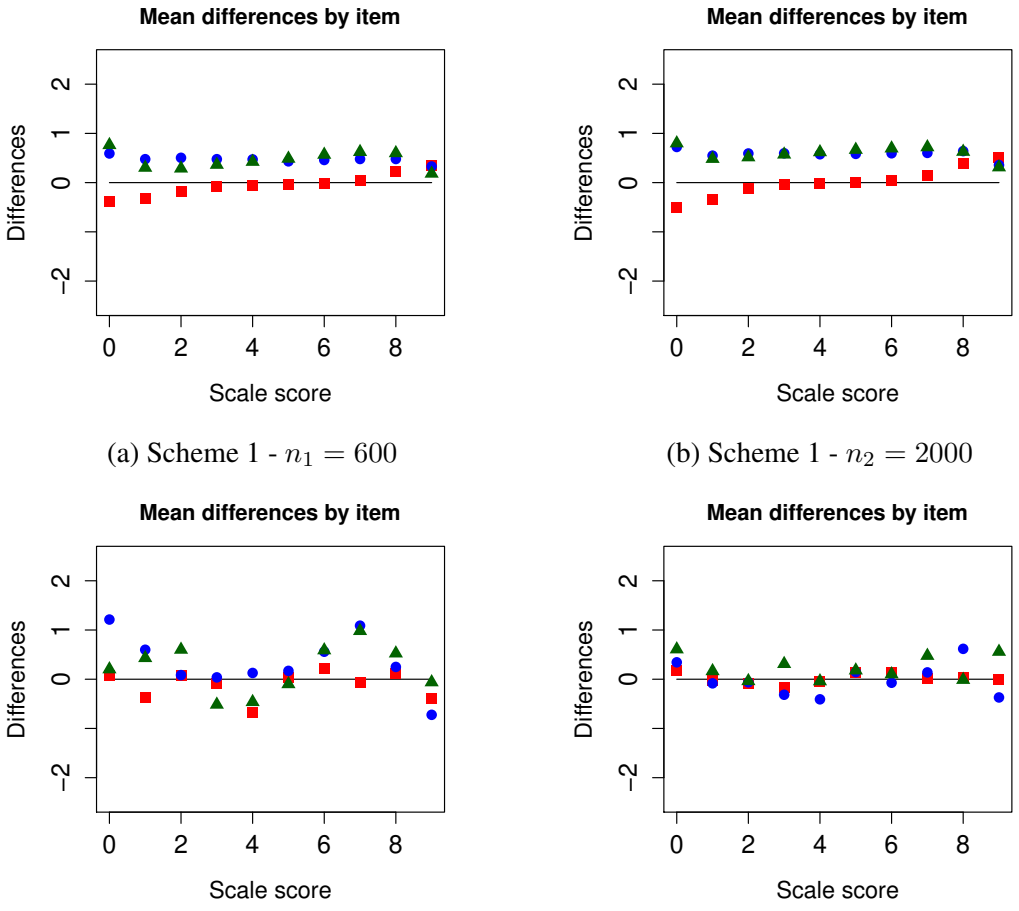
following procedure was carried out following ideas proposed in González et al. (2015b). The nominal feature of the gender covariate allows to integrate it out, by using the sample proportions of each gender category within each instrument, allowing to obtain a linking procedure between the instruments without considering covariates. Then, inferences about the linking procedure are based on the three steps defined in Section 3.4.1 where the posterior samples of the equipercentile function in the latent setting are obtained based on the expression:

$$F_{v_i} = \int F_{\mathbf{x}_i}(\cdot)\hat{F}_n(dg) \,,$$

where $\hat{F}_n(\cdot)$ denotes the empirical distribution of covariate v. In addition, a discrete version of the equated scores from traditional equating methods is considered to fairly compare all the methods. The equated scores are rounded to the largest integer not greater than the corresponding estimated equated score. Results are summarised in Figure 3.4, where the difference between true and estimated measurements are shown on each possible value of the scale for both samples sizes $n_1 = 600$ and $n_2 = 2000$ within each scheme. In Scheme 1, the estimations from the proposed method underestimate the true values in almost all the scale. Discrete versions of traditional equating methods overestimate the true values. In Scheme 2 there are not specific patterns in the estimation of the linked measurements, however our results are better in terms of absolute values along all the scale of the measurements.

3.6 Application

The Beck Depression Inventory (BDI, Beck et al., 1961) and the Outcome Questionnaire (OQ-45.2, Lambert et al., 1996) are two self-report instruments used to evaluate depression symptoms. The BDI is a 21-item self-report inventory that assesses symptoms of depression underlying one factor. Each item is rated from 0 to 3 according to severity of difficulty experienced. The sum of scores (total score) is considered which range from 0 to 63. Depending on the score, the depression is categorised into four levels of severity:



(a) Scheme 1 - $n_1 = 600$

(b) Scheme 1 - $n_2 = 2000$

(c) Scheme 2 - $n_1 = 600$

(d) Scheme 2 - $n_2 = 2000$

Figure (3.4) Simulated data: Mean differences between real discrete linked measurements and its estimations considering the linking method (red squares) and discrete version of Gaussian Kernel equating (blue dots) and Equipercentile equating (green triangles), for sample sizes $n_1 = 600$ and $n_2 = 2000$ within Schemes 1 and 2.

minimal depression (total score, 0-13); mild depression (total score, 14-19); moderate depression (total score, 20-28); severe depression (total score ≥ 29; Beck et al., 1996). The OQ-45.2 measures progress in psychological functioning during treatment on three dimensions: subjective discomfort, interpersonal relationships, and social role performance

(Lambert et al., 1983). These dimensions are intended to monitor an overall performance of the patient, but are not intended as a diagnostic tool. It is of interest to relate information from these two instruments in terms of finding equivalent scores between them.

On one hand, a total of 672 people involved on a suicidal analysis were evaluated with the OQ-45.2 instrument. On the other hand, an independent sample of 701 people who were involved in a research of intercultural-relations within the Chilean population were evaluated with the BDI instrument. The gender of each person evaluated with either of these two instruments is recorded as additional information. A summary about the composition of the sample considered in the analysis is shown in Table 3.3. Similar proportion of males and females were evaluated with the instrument BDI whereas a higher proportion of males where evaluated with the instrument OQ-45.2.

The distribution of the scores using these two instruments is shown in Figure 3.5. It can be seen that there are no much differences in the distribution of the BDI scores between males and females. In contrast, the distribution of the scores is not so similar for people evaluated with the OQ-45.2 instrument, even though the proportion of male and female are not similar.

In order to obtain a link between the scores of these two instruments, the proposed method for linking measurements was applied. We used the model (3.7) where the vector of covariates was defined as $\mathbf{x} = (v, g)$ where v indicates the instrument used to measure depression symptoms and g represents the gender. In all cases, the link function to be estimated considers a relation from BDI scores to OQ-45.2 scores. The hyperparameters

Table (3.3) Frequency of male and female within each group of patients evaluated with the Beck Depression index (BDI) and the Outcome questionnaire (OQ-45.2).

	Gender	
Instrument	Male	Female
BDI	391	310
OQ-45.2	534	138

of the model were fixed as in the simulation study (see Section 3.4.1).

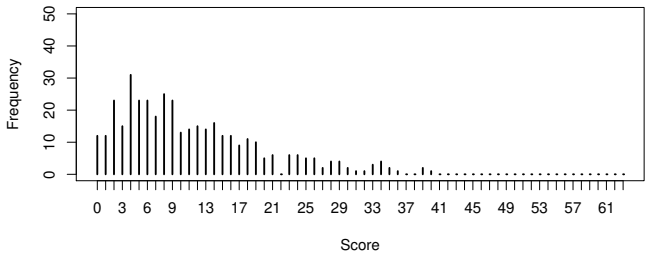
Figure 3.6 shows the estimated latent equipercentile function obtained after applying the proposed method to the depression datasets. In particular, Figure 3.6-(a) shows the result for linking measurements from the group of males (M) evaluated under BDI to the group of males evaluated under OQ-45.2. Figure 3.6-(b) shows the result of linking the group of females (F) evaluated under BDI to the group of females evaluated under OQ-45.2. For both cases, the 95% credibility intervals for the latent equipercentile functions are very thin in almost all the scale of the latent variable, with higher variability at the end of its scale. From our point of view, this result is explained by the low frequency observed at high measurements on both scales, as is illustrated in Figure 3.5.

The discrete linked measurements obtained for linking the groups of males evaluated under BDI to the same group evaluated using the OQ-45.2 scale is displayed in Figure 3.7-(a). The results are equal after linking the group of females evaluated under the BDI to the group of females evaluated under the OQ-45.2, shown in Figure 3.7-(b). The standard errors obtained for each linking procedure are shown in Figure 3.8. The values are very similar for the range $0-40$ BDI scores. For BDI-scores higher than 40, the errors increase in both cases but greater values are found when linking the two scales considering the group of females. We believe that the higher variability in the estimation at the end of the scale is because there is not enough information from the data to reduce the variability in the estimation process.

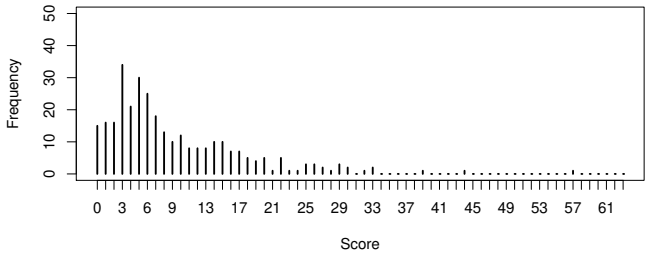
The LPML was computed to compare the fitting of the model considering both the instrument and the gender in contrast to the model that only consider the test as a covariate. The values were 28.187 and 29.031 showing no differences in the fitting of both models, i.e., the gender does not produce a change on the fitted model. Thus, linked measurements from BDI to OQ-45.2. are the same for both females and males (see Figure 3.7). The prior information of higher prevalence of depression in females in the Chilean population does not represent an effect on the linked measurements of depression symptoms, at least using the BDI and OQ-45.2 scales.

3.7 Concluding remarks

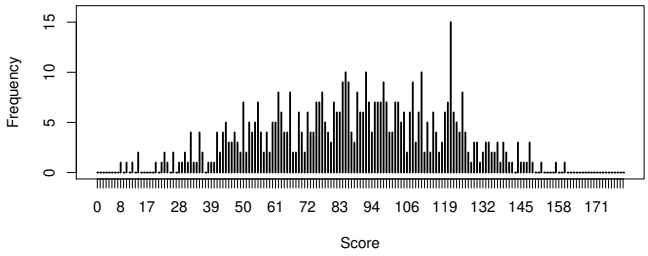
The problem to obtain comparable measurements form different measurement instruments has been described in this chapter. The equipercentile function developed in linking and equating methods has been the basic element in the proposed method. By considering measurements as ordinal random variables, its latent representation has been used to estimate the equipercentile function in the latent setting. To that purpose, a flexible Bayesian nonparametric model was considered which allow to model customised linking functions between measurements defined by different instruments and additional information of covariates. The performance of the proposed method was evaluated by a simulation study showing that it estimates accurately the linking functions between measurements of different subgroups when there are(are not) differences in the link functions due to the covariates. Moreover, it was illustrated, by comparison criterion models, that the covariate-dependent Bayesian model was preferred over the version of the model when information from additional covariates is omitted in the model. The proposed method for linking measurements was applied to a real dataset of depression symptoms in the Chilean population. Results of the linking procedure shown that there is no effect of the gender when linking measurements from the BDI to the OQ-45.2 scales.



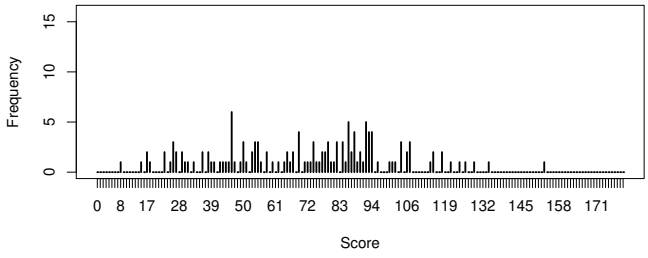
(a) BDI - Male



(b) BDI - Female

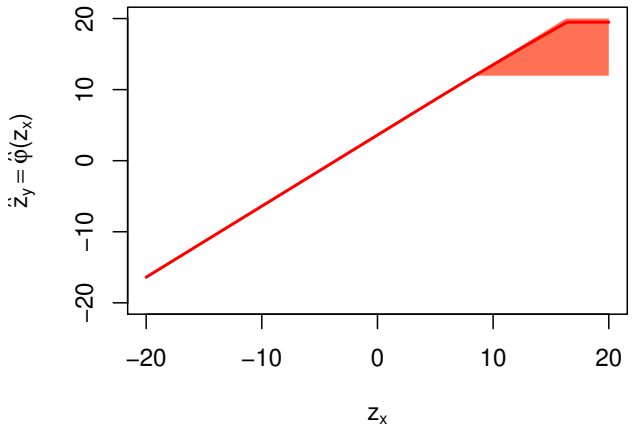


(c) OQ-45.2 - Male

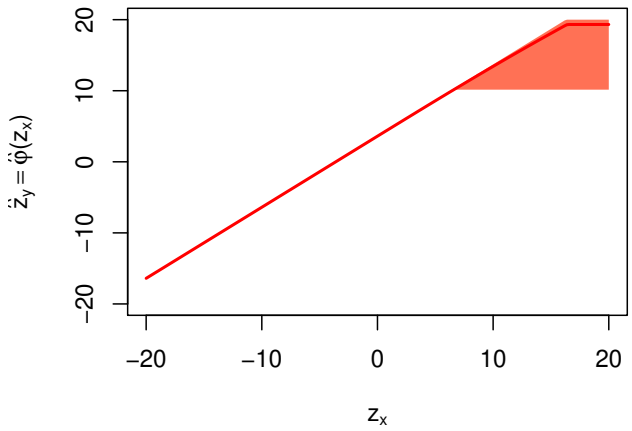


(d) OQ-45.2 - Female

Figure (3.5) Depression instruments: Distribution of the scores for patients evaluated with the BDI and the OQ-45.2 instrument.

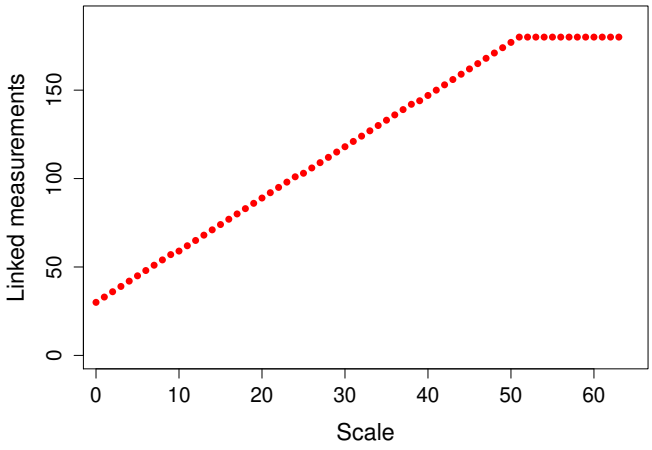


(a) BDI-M to OQ-45.2-M

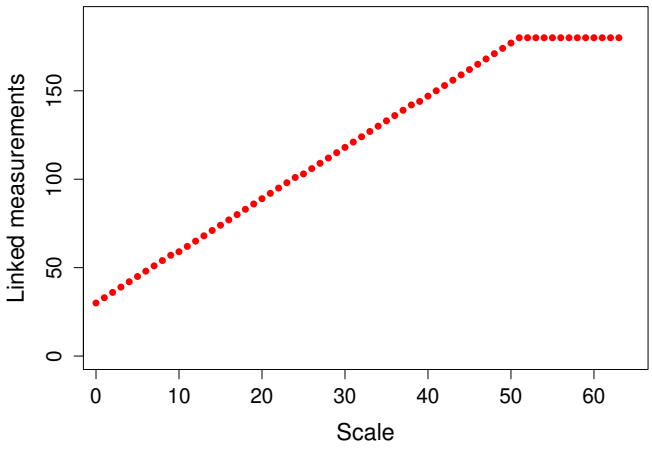


(b) BDI-F to OQ-45.2-F

Figure (3.6) Depression instruments: Estimated latent equipercentile function after linking group males evaluated under BDI to the group of males evaluated under OQ-45.2 (BDI-M to OQ-45.2-M) and the group of female evaluated under BDI to the group of females evaluated under OQ-45.2 (BDI-F to OQ-45.2-F).

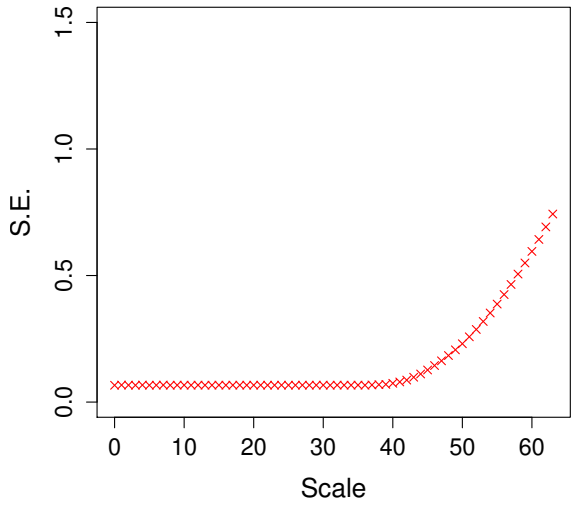


(a) BDI-M to OQ-45.2-M

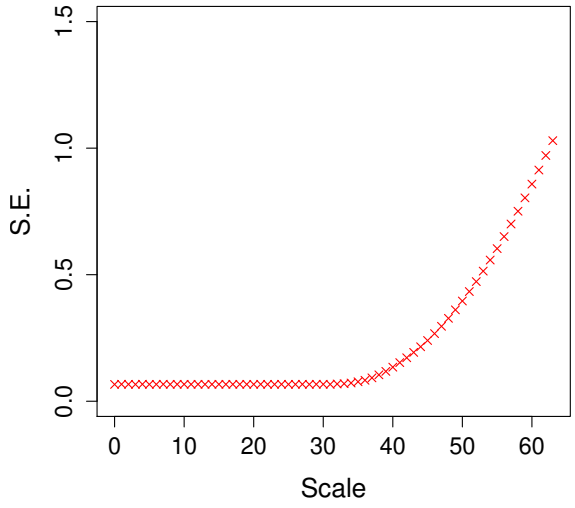


(b) BDI-F to OQ-45.2-F

Figure (3.7) Depression instruments: Discrete linked measurements. (a) Linking the group of males evaluated under BDI to the group of males evaluated under OQ-45.2 (BDI-M to OQ-45.2-M). (b) Linking the group of females evaluated under BDI to the group of females evaluated under OQ-45.2 (BDI-F to OQ-45.2-F)



(a) BDI-M to OQ-45.2-M



(b) BDI-F to OQ-45.2 -F

Figure (3.8) Depression instruments: Standard errors. (a) Linking the group of males evaluated under BDI to the group of males evaluated under OQ-45.2 (BDI-M to OQ-45.2-M). (b) Linking the group of females evaluated under BDI to the group of females evaluated under OQ-45.2 (BDI-F to OQ-45.2-F)

Chapter 4

Conclusions and discussion

"What we know is not much. What we don't know is enormous."
Laplace. Quoted in De Morgan (1866).

The increasing development of new instruments to recover similar or even equal information is a common situation in almost all scientific areas. From an statistical perspective, this fact represents a challenge in the sense that new methods allowing the comparability of several measurements obtained from different instruments need to be developed. There are different statistical approaches to deal with this situation such as those used in measuring agreement analyses, briefly described in the first chapter of this dissertation. However, as it was discussed, when the faced problem is to obtain equivalent measurements from cognitive variables, the adequacy of these approaches is questionable.

In the context of educational measurement, linking and equating methods (Angoff, 1971; Braun and Holland, 1982; Kolen and Brennan, 2014) have been developed to obtain comparable measurements from different forms of a test measuring the same construct. The equipercentile function is defined as a function relating the scales where scores are defined. Different assumptions related to the score distributions result to different parametric, semiparametric and nonparametric estimations of the equipercentile function (González and von Davier, 2013). Since its definition, given by Braun and Holland (1982), the equiper-

centile function was obtained under the assumption of continuous distribution functions for the score distributions. However, because the most used scoring approaches are based on discrete scores, such as the sum of the scores on each item, continuization does not guarantee that equated scores lie into the original scale (for measurements defined on discrete scales), and also that they could be outside from the range defined for the scores. Additionally, even when educational tests forms are build satisfying the requirement of equal construct, traditional equating methods consider scores distributions as independent.

In this dissertation we extended ideas of equating methods to a general approach to link measurements obtained from different measurement instruments. Our interest was to find measurements with equal meaning on the scales defined by the instruments. To that purpose, ideas of equating and linking methods were considered to define "the same meaning" of measurements from different instruments. Our method is based on the fact that measurements define an order relation between them, so that those defined on discrete scales can be considered as ordinal random variables. The latent representation of this kind of variables allows to consider flexible models for the latent variables so that the equipercentile function can be estimated under continuous cdf's. Discrete measurement estimations result by using the one-to-one relation between the latent variable and the ordinal one. Following results of Kottas et al. (2005), a Bayesian nonparametric model is considered for the latent variables, which is flexible enough for modelling ordinal data. All these features of the proposed method tackle some of the disadvantages discussed in the linking and equating literature. In particular, the assumption of ordinal measurements represents an alternative to the continuization step of equating methods to deal with equated scores that do not lie into the scale they are defined. In fact, as it was shown in this dissertation, the latent representation ensures that all linked measurements are properly defined on their corresponding scales.

The ANOVA-type fashion dependent model for the latent variables of the discrete measurements adds an advantage to the modelling process. The dependent relation among the latent variables induces a dependent structure for the scores distributions. This approach agrees with comments in Karabatsos and Walker (2011) that equated tests, measuring the

same construct, should be related.

Even though the model-based approach for linking measurements considered in this dissertation extended and tackled some drawbacks of traditional equating methods, there are some open questions to be discussed. In the context of linking and equating methods, Karabatsos and Walker (2009a) proposed to model the scores distributions considering a bivariate-Bayesian nonparametric model. In addition, González et al. (2015b) proposed a covariate-dependent Bayesian nonparametric model for the score distributions based on Bernstein polynomials. These approaches solve the problem of the permitted range of equated scores, however, both result to continuous equated scores a problem that is tackled by the linking measurement method proposed in this dissertation. Notwithstanding, it could be interesting to evaluate the proposal's performance with respect to these methods.

Several directions can be considered as future work for improving the modelling process. The first attempt is to compare results with an ANOVA-Poisson Dirichlet process for the mixing distribution assumed for the latent variables. Because Poisson-Dirichlet processes allow to define a different structure for the clustering feature, the results obtained for both the latent and the discrete scores could be improved.

To deal with the confounding effect of the difficulty of the tests and the ability of the test takers, as mentioned in Section 1.2, there are several ways to collect the score data and thus to consider the sampling process in the context of equating and linking methods. However, those sampling strategies are not only considered in educational measurement contexts but also in health related areas. For example, in Adroher et al. (2019) the interest is to link measurements of sleep scales, where the instruments have some common questions. Then, the next natural step is to apply the proposed method to different sampling designs such as for instance, for the single group design and the NEAT design. To this aim, we propose to use another prior model for the mixing distribution of the latent variables. Müller et al. (2004) proposed to relate hierarchical models where each model is nonparametric. Let us consider $\boldsymbol{z}_j = \{z_{ij}, i = 1, \ldots, n_j\}$ denotes the vector of latent variables associated to the instrument j, for $j = 1, \ldots, J$ and J is the total number of instruments, such that

$$\boldsymbol{z}_j \sim H_j \,, \qquad H_j \sim p(H_j \mid \boldsymbol{\eta}) \,.$$

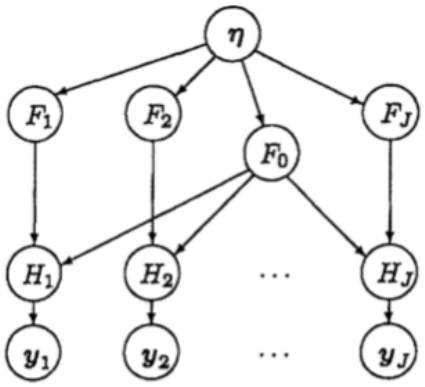
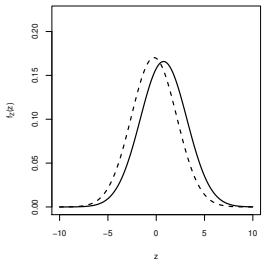


Figure (4.1) Hierarchical model for relating distribution. This picture corresponds to Fig. 2 of Müller et al. (2004)

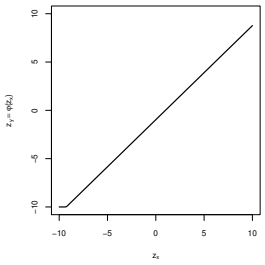
The model considers a prior defined as a mixture between a common measure and a specific measure for each instrument. The idea of the model is summarised in Figure 4.1, where the measure F_0 is shared by all the instruments and the random probability measures F_j characterise the instrument j, i.e., $H_j = \epsilon F_0 + (1-\epsilon)F_j$, where $F_j \sim p(F_j \mid \boldsymbol{\beta})$, for $j = 1, \ldots, J$ and $\epsilon \in [0, 1]$ represents the level of relation between the instruments. In the case of the NEAT design, F_0 could be understood as the information from the anchor test and F_j contains information about the specific instrument j.

Appendices

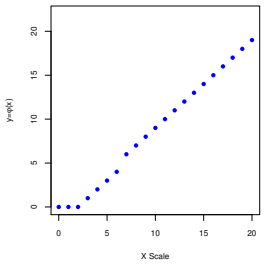
A Simulated Schemes Scenario I



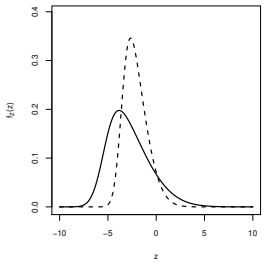
(a) True f_{Z_X} and f_{Z_Y}



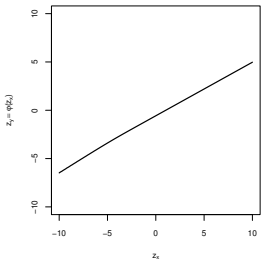
(b) Equipercentile function



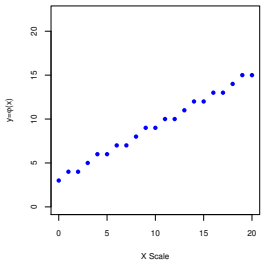
(c) Discrete equated scores



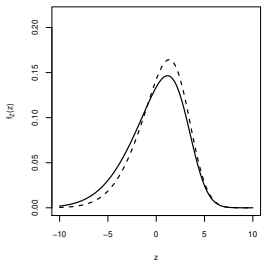
(d) True f_{Z_X} and f_{Z_Y}



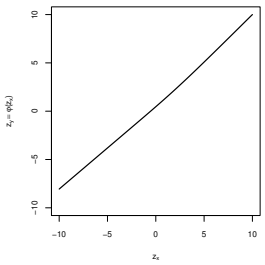
(e) Equipercentile function



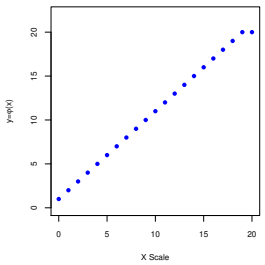
(f) Discrete equated scores



(g) True f_{Z_X} and f_{Z_Y}



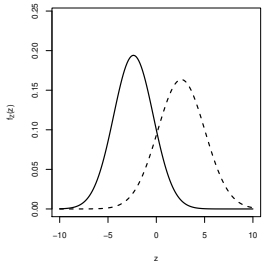
(h) Equipercentile function



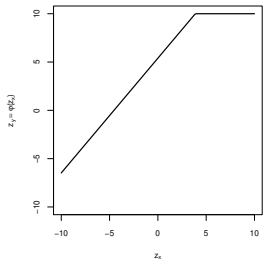
(i) Discrete equated scores

Figure (2) Scenario I: Scheme 1 (Figures (a), (b) and (c)). Scheme 2 (Figures (d), (e) and (f)). Scheme 3 (Figures (g), (h) and (i)). True pdf of Z_X (continuous line) and Z_Y (dashed line).

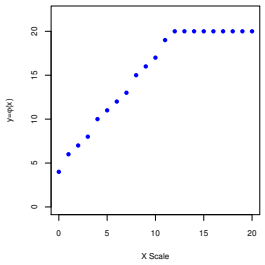
B Simulated Schemes Scenario II



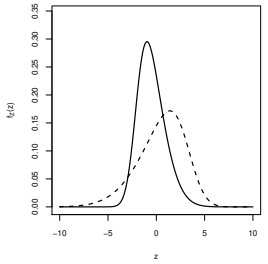
(a) True f_{Z_X} and f_{Z_Y}



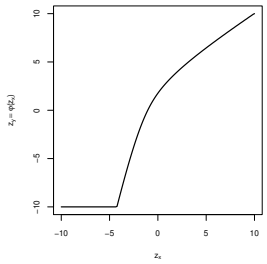
(b) Equipercentile function



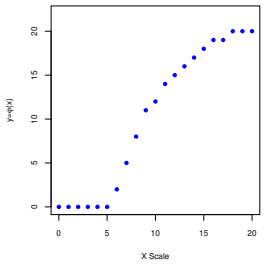
(c) Discrete equated scores



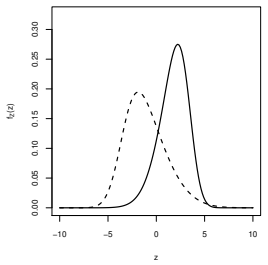
(d) True f_{Z_X} and f_{Z_Y}



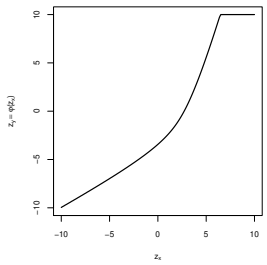
(e) Equipercentile function



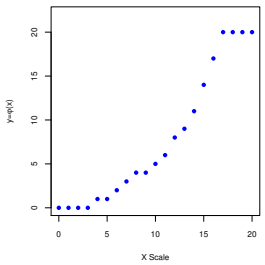
(f) Discrete equated scores



(g) True f_{Z_X} and f_{Z_Y}



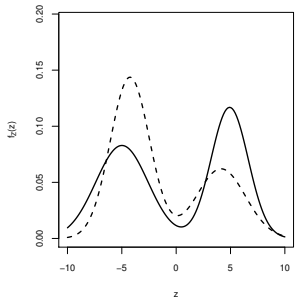
(h) Equipercentile function



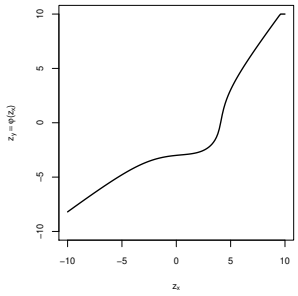
(i) Discrete equated scores

Figure (3) Scenario II: Scheme 4 (Figures (a), (b) and (c)). Scheme 5 (Figures (d), (e) and (f)). Scheme 6 (Figures (g), (h) and (i)). True pdf of Z_X (continuous line) and Z_Y (dashed line).

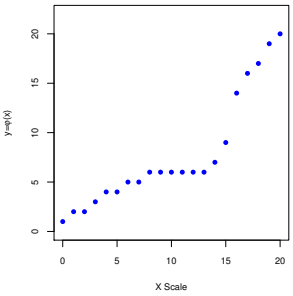
C Simulated Bimodal Latent Distributions



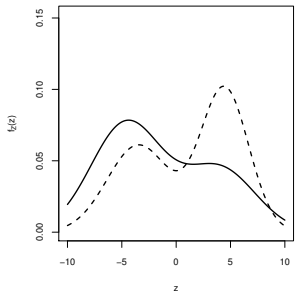
(a) True f_{Z_X} and f_{Z_Y}



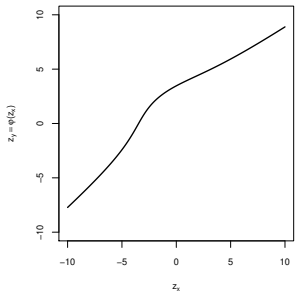
(b) Equipercentile function



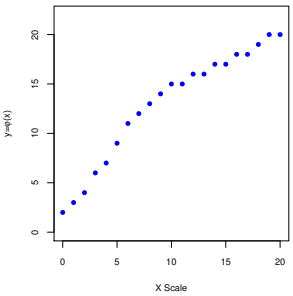
(c) Discrete equated scores



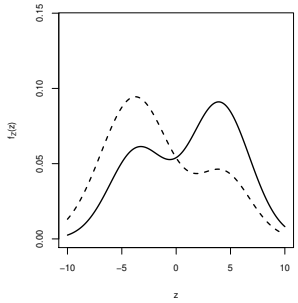
(d) True f_{Z_X} and f_{Z_Y}



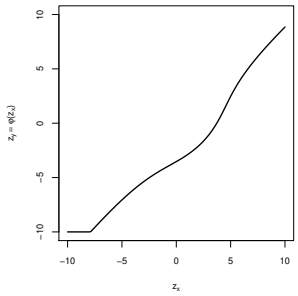
(e) Equipercentile function



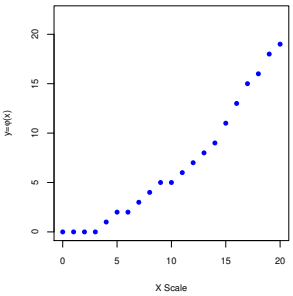
(f) Discrete equated scores



(g) True f_{Z_X} and f_{Z_Y}



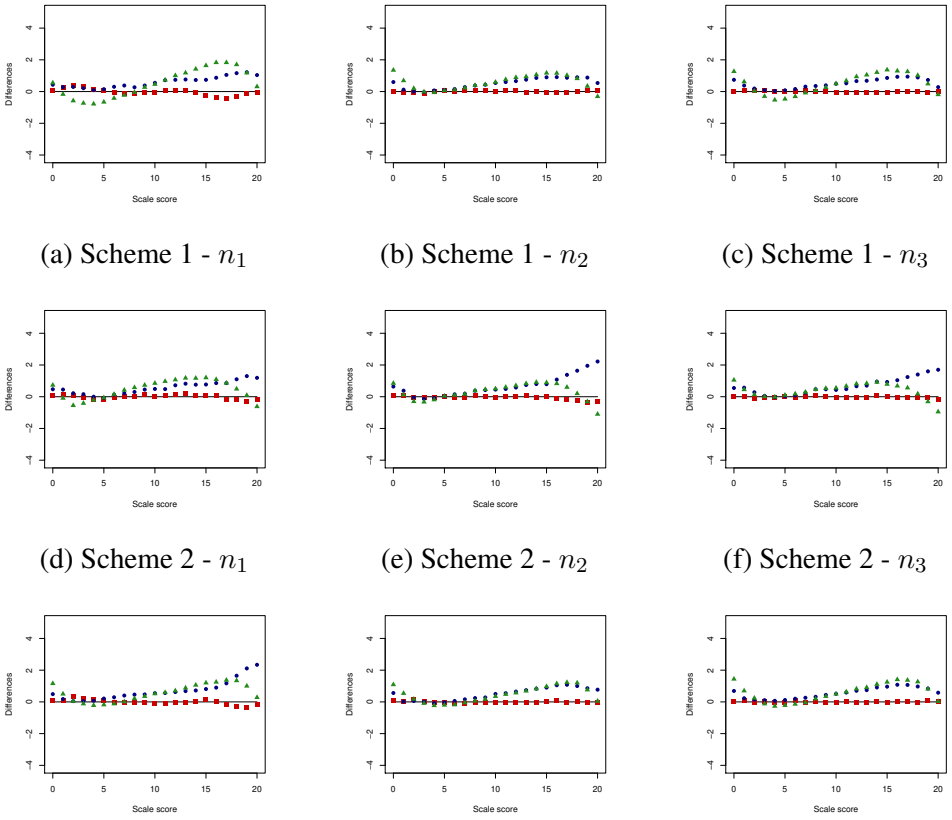
(h) Equipercentile function



(i) Discrete equated scores

Figure (4) Bimodal latent distributions: True pdf of Z_X (continuous line) and Z_Y (dashed line).

D Comparison of discrete equated scores



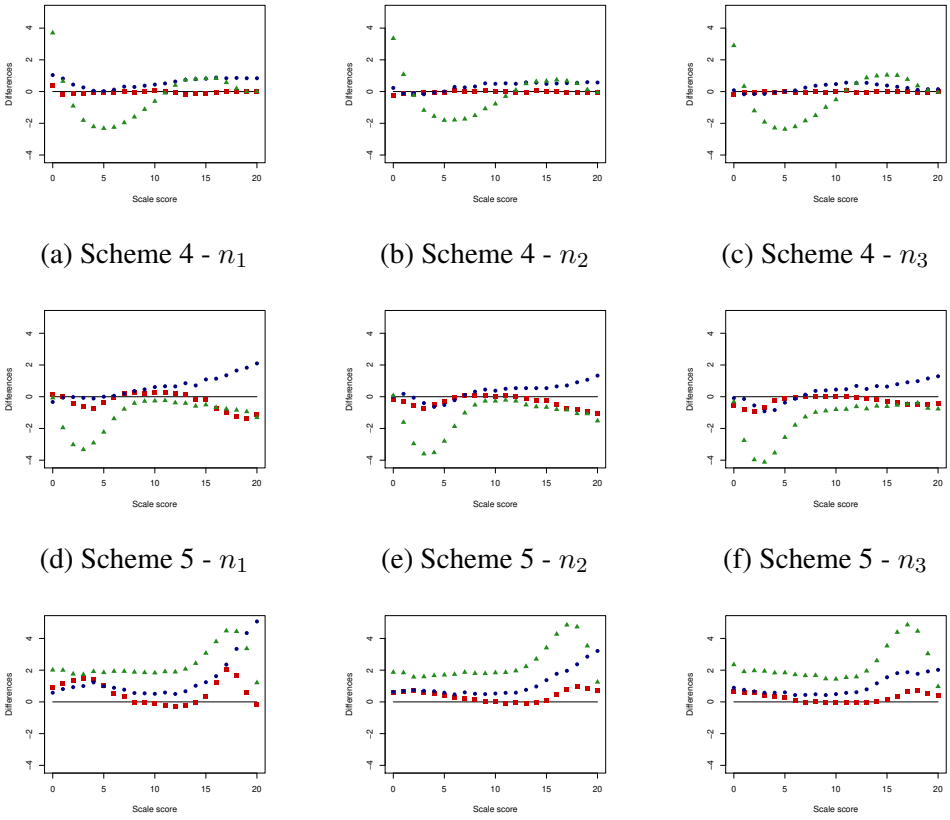
(a) Scheme 1 - n_1 (b) Scheme 1 - n_2 (c) Scheme 1 - n_3

(d) Scheme 2 - n_1 (e) Scheme 2 - n_2 (f) Scheme 2 - n_3

(g) Scheme 3 - n_1 (h) Scheme 3 - n_2 (i) Scheme 3 - n_3

Figure (5) Scenario I: On each possible score scale, the expected value of the difference between true equated scores and estimated discrete equated score for three equating methods: Latent equating (red squares), Equipercentile equating (blue circles) and Gaussian kernel equating (green triangles) for sample sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$, $n_3 = (1500, 1450)$.

E Comparison of discrete equated scores



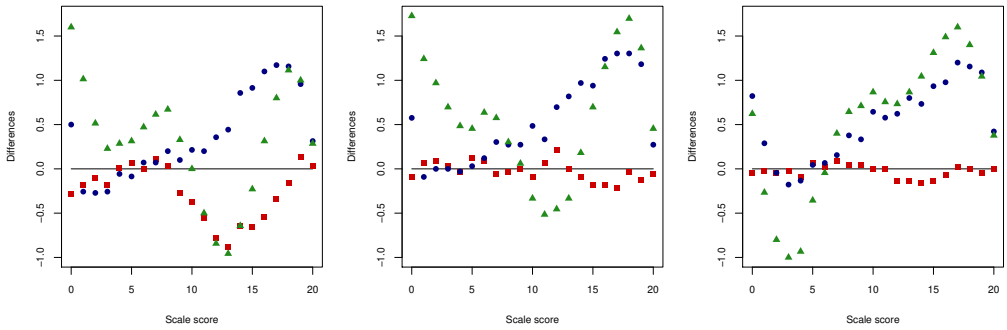
(a) Scheme 4 - n_1 (b) Scheme 4 - n_2 (c) Scheme 4 - n_3

(d) Scheme 5 - n_1 (e) Scheme 5 - n_2 (f) Scheme 5 - n_3

(g) Scheme 6 - n_1 (h) Scheme 6 - n_2 (i) Scheme 6 - n_3

Figure (6) Scenario II: On each possible score scale, the expected value of the difference between true equated scores and estimated discrete equated score for three equating methods: Latent equating (red squares), Equipercentile equating (blue circles) and Gaussian kernel equating (green triangles) for sample sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$, $n_3 = (1500, 1450)$.

F Bimodal latent distributions: comparison of discrete equated scores



(a) Sample size n_1 (b) Sample size - n_2 (c) Sample size - n_3

Figure (7) Bimodal latent distributions: On each possible scale score, the expected value of the difference between true equated scores and estimated discrete equated score for three equating methods: Latent equating (red squares), Equipercentile equating (blue circles) and Gaussian kernel equating (green triangles) for sample sizes (n_X, n_Y): $n_1 = (80, 100)$, $n_2 = (500, 500)$, $n_3 = (1500, 1450)$.

G Posterior computation

Posterior inference for the proposed model is based on the blocked Gibbs sampler algorithm (Ishwaran and James, 2001), where all parameters of the model are updated using the conditional posterior distributions. In what follows we describe the posterior distributions for all the parameters of the model for tests scores from test X but a similar formulation is made for the model of scores from test Y.

Let $\boldsymbol{K}^\star = \{K_1^\star, \ldots, K_m^\star\}$ be the unique values of $\boldsymbol{K} = \{1, \ldots, C_X + 1\}$, $N_j = \{i : K_i = K_j^\star\}$ the number of indexes equal to $K_j^\star$, for $j = 1, \ldots, m$. The algorithm for the estimation, is given by the following steps:

- Updating Z_i: For $i = 1, \ldots, n_X$,

$$Z_i \mid \ldots \sim N_T(\mu_{K_i}, 1/\sigma^2_{K_i}, \gamma_h, \gamma_{h+1})$$

  where $N_T(a, b, c, d)$ denotes a truncated normal distribution with location parameter a, scale parameter b, left truncated value c and right truncated value d. In this case, the values γ_h and γ_{h+1} depend on the value of X_i.

- Updating K_i: $i = 1, \ldots, n_X$,

$$K_i \mid \ldots \overset{ind}{\sim} \sum_{l=1}^{C_X+1} p_{l,i}\delta_l(\cdot)$$

  where

$$(p_{1,i}, \ldots, p_{C_{X+1},i}) \propto (p_1 f(z_i \mid \theta_1), \ldots, p_{C_X+1} f(z_i \mid \theta_{C_{X+1}}))$$

  and $(p_{1,i}, \ldots, p_{C_{X+1},i})$ is updated as follows:

$$\begin{aligned} p_1 &= V_1^\star, \qquad p_k = \prod_{l=1}^{k-1}(1 - V_l^\star)V_k^\star \\ V_k^\star &\sim Beta\left(1 + N_k, M + \sum_{l=k+1}^{C_X+1} N_l\right), \quad V_{C_X+1}^\star = 1, \quad k = 1, \ldots, C_X \end{aligned}$$

- Updating θ: For $k \in \boldsymbol{K} - \boldsymbol{K}^{\star}$,

$$\theta_k \mid \ldots \overset{ind}{\sim} N(\theta_k \mid \lambda, \tau/\sigma_k^2) Gamma(\sigma_k^2 \mid \alpha_0, \beta)$$

and for $j = 1, \ldots, m$

$$\theta_{K_j^{\star}} \mid \ldots \overset{ind}{\sim} N(\mu_{K_j^{\star}} \mid \eta_{K_j^{\star}}^{\star}, \omega_{K_j^{\star}}^{\star}) Gamma(\sigma_{K_j^{\star}}^{2\star} \mid \alpha_{K_j^{\star}}^{\star}, \beta_{K_j^{\star}}^{\star})$$

where

$$\begin{aligned} \eta_{K_j^{\star}}^{\star} &= \left(N_j + \frac{1}{\tau}\right)^{-1} \left(\frac{\lambda}{\tau} + N_j \bar{z}_j\right), \qquad \omega_{K_j^{\star}}^{\star} = \sigma_{K_j^{\star}}^{2\star} \left(N_j + \frac{1}{\tau}\right)^{-1}, \\ \alpha_{K_j^{\star}}^{\star} &= \frac{N_j + 1}{2} + \alpha_0, \qquad \beta_{K_j^{\star}}^{\star} = \beta + \frac{(N_j - 1)s_j^2}{2} + \frac{N_j}{N_j + 1} \frac{(\bar{z}_j - \lambda)^2}{2} \end{aligned}$$

and $\bar{z}_j$ and s_j^2 are the mean and the sample variance of the set $\{z_i : \ K_i = K_j^{\star}\}$, respectively.

- Updating τ:

$$\tau \mid \ldots \ \sim \ IG\left(\tau \mid w_0 + \frac{C_X + 1}{2}, W_0 + \frac{1}{2} \sum_{j=1}^{C_X+1} \sigma_j^2 (\mu_j - \lambda)^2\right)$$

- Updating $\phi = (\lambda, \beta, M)$:

$$\begin{aligned} \lambda \mid \ldots &\sim N\left(\lambda \mid \left(\sum_{j=1}^{C_X+1} \frac{\sigma_j^2}{\tau} + \frac{1}{Q_0}\right)^{-1} \left(\frac{q_0}{Q_0} + \sum_{j=1}^{C_X+1} \frac{\mu_j \sigma_j^2}{\tau}\right), \left(\sum_{j=1}^{C_X+1} \frac{\sigma_j^2}{\tau} + \frac{1}{Q_0}\right)^{-1}\right) \\ \beta \mid \ldots &\sim Gamma\left(\beta \mid c_0 + (C_X + 1)\alpha_0, C_0 + \sum_{j=1}^{C_x+1} \sigma_j^2\right) \\ M \mid \ldots &\sim Gamma\left(M \mid C_X + a_0, b_0 - \sum_{j=1}^{C_X} log(1 - p_j)\right) \end{aligned}$$

We use the posterior predictive distribution to compute the probability of a new unobserved value of Z_X given the observed sample value δ_h. This distribution is calculated by integrating the density of a new observation over the posterior distribution of the parameters that define the model. As the model involves probabilities associated with the cdf of Z_X, the cumulative predictive distribution function is given by:

$$F_{Z_X}(z) \ = \ \int_{-\infty}^{z} \sum_{l=1}^{C_X+1} p_l N(s \mid \mu_l, 1/\sigma_l^2) ds \ .$$

H Evaluation of estimated latent equipercentile functions

An important step of the proposed latent equating method consists of estimating the equipercentile function in the latent setting. In section 2.2 was described how the estimation is obtained.

The criterion used to evaluate the results in the simulation study is the expected value of the L_2 norm between the estimation and the real equipercentile function, with respect to the sampling distribution.

The expected L_2 distance is obtained as

$$\mathbb{E}[\|\varphi_0 - \hat{\varphi}\|_2] \approx \frac{1}{100}\sum_{i=1}^{100}\|\varphi_0 - \hat{\varphi}_{(i)}\|_2 ,$$

where $\mathbb{E}$ denotes the expected value with respect to the sampling distribution, $\|\cdot\|_2$ is the L_2 norm, φ_0 is the true equating function, $\hat{\varphi}$ is the estimator obtained using the proposed method and $\hat{\varphi}_{(i)}$ is the estimation of φ using $\hat{\varphi}$ at the $i-$th replicate. The estimator associated with the nonparametric procedure correspond to

$$\hat{\varphi} = \mathbb{E}[F_{Z_Y}^{-1}(F_{Z_X}(\cdot)) \mid \text{data}] ,$$

where $\mathbb{E}[\cdot \mid \text{data}]$ denotes the posterior mean and, F_{Z_Y} and F_{Z_X} are the posterior predictive cdfs of the DPM proposed model for each latent variable. This expectation was approximated by using the 100 replicates and the Monte Carlo method.

I Evaluation of estimated discrete equated scores

The final step of the proposal is to obtain equated scores that belong in the original scale score of the test as the result of applying the strategy defined in see section 2.2.

The performance of the proposed equating method at this step in the simulation study was developed by considering the statistic Ψ_2 which we define as follows. Let us consider W_0 the vector with true discrete equated values and $\hat{W}$ the vector of estimated discrete equated scores in the whole scale under the proposed model. Then,

$$\Psi_2 \quad = \quad \mathbb{E}[\|W_0 - \hat{W}\|_2] \, ,$$

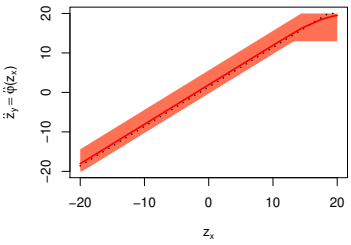
where $\mathbb{E}$ denotes the expected value with respect to the sampling distribution, $\|\cdot\|_2$ is the L_2 norm. This quantity was approximated throughout the MCMC method and the 100 replicates generated for each scheme, such that,

$$\mathbb{E}[\|W_0 - \hat{W}\|_2] \quad \approx \quad \frac{1}{100}\sum_{i=1}^{100} \|W_0 - \hat{W}_{(i)}\|_2 \, ,$$

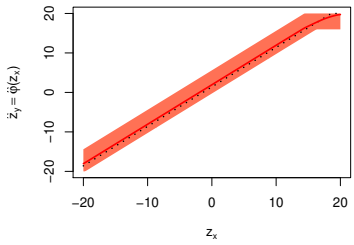
where $\hat{W}_{(i)}$ is the estimation of W_0 at the $i-$th replicate.

J Illustration Linking discrete measurements

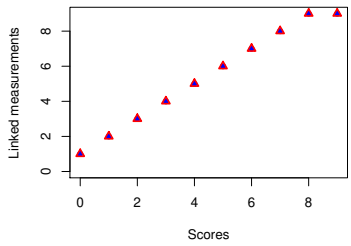
To illustrate results of the proposed method, a random sample from the 100 datasets simulated for each scheme and sample sizes was chosen. The estimation of the latent equipercentile function as well as the discrete linked measurements estimated by the method are in Figures 8 and 9 for the Scheme 1. For the Scheme 2, the corresponding results are shown in Figures 10 and 11. In both schemes results are similar, where the latent equipercentile function is covered by the 95% HDP intervals. The discrete estimated linked measurements are well estimated from the method in almost all the range of the scales. Additionally, the standard errors are similar along the scale for all the illustration with higher values for the last measurements in the scale. Note that in Scheme 1, an increase in the sample size, the standard errors are reduce almost a half than for small sample size.



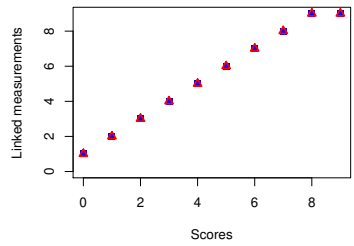
(a) Scheme 1 - $G = 0$



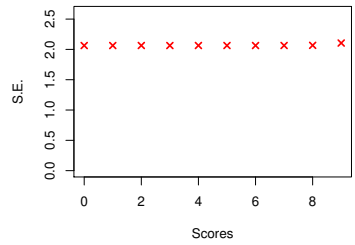
(b) Scheme 1 - $G = 1$



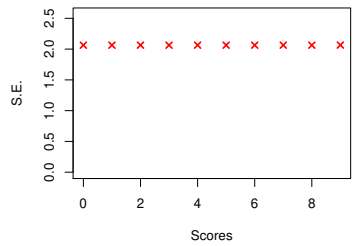
(c) Scheme 1 - $G = 0$



(d) Scheme 1 - $G = 1$

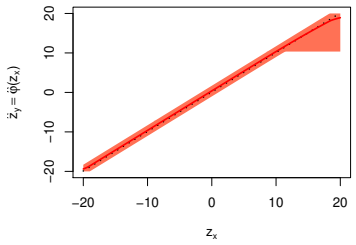


(e) Scheme 1 - $G = 0$

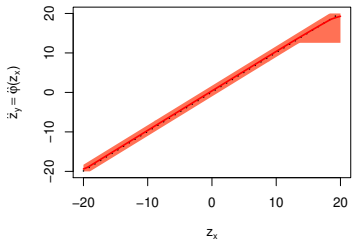


(f) Scheme 1 - $G = 1$

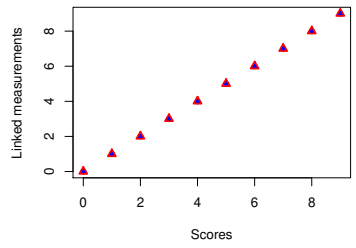
Figure (8) Linking Instrument 1 to Instrument 2: (a)-(b) True (dashed line) equipercentile function and its estimation (red line) for sample size $n_1 = 600$ on Scheme 1. The point-wise 95% HPD interval is displayed as the coloured area. (c)-(d) Estimated linked measurements. True linked measures (blue dot) and estimated linked measurements (red triangles). (e)-(f) Standard errors for estimated linked measurements along the scale of the instrument.



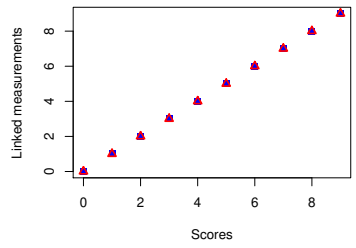
(a) Scheme 1 - $G = 0$



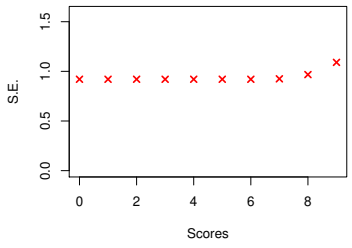
(b) Scheme 1 - $G = 1$



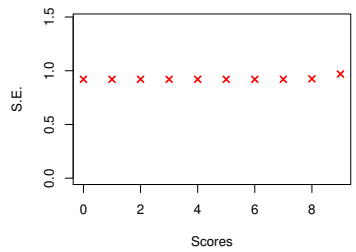
(c) Scheme 1 - $G = 0$



(d) Scheme 1 - $G = 1$

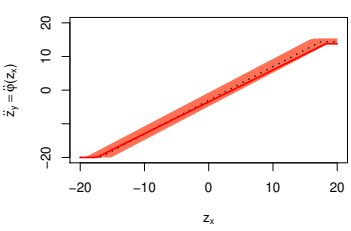


(e) Scheme 1 - $G = 0$

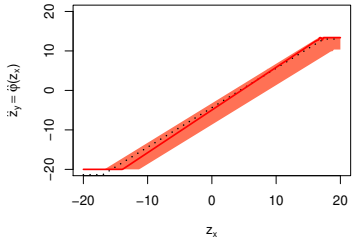


(f) Scheme 1 - $G = 1$

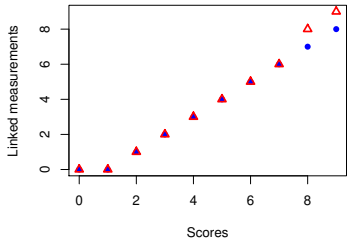
Figure (9) Linking Instrument 1 to Instrument 2: (a)-(b) True (dashed line) equipercentile function and its estimation (red line) for sample size $n_2 = 2000$ on Scheme 1. The point-wise 95% HPD interval is displayed as the coloured area. (c)-(d) Estimated linked measurements. True linked measures (blue dot) and estimated linked measurements (red triangles). (e)-(f) Standard errors for estimated linked measurements along the scale of the instrument.



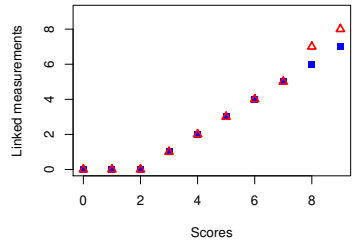
(a) Scheme 2 - $G = 0$



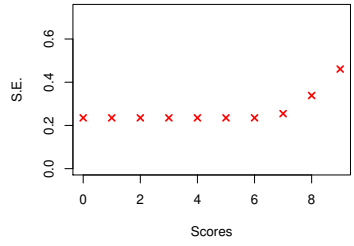
(b) Scheme 2 - $G = 1$



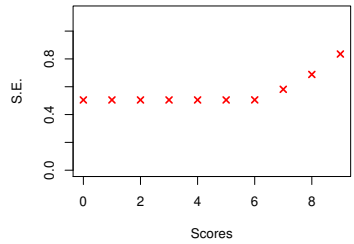
(c) Scheme 2 - $G = 0$



(d) Scheme 2 - $G = 1$

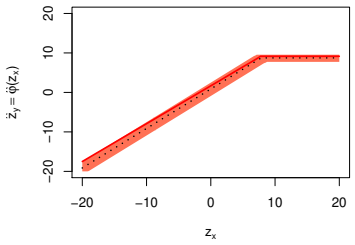


(e) Scheme 2 - $G = 0$

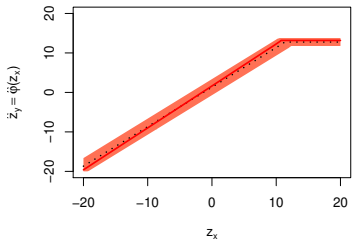


(f) Scheme 2 - $G = 1$

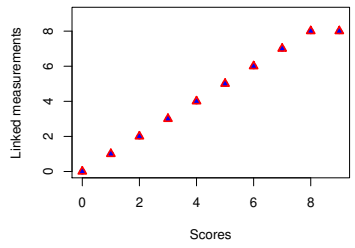
Figure (10) Linking Instrument 1 to Instrument 2: (a)-(b) True (dashed line) equipercentile function and its estimation (red line) for sample size $n_1 = 600$ on Scheme 2. The point-wise 95% HPD interval is displayed as the coloured area. (c)-(d) Estimated linked measurements. True linked measures (blue dot) and estimated linked measurements (red triangles). (e)-(f) Standard errors for estimated linked measurements along the scale of the instrument.



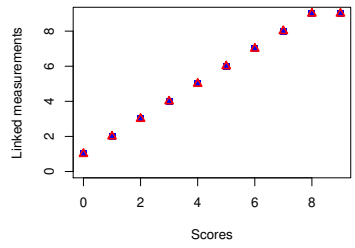
(a) Scheme 2 - $G = 0$



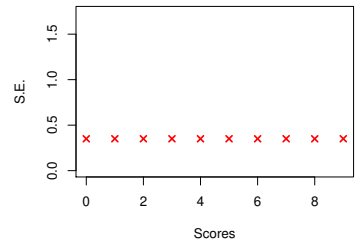
(b) Scheme 2 - $G = 1$



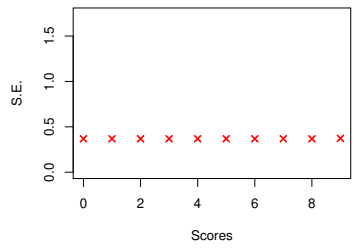
(c) Scheme 2 - $G = 0$



(d) Scheme 2 - $G = 1$



(e) Scheme 2 - $G = 0$



(f) Scheme 2 - $G = 1$

Figure (11) Linking Instrument 1 to Instrument 2: (a)-(b) True (dashed line) equipercentile function and its estimation (red line) for sample size $n_2 = 2000$ on Scheme 2. The point-wise 95% HPD interval is displayed as the coloured area. (c)-(d) Estimated linked measurements. True linked measures (blue dot) and estimated linked measurements (red triangles). (e)-(f) Standard errors for estimated linked measurements along the scale of the instrument.

Bibliography

Adroher, N. D., S. Kreiner, C. Young, R. Mills, and A. Tennant (2019). Test equating sleep scales: applying the Leunbach's model. *BMC Medical Research Methodology 19*(1), 141–153.

Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*(422), 669–679.

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. (Reprinted as Angoff WH (1984). *Scales, Norms and Equivalent Scores*. Princeton, NJ: Educational Testing Service.).

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist. 2*(6), 1152–1174.

Barnhart, H. X. and A. R. Sampson (1994). Overview of multinomial models for ordinal data. *Communications in Statistics - Theory and Methods 23*(12), 3395–3416.

Barrientos, A. F., A. Jara, and F. A. Quintana (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis 7*, 277–310.

Barrientos, A. F., A. Jara, and F. A. Quintana (2017). Fully nonparametric regression for bounded data using dependent Bernstein polynomials. *Journal of the American Statistical Association 112*(518), 806–825.

Beck, A., R. Steer, and G. Brown (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Beck, A. T., C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh (1961). An Inventory for Measuring Depression. *JAMA Psychiatry 4*(6), 561–571.

Beck, A. T., C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh (2002). OQ-45.2, Cuestionario para la evaluación de resultados y evolución en psicoterapia: Adaptación, validación e indicaciones para su aplicación. *Terapia Psicológica 20*(2), 161–176.

Bellemare, C., B. Melenberg, and A. van Soest (2002). Semi-parametric models for satisfaction with income. Workingpaper, Econometrics. Pagination: 39.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics 1*(2), 353–355.

Boes, S. and R. Winkelmann (2006). Ordered response models. *Advances in Statistical Analysis 90*, 165–179.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association 71*(356), 791–799.

Braun, H. and P. Holland (1982). Observed-score test equating: a mathematical analysis of some ets equating procedures. In P. Holland and D. Rubin (Eds.), *Test equating*, Volume 1, pp. 9–49. New York: Academic Press.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*(4), 434–455.

Casaletto, K. and R. Heaton (2017). A review of screening tests for cognitive impairment. *Journal of the International Neuropsychological Society 23*, 778–790.

Chen, M.-H. and D. K. Dey (2000). Bayesian analysis for correlated ordinal data models. In D. K. Dey, S. Ghosh, and B. K. Mallick (Eds.), *Genelized Linear Models: A Bayesian perspective*, pp. 135–162. New York: Marcel Dekker.

Choudhary, P. K. and H. N. Nagaraja (2017). *Measurment Agreement: Models, methods and applications* (1st ed.). Hoboken, NJ: Wiley.

Clogg, C. C. and E. S. Shihadeh (1994). *Statistical models for Ordinal variables* (1st ed.). Thousand Oaks.

Cullen, B., B. O'Neill, J. Evans., R. Coen, and B. Lawlor (2007). A review of screening tests for cognitive impairment. *Journal of Neurology, Neurosurgery, and Psychiatry 78*(8), 790–799.

Daniels, M. J. and A. R. Linero (2015). *Bayesian Nonparametrics for Missing Data in Longitudinal Clinical Trials*, pp. 423–446. Cham: Springer International Publishing.

De Iorio, M., W. O. Johnson, P. Müller, and G. L. Rosner (2009). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Biometrics 65*(3), 762–771.

De Iorio, M., P. Müller, G. Rosner, and S. MacEachern (2004). An ANOVA model for dependent random measures. *Journal of American Statistical Association 99*, 205–215.

De la Cruz, R., A. A. Quintana, and P. Müller (2007). Semiparametric Bayesian classification with longitudinal markers. *Applied Statistics 56*(2), 119–137.

De Morgan, A. (1866). *A Budget of Paradoxes*. London: Longmans, Green.

Dorans, N. and P. Holland (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement 37*(4), 281–306.

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality Life Research 16*, 85–94.

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

Duan, J. A., M. Guindani, and A. E. Gelfand (2007). Generalized spatial Dirichlet process models. *Biometrika 94*(4), 809–825.

Dunson, D. and A. H. Herring (2006). Semiparametric Bayesian latent trajectory models. Technical report, ISDS Discussion paper 16, Duke University, NC, USA.

Dunson, D. B., N. Pillai, and J.-H. Park (2007). Bayesian density regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 69*(2), 163–183.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics 1*, 209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In D. Siegmund, J. Rustage, and G. G. Rizvi (Eds.), *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, pp. 287–302. Bibliohound.

Fried, E. I. (2016). Are more responsive depression scales really superior depression scales? *Journal of Clinical Epidemiology 77*, 4–6.

Geisser, S. and W. F. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association 74*(365), 153–160.

Gelfand, A. E. and D. K. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(3), 501–514.

Gelfand, A. E., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association 100*(471), 1021–1035.

Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science 7*, 457–511.

Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In N. Hjort, P. Holmes, C.and Müller, and S. G. Walker (Eds.), *Bayesian Nonparametrics*, pp. 22–34. Cambridge University Press.

Ghosal, S. and A. W. Van der Vaart (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics 35*, 697–723.

Ghosh, S. K., C. B. Burns, D. L. Prager, L. Zhang, and G. Hui (2018). On nonparametric estimation of the latent distribution for ordinal data. *Computational Statistics and Data Analysis 119*, 86–98.

González, J. (2014). SNSequate: Standard and nonstandard statistical models and methods for test equating. *Journal of Statistical Software 59*(7), 1–30.

González, J., A. F. Barrientos, and F. A. Quintana (2015a). A Dependent Bayesian Nonparametric Model for Test Equating. In R. Millsap, D. Bolt, L. van der Ark, and W.-C. Wang (Eds.), *Quantitative Psychology Research*, pp. 213–226. Springer International Publishing.

González, J., A. F. Barrientos, and F. A. Quintana (2015b). Bayesian nonparametric estimation of test equating functions with covariates. *Computational Statistics & Data Analysis 89*, 222–244.

González, J. and M. von Davier (2013). Statistical models and inference for the true equating transformation in the context of local equating. *Journal of Educational Measurement 50*(3), 315–320.

González, J. and M. Wiberg (2017). *Applying Test Equating Methods Using R*. Springer.

Griffin, J. E. and M. F. J. Steel (2006). Order-based Dependent Dirichlet Processes. *Journal of the American Statistical Association* (101), 179–194.

Haber, M. (1985). Maximum likelihood methods for linear and log-linear models in categorical data. *Computational Statistics and Data Analysis 3*, 1–10.

Hanson, T. and W. Johnson (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association 97*(460), 1020–1033.

Hjort, N. L., C. Holmes, P. Müller, and S. Walker (2010). *Bayesian nonparametrics*. Cambridge, UK: Cambridge University Press.

Holland, P. and D. Rubin (1982). *Test equating*. New York: Academic Press.

Holland, P. and D. Thayer (1989). The kernel method of equating score distributions. Technical report, Princeton, NJ: Educational Testing Service.

Holland, P. W. and N. J. Dorans (2006). *Linking and equating*, pp. 187–220. Westport, CT:: Praeger.

Holland, P. W. and M. Hoskens (2003, Mar). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika 68*(1), 123–149.

Inácio de Carvalho, V., A. Jara, and M. de Carvalho (2015). *Bayesian Nonparametric Approaches for ROC Curve Inference*, pp. 327–344. Cham: Springer International Publishing.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), 161–173.

Johnson, V. and J. Albert (1999). *Ordinal data modeling* (1st ed.). New York: Springer.

Karabatsos, G. and S. Walker (2009a). A Bayesian nonparametric approach to test equating. *Psychometrika 74*(2), 211–232.

Karabatsos, G. and S. Walker (2009b). Coherent psychometric modelling with Bayesian nonparametrics. *British Journal of Mathematical and Statistical Psychology 62*(1), 1–20.

Karabatsos, G. and S. Walker (2011). A Bayesian Nonparametric Model for Test Equating. In A. von Davier (Ed.), *Statistical Models for Test Equating, Scaling, and Linking*, Volume 1, pp. 175–184. New York: Springer.

Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998). Markov Chain Monte Carlo in practice: A roundtable discussion. *The American Statistician 52*(2), 93–100.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: World Book Co.

Kolen, M. and R. Brennan (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

Kottas, A., P. Muller, and F. Quintana (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphics Statistics 14 14*(3), 610–625.

Lambert, M., E. Christensen, and S. DeJulio (1983). *The Assessment of psychotherapy outcome*. New York, NY: John Wiley and Sons.

Lambert, M. J., N. B. Hansen, V. Umphress, and O. J. B. G. M. . R. C. W. Lunnen, K. (1996). *Administration and scoring manual for the Outcome Question- naire (OQ–45.2)*. Wilmington, DE: American Professional Credentialing Services.

Lavine, M. (1992). Some aspects of polya tree distributions for statistical modelling. *The Annals of Statistics 20*(3), 1222–1235.

Lijoi, A. and I. Prünster (2010). Models beyond the Dirichlet process. In N. Hjort, P. Holmes, C.and Müller, and S. G. Walker (Eds.), *Bayesian Nonparametrics*, pp. 80–136. Cambridge University Press.

Lin, L., A. S. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement. *Journal of the American Statistical Association 97*(457), 257–270.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics 45*(1), 255–268.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *The Annals of Statistics 12*, 351–357.

Lord, F. (1964). Nominally and rigorously parallel test forms. *Psychometrika 29*(4), 335–345.

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

Lord, F. (1982). The standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics 7*(3), 165.

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

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Department of Statistics, The Ohio State University.

MacEachern, S. N. and P. Müller (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics 7*(2), 223–238.

Mauldin, R., W. Sudderth, and S. Williams (1992). Polya trees and random distributions. *The Annals of Statistics 20*(3), 1203–1221.

McCullagh, P. (1980). Regression models for ordinal data. *J. R. Stat. Soc. Ser. B (Stat. Methodology) 42*(2), 109–142.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall: London.

Mitra, R. and P. Müller (2015). *Nonparametric Bayesian Inference in Biostatistics*. Springer International Publishing Switzerland.

Müller, P., F. Quintana, A. Jara, and T. Hanson (2015). *Bayesian Nonparametric Data Analysis*. New York: Springer.

Müller, P., F. Quintana, and G. Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society, Series B 66*(3), 735–749.

Neal, R. M. (2000). Markov Chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.

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

Pommerich, M. and N. J. Dorans (2004). Linking scores via concordance: Introduction to the special issue. *Applied Psychological Measurement 28*(4), 216–218.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rodríguez, A. and D. Dunson (2011). Nonparametric Bayesian models through probit stick-breaking processes. *BMC Medical Research Methodology* (6), 145–178.

Santor, D., M. Gregus, and A. Welch (2009). Eight decades of measurement in depression. *Measurement 4*, 135–155.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, 639–650.

Shah, D. A. and L. V. Madden (2004). Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology 91*(1), 33–43.

Stewart, M. (2004). A comparison of semiparametric estimators for the ordered response model. *Computational Statistics and Data Analysis 49*, 555–573.

Titov, N., F. D. Blake, D. McMillan, T. Anderson, J. Zou, and M. Sunderland (2011). Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cognitive Behaviour Therapy 40*(2), 126–136.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics 33*(1), 1–167.

Valdés, C., I. Morales-Reyes, J. C. Pérez, A. Medellín, G. Rojas, and M. Krause (2017). Propiedades psicométricas del inventario de depresión de Beck IA para la población chilena. *Revista médica de Chile 145*, 1005 – 1012.

van der Linden, W. J. (2011). Local observed-score equating. In A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking*, pp. 201–223. New York: Springer.

van der Linden, W. J. (2013). Some conceptual issues in observed-score equating. *Journal of Educational Measurement 50*(3), 249–285.

van Steenoven, I., D. Aarsland, H. Hurtig, A. Chen-Plotkin, J. E. Duda, J. Rick, M. C. Lama, N. Dahodwala, J. Q. Trojanowski, R. D. R., P. J. Moberg, and D. Weintraub (2014). Conversion between Mini-Mental State Examination, Montreal Cognitive Assessment, and Dementia Rating Scale-2 scores in Parkinson's disease. *Movement Disorders 29*(14), 1809–1815.

Varas, I., J. González, and F. A. Quintana (2019). A new equating method through latent variables. In M. Wiberg, S. A. Culpepper, R. Janssen, J. González, and D. Molenaar (Eds.), *Quantitative psychology*, pp. 343–353. Cham: Springer.

von Davier, A. (2011). *Statistical Models for Test Equating, Scaling, and Linking*. New York: Springer.

von Davier, A. A., P. Holland, and D. Thayer (2004). *The Kernel method of Test Equating*. New York: Springer.

Wiberg, M. and K. Bränberg (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement 39*(5), 349–361.

Wiberg, M., W. J. van der Linden, and A. A. von Davier (2014). Local observed-score kernel equating. *Journal of Educational Measurement 51*, 57–74.

Wilk, M. and R. Gnanadesikan (1968). Probability plotting methods for the analysis of data. *Biometrika 55*(1), 1–17.

Winship, C. and R. D. Mare (1984). Regression models with ordinal variables. *American Sociological Review 49*(4), 512–525.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association 86*(413), 79–86.