Laboratorio
Interdisciplinario de
Estadística Social

Discussion Paper N° 2020|01

How Fair is to be Fair?
Revisiting Test Equating under the NEAT Design

Ernesto San Martín & Jorge González

Facultad de Matemáticas
Pontificia Universidad Católica de Chile
Av. Vicuña Mackenna 4860, Macul
lies.mat.uc.cl/

How Fair is to be Fair?

Revisiting Test Equating under the NEAT Design

Ernesto San Martín[1, 2, 3], and Jorge González[1, 2]

[1] *Faculty of Mathematics, Pontificia Universidad Católica de Chile, Chile*

[2] *Interdisciplinary Laboratory of Social Statistics*

[3] *The Economics School of Louvain, Université Catholique de Louvain, Belgium*

January 7, 2020

Abstract

The nonequivalent groups with anchor test design (NEAT) is widely used in test equating. Under this design, two groups of examinees are administered different test forms with each test form containing a subset of common items. Because test takers from different groups are assigned only one test form, missing score data emerge by design rendering some of the score distributions unavailable. The partial observability of the score data formally leads to an identifiability problem which has not been recognized as such in the equating literature, and has been faced from different perspectives, all of them making different assumptions in order to estimate the unidentified score distributions.

In this paper, we formally specify the statistical model underlying the NEAT design and unveil the lack of identifiability of the parameters of interest that compose the equating transformation. We use the theory of partial identification to offer alternatives to traditional practices used to point identify the score distributions when conducting equating under the NEAT design.

Key words: Partial observability; Missing observations; Strong ignorability; Equity; Interchangeable scores; Quantiles.

1 Introduction

Test equating is conducted to adjust the scores of different test forms in order to compensate for differences in relative difficulty and thus make scores equivalent and comparable (Angoff, 1984; Kolen & Brennan, 2014). The *equating problem* can be considered as an statistical problem in which scores defined on one scale are to be mapped into their equivalents on the other scale. Such mapping is achieved using what is called an equating transformation function (for details, see González & Wiberg, 2017, Chapter 1).

Because score differences can also be due to ability differences of test takers, comparable groups of examinees must be used when collecting score data to estimate the equating function. Different strategies for collecting score data has been proposed in the literature, leading to what are called equating designs (see, e.g., von Davier, Holland, & Thayer, 2004, Chapter 2); (Kolen & Brennan, 2014, Section 1.4) and González and Wiberg (2017, Section 1.3.1)). These designs differ in that either common persons or common items are used to perform the score transformation. In this paper we will focus the attention on the nonequivalent groups with anchor test design (NEAT).

The NEAT design (a.k.a the common item nonequivalent group design, CINEG) is widely used in test equating. Under this design, two groups of test takers are administered different test forms with each test form containing a subset of common items. Because test takers from different groups are assigned only one test form, *missing* score data emerge by design rendering some of the score distributions unavailable. The partial observability of the score data formally leads to an identifiability problem, which has not been recognized as such in the equating literature, and has been faced from different perspectives, all of them making different assumptions in order to estimate the unidentified score distributions; see Sinharay and Holland (2010), Miyazaki, Hoshino, Mayekawa, and Shigemasu (2009), Liou and Cheng (1995), Holland, Sinharay, von Davier, and Han (2008), among others. Among these approaches, considering the conditional probability distributions of test scores given the anchor scores and imposing the strong ignorability condition to point identify the score distributions has been the traditional practice when performing equating under the NEAT design.

In this paper, we formally specify the statistical model underlying the NEAT design and unveil the lack of identifiability of the parameters of interest that compose the equating transformation. Three possible solutions to the identification problem are shown including the traditional practice of point identify the score distributions by imposing the strong ignorability condition. We follow a partial identification approach (Manski, 2003) under which the score probability distributions used to obtain the equating transformation are bounded on a region where they are identified by the data. We offer two solutions, one in which no particular assumptions are needed to obtain the identification bounds, and another one that considers the NEAT design under a self-selection process. The results show that the uncertainty about the score probability distributions, reflected on the width of the identification bounds, can be very large leading to question how fair is to be fair conducting equating under the NEAT design.

The paper is organized as follows: in Section 2 the basics elements involved in the statistical model underlying the NEAT design are defined and the identifiability problem is described. In Section 3 we offer an alternative to point identifiability based on the theory of partial identifiability. The severity of assuming strong ignorability on equating is illustrated considering both the score probability distributions and the quantiles of them. Section 4 illustrates the partial identification approach under an optimist self-selection assumption in which examinees make a rational choice about the test to take based on their beliefs of better performance. The paper finalizes in Section 5 summarizing the main points and discussing on how severe can it be the identifiability problem in the context of test equating.

2 The Statistical Model underlying the NEAT Design

2.1 Basic definitions

A *statistical model* $\mathfrak{E}$ involves three components: i) observations of a population of interest lying on a sample space $\mathcal{X}$; ii) probability distributions, F^θ, defined on $\mathcal{X}$ and indexed by parameters θ; and iii) a parameter space Θ, such that $\theta \in \Theta$. The statistical model can compactly be written as

$$\mathfrak{E} = \{\mathcal{X}, F^\theta : \theta \in \Theta\}. \tag{2.1}$$

For details, see Fisher (1922), Cox and Hinkley (1979), McCullagh (2002), San Martín, González, and Tuerlinckx (2015), and San Martín (2016).

A statistical model describes the observed population in the sense that the body of observations is fully characterized by the statistical model. This is why the specification of the statistical model requires "an understanding of the way in which the data are supposed to, or did in fact, originate" (Fisher, 1973, p.8). In this case, the characteristics of a probability distribution F^θ correspond to specific characteristic of such population. These characteristics are functionals of the probability distribution F^θ, such as the mean, variance, kurtosis or quantiles. Consequently, the parameters indexing a probability distribution are functionals of it and, accordingly, "describe [the population] exhaustively in respect of all qualities under discussion" (Fisher, 1922, p.311). It is important to emphasize that all the statistical models are *parametric*: when the parameter space Θ is a finite dimensional set, the statistical model is called *a parametric model*; when Θ is an infinite dimensional one, it is called *a non-parametric model*; and when Θ is a Cartesian product of a finite set and an infinite set, it is called *semi-parametric model*.

2.2 Characteristics of the sampling process underlying the NEAT design

The observed score data originated under the NEAT design come from a sampling process characterized by two properties: Form X and Form Y are administered to different (nonequivalent) groups of examinees sampled from populations denoted here as $\mathcal{P}$ and $\mathcal{Q}$, respectively. In addition, an anchor test A is taken by both groups and is used to study the differences in ability between them. This design is widely used in applications for different reasons:

1. For security reasons, only one form needs to be administered on a given test date. This is the case for high-stakes tests as those of the Advanced Placement Program Examinations or the SAT: for details and references, see Kolen and Hendrickson (2011).
2. In some evaluation systems, where a different test form is administered on each test date, examinees can decide the occasion when taking the test. In this case, the NEAT design can be viewed as an observational study where there are non-randomized groups that are possibly subjected to varying amounts of self-selection (Holland, Dorans, & Petersen, 2007, p.179).

In spite of these advantages, in what follows we will see that the probability distributions underlying the NEAT design are non-identified which bounds its applications.

2.3 The equating function

Following Lord (1950) and Angoff (1984), "two scores, one on Form X and the other on Form Y [. . .] may be considered equivalent if their corresponding percentile ranks in any given group are equal" (Angoff, 1984, p.86). The equating transformation accordingly establishes equated scores as those that have the same percentile rank in a given group of test takers.

Formally, let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be the random variables representing test scores from tests forms X and Y, respectively; and let F_X and F_Y be their respective distribution functions. The X- and Y-scores are comparable if for each $x \in \mathcal{X}$, there exists a unique $y \in \mathcal{Y}$ such that

$$F_X(x) = P(X \leq x) = p = P(Y \leq y) = F_Y(y),\ p \in (0,1).$$

Assuming that both F_X and F_Y are strictly increasing functions, *the equating function* $\varphi : \mathcal{X} \longrightarrow \mathcal{Y}$ is defined as

$$\mathcal{Y} \ni y = \varphi(x) \doteq F_Y^{-1}(F_X(x)) \qquad \forall\, x \in \mathcal{X}. \tag{2.2}$$

It should be noted that this definition implicitly assumes that both $\mathcal{X}$ and $\mathcal{Y}$ are "of the same nature", that is, both on them are subset of a common set which can be considered (an interval of) the real line.

Although most of the time scores are assumed to take values as consecutive integers (i.e., the total number of correct answers), which implies that both F_X and F_Y are discrete distribution functions, all practical applications of test equating make use of what is called a *continuization* step; for details, see Braun and Holland (1982), Kolen and Brennan (2014) and González and Wiberg (2017). In passing, let us mention that Braun and Holland (1982, p.15) state that "to define the inverse of [a probability distribution . . .] is only problematic if the number of possible values of the raw scores X and Y is small".

In practice, score data are needed for the estimation of φ. The actually observed score data are assumed to be realizations of the random variables X and Y, and are used to estimate F_X and F_Y. For details, see, González and Wiberg (2017, Chapter 1).

2.4 The parameters of interest underlying the NEAT design

2.4.1 Construction of the probability distributions underlying the NEAT design

According to the description of the NEAT design, three elements can be distinguished:

1. Two populations of examinees, $\mathcal{P}$ and $\mathcal{Q}$: examinees of population $\mathcal{P}$ are administered Form X, whereas examinees of population $\mathcal{Q}$, the Form Y.

2. The relative size (i.e., the proportion of examinees) of populations $\mathcal{P}$ and $\mathcal{Q}$.

3. A set of common items is administered to both populations. These items are gathered in the Form A.

Using this information, it is possible to define the following probability distributions:

1. Under the condition that the test takers belong to population $\mathcal{P}$, we can define a probability distribution on the set $\mathcal{A}$ of all possible scores from test A, and a probability distribution on the set $\mathcal{X}$ of all possible scores from test X.

2. Under the condition that the test takers belong to population $\mathcal{Q}$, we can define a probability distribution on the set $\mathcal{A}$ of all possible scores from test A, and a probability distribution on the set $\mathcal{Y}$ of all possible scores from test Y.

These probability distributions are by design *conditional distributions*, the first ones conditional on the event *test takers of population* $\mathcal{P}$, the second ones conditional on the event *test takers of population* $\mathcal{Q}$.

In order to properly write these distributions, we need to introduce a binary random variable Z defined as

$$Z = \begin{cases} 1, & \text{if test taker is admnistered X and A in } \mathcal{P}; \\ 0, & \text{if test taker is admnistered Y and A in } \mathcal{Q}. \end{cases}$$

This random variable makes explicit the fact that both population are mutually exclusive. The relative size of population $\mathcal{P}$ is accordingly given by $P(Z = 1)$, and the relative size of population $\mathcal{Q}$, by $P(Z = 0)$. The distribution of scores $x \in \mathcal{X}$ under the condition that test takers belong to $\mathcal{P}$ is given by $F_{X|Z=1}(x) = P(X \leq x \mid Z = 1)$ for $x \in \mathcal{X}$: here X is a random variable defined on $\mathcal{X}$. The distribution of scores $y \in \mathcal{Y}$ under the condition that test takers belong to $\mathcal{Q}$ is given by $F_{Y|Z=0}(y) = P(Y \leq y \mid Z = 0)$ for $y \in \mathcal{Y}$: here Y is a random variable defined on $\mathcal{Y}$. Similarly, if A denotes a random variable defined on $\mathcal{A}$, the distribution of scores $a \in \mathcal{A}$ under the condition that test takers belong to $\mathcal{P}$ is given by $F_{A|Z=1}(a) = P(A \leq a \mid Z = 1)$; and the distribution of scores $a \in \mathcal{A}$ under the condition that test takers belong to $\mathcal{Q}$ is given by $F_{A|Z=0}(a) = P(A \leq a \mid Z = 0)$.

Summarizing, the observations collected under the NEAT design are fully characterized by the following five probability distributions:

$$\begin{aligned} &P(Z = z), \;\; z \in \{0, 1\}; \quad F_{A|Z=1}(a), \;\; a \in \mathcal{A}; \quad F_{A|Z=0}(a), \;\; a \in \mathcal{A}; \\ &F_{X|Z=1}(x), \;\; x \in \mathcal{X}; \quad F_{Y|Z=0}(y), \;\; y \in \mathcal{Y}. \end{aligned} \tag{2.3}$$

Using these probability distributions, it is possible to make explicit the statistical model underlying the NEAT design. To this end, it is enough to decompose the joint distribution generating the available observations, namely $P(X \leq x, Y \leq y, A \leq a, Z = z)$ for all $(x, y, a, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \times \{0, 1\}$.

Taking into account the NEAT design itself, we can see that first Z is generated which means the self-selection or assignment of Forms X and Y. Thereafter, A is generated because the Form A is applied to both populations. Conditionally on $Z = 1$, it is generated X, and conditionally on $Z = 0$, it is generated Y. Therefore, the joint distribution generating (X, Y, Z, A) is decomposed as

$$\begin{aligned} &P(X \leq x, Y \leq y, A \leq a, Z = z) = \\ &\qquad P(X \leq x \mid A \leq a, Z = 1)\, P(Y \leq y \mid A \leq a, Z = 0)\, P(A \leq a \mid Z)\, P(Z = z) \end{aligned}$$

for all $(x, y, a, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \times \{0, 1\}$, which means to assume the following conditions:

$$\begin{aligned} &\text{(i)} \quad X \perp\!\!\!\perp Y \mid (A, Z). \\ &\text{(ii)} \quad X \perp\!\!\!\perp \mathbb{1}_{\{Z=0\}} \mid (A, \mathbb{1}_{\{Z=1\}}). \\ &\text{(iii)} \quad Y \perp\!\!\!\perp \mathbb{1}_{\{Z=1\}} \mid (A, \mathbb{1}_{\{Z=0\}}). \end{aligned} \tag{2.4}$$

Thus, the statistical model underlying the NEAT design is given by the joint distribution generating (X, Y, A, Z) which is characterized by conditions (2.4). It should be noticed that in this case the involved probabilities are at the same time the parameters of the statistical model.

Remark 2.1 The random variable Z could be considered as a *random assignment treatment variable*. According to this interpretation, the only situation in which the NEAT design would be valid is when the examinees are randomly assigned to the forms X and Y. We argue that this interpretation is based on a notion of *randomness* as if it were a basic concept of the Theory of Probability as axiomatically developed by Kolmogorov (1956). However, correctly understood, *random events* and *probability function* are mere terms designating, on the one hand the elements of a σ-field and, on the other hand, a function satisfying the axioms of a set function (namely, a normalization condition and the σ-additivity condition); see (Kolmogorov, 1956, p.2). Under a Kolmogorovian perspective, the relative size of populations $\mathcal{P}$ and $\mathcal{Q}$, namely $r_{\mathcal{P}}$ and $r_{\mathcal{Q}}$, can be used to construct the probability distribution of Z as follows:

$$P(Z = 1) = r_{\mathcal{P}}, \qquad P(Z = 0) = r_{\mathcal{Q}};$$

see Kolmogorov (1956, Section 2). By doing so, all the possible NEAT designs (see Section 2.2) can be embedded in the proposed formalization. ■

2.4.2 Parameters of interest and their lack of identifiability

The challenge of the NEAT design is to equate the scores obtained by the examinees of population $\mathcal{P}$ with those of examinees of population $\mathcal{Q}$. This requires to define the equating transformation in terms of the marginal probability distributions $F_X(x)$ and $F_Y(y)$: these distributions correspond to the parameters of interest underlying the NEAT design.

As it was discussed in Section 2.1, the data generating process is fully characterized by a family of probability distributions defined on the sample space: their parameters correspond to specific characteristics of the population under study. However, a fundamental question is the following: Is a specific characteristic of the population of interest a parameter of the probability distribution generating it? This is precisely an identification problem, which is formalized through the injectivity of the mapping $\Theta \ni \theta \longmapsto F^{\theta} \in \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of the probability distributions defined on the sample space $\mathcal{X}$; for details, discussion and references, see San Martín (2018).

In the case of the NEAT design the question is whether F_X and F_Y are characteristics of the observations under analysis. The answer is negative because these distributions can not be expressed as a function of the statistical model (2.4). As a matter of fact, on the one hand

$$F_X(x) = E[P(X \leq x \mid A)] \;\; x \in \mathcal{X}, \qquad F_Y(y) = E[P(Y \leq y \mid A)] \;\; y \in \mathcal{Y}.$$

On the other hand, using the Law of Total Probability (Kolmogorov, 1956), the conditional probabilities inside of the above expectations can be decomposed as follows:

$$F_{X|A}(x) \;\; = \;\; F_{X|A,Z=1}(x)P(Z=1 \mid A) + F_{X|A,Z=0}(x)P(Z=0 \mid A), \;\; x \in \mathcal{X} \qquad (2.5)$$

$$F_{Y|A}(y) \;\; = \;\; F_{Y|A,Z=1}(y)P(Z=1 \mid A) + F_{Y|A,Z=0}(y)P(Z=0 \mid A), \;\; y \in \mathcal{Y} \qquad (2.6)$$

Nevertheless, $F_{X|A,Z=0}(x)$ and $F_{Y|A,Z=1}(y)$ are not identified because of the partial observability inherent to the NEAT design. This implies that both conditional probabilities can not be uniquely determined, which in turn implies the non identifiability of $F_{X|A}$ and $F_{Y|A}$. The conclusion follows.

2.4.3 Questioning the notion of synthetic population

To make explicit the statistical model underlying the NEAT design not only allows us to make explicit the parameters of interest and their lack of identifiability, but also to questioning the concept of *synthetic* *population* introduced by Braun and Holland (1982). As a matter of fact, in a statistical model, the parameters of interest make explicit the characteristics of the population under study and, by extension, *the* population of interest. In the NEAT design, the parameters of interest are the distributions F_X and F_Y, which can be decomposed in a *unique* way as

$$F_X(x) \quad = \quad F_{X|Z=1}(x)P(Z=1) + F_{X|Z=0}(x)P(Z=0), \quad x \in \mathcal{X} \tag{2.7}$$

$$F_Y(y) \quad = \quad F_{Y|Z=1}(y)P(Z=1) + F_{Y|Z=0}(y)P(Z=0), \quad y \in \mathcal{Y}. \tag{2.8}$$

these decompositions follow from (2.5) and (2.6), respectively, after marginalizing A. This means that the weighs of the populations $\mathcal{P}$ and $\mathcal{Q}$ *are determined by the observations rather than chosen arbitrarily*, as suggested, for instance, by Brennan and Kolen (1987) and Kolen and Brennan (2014, Section 4.5.2); and that the population of interest is just both populations $\mathcal{P}$ and $\mathcal{Q}$ considered simultaneously. Thus, the notion of synthetic population can be considered as a valid concept only if it refers to the basic decompositions (2.5) and (2.6).

3 Solutions to the Identification Problem

3.1 Strong ignorability condition

The equating literature typically considers the lack of identifiability of the marginal distributions F_X and F_Y as a missing data problem; see, among many others, Liou and Cheng (1995), Bolsinova and Maris (2016), and Sinharay and Holland (2010). To solve the problem, it is typically assumed that the anchor items A are informative enough such that

$$F_{X|Z=1,A}(x) = F_{X|Z=0,A}(x) \quad \forall\, x \in \mathcal{X}, \qquad F_{Y|Z=1,A} = F_{Y|Z=0,A} \quad \forall\, y \in \mathcal{Y},$$

which can compactly be rewritten as

$$(X,Y) \perp\!\!\!\perp Z \mid A; \tag{3.1}$$

see Braun and Holland (1982), Kolen and Brennan (2014), von Davier et al. (2004) and González and Wiberg (2017). This condition, known as *switching condition* (Maddala, 1983) or *strong ignorability condition* (Rosenbaum & Rubin, 1983), mimics the definition of a randomized experiment (Fisher, 1935; Cochran & Chambers, 1965; Angrist & Pischke, 2008), but conditionally on a set of covariates.

Condition (3.1) is not empirically refutable (Manski, 1995, 2007), but only justified in the context of specific applications, the justification essentially depending on the choice of those covariates.

What is important to emphasize is that condition (3.1) is *an identification restriction* allowing to point identify $F_{X|A}$ and $F_{Y|A}$ and, by extension, F_X and F_Y. As a matter of fact, under condition (3.1), decompositions (2.5) and (2.6) imply that

$$F_{X|A}(x) = F_{X|Z=1,A}(x) \quad \forall\, x \in \mathcal{X}, \qquad F_{Y|A}(y) = F_{Y|Z=0,A}(y) \quad \forall\, x \in \mathcal{Y},$$

and, therefore,

$$F_X(x) = E[F_{X|Z=1,A}(x)] \quad \forall\, x \in \mathcal{X}, \qquad F_Y(y) = E[F_{Y|Z=0,A}(y)] \quad \forall\, x \in \mathcal{Y}.$$

It is known that an identification restriction reduces the parameter space of the corresponding statistical model. It should be asked how can this be evaluated in the NEAT design, which means to answer the following question: how strong is the strong identification condition?

3.2 Partial identification analysis

A natural starting point is to ask what the data alone reveal on $F_{X|A}$ and $F_{Y|A}$. This can be done by means of a partial identification strategy, widely used in empirical research; see, among many others, Pepper (2000), Kreider and Pepper (2007), Blundell, Gosling, Ichimura, and Meghir (2007), Gundersen and Kreider (2008), Gundersen and Kreider (2009), Molinari (2010), Gundersen, Kreider, and Pepper (2012) and Manski and Pepper (2013).

3.2.1 Partial identification of the parameters of interest

In order to make explicit what the data alone reveal on $F_{X|A}$, it is enough to notice that the unidentified conditional probability distribution $F_{X|Z=0,A}$ lies between 0 and 1. This means to make explicit two polar cases:

1. None of the examinees in $\mathcal{P}$ with a common anchor score equal to a would have scored at most y in Form Y; that is, $F_{X|Z=0,A=a}(y) = 0$.

2. All examinees of $\mathcal{P}$ with a common anchor score equal to a would have scored at most y points in

the Form Y; that is, $F_{X|Z=0,A=a}(y) = 1$.

Thus, without additional assumptions on $F_{X|Z=0,A}$, decomposition (2.5) implies that the conditional probability $F_{X|A}$ lies in the interval

$$F_{X|Z=1,A}(x)P(Z = 1 \mid A) \leq F_{X|A}(x) \leq F_{X|Z=1,A}(x)P(Z = 1 \mid A) + P(Z = 0 \mid A) \tag{3.2}$$

for all $x \in \mathcal{X}$. Similarly, $F_{Y|Z=1,A}$ lies between 0 and 1 and, therefore, decomposition (2.6) implies that the conditional probability $F_{Y|A}$ lies in the interval

$$F_{Y|Z=0,A}(y)P(Z = 0 \mid A) \leq F_{Y|A}(y) \leq F_{Y|Z=1,A}(y)P(Z = 0 \mid A) + P(Z = 1 \mid A) \tag{3.3}$$

for all $y \in \mathcal{Y}$. After integrating out with respect to A, these inequalities provide the partial identification intervals of the parameters of interest, which is summarized in the following theorem:

Theorem 3.1 *In the NEAT design, the parameters of interest F_X and F_Y are partially identified by the following intervals:*

$$\begin{aligned} &\textit{(i)} \quad F_{X|Z=1}(x)P(Z = 1) \leq F_X(x) \leq F_{X|Z=1}(x)P(Z = 1) + P(Z = 0), \quad x \in \mathcal{X}; \\ &\textit{(ii)} \quad F_{Y|Z=0}(y)P(Z = 0) \leq F_Y(y) \leq F_{Y|Z=0}(y)P(Z = 0) + P(Z = 1), \quad y \in \mathcal{Y}. \end{aligned} \tag{3.4}$$

This theorem deserves the following comments:

1. The partial identification intervals (3.4.i) and (3.4.ii) do not depend on any assumption; they accordingly provide all the plausible values of F_X and F_Y coherent with the data. This means that all the solutions based on specific alternative assumptions must be contained in these intervals. In particular, it can empirically be shown that the standard solution based on *target* score distributions (denoted as F_{XT} and F_{YT}) computed using the concept of synthetic population, is a plausible one, as it is shown in Figure 1.

2. Nevertheless, it is important to emphasize that the target score distribution F_{XT} does not always coincide with F_X as decomposed in (2.7), except when $\omega = P(Z = 1)$ (see (3.5) below). This is due to the fact that the target distribution F_{XT} is defined *arbitrarily* by a convex linear combination

of the conditional distributions $F_{X|Z=1}$ and $F_{X|Z=0}$, namely

$$\omega F_{X|Z=1}(t) + (1-\omega)F_{X|Z=0}(t), \quad \omega \in [0,1]. \tag{3.5}$$

As a matter of fact, (3.5) can not be represented by a conditional probability distribution, as it is the case for F_X in (2.7). The key issue of the Law of Total Probability is precisely to show how F_X can be *represented by a conditional probability distribution* (which corresponds to the Theorem of Radon-Nikodym):

$$\begin{aligned}
E\left(1\!\!1_{\{X\leq t\}}\right) &= F_X(t) \\
&= F_{X|Z=1}(t)\,P(Z=1) + F_{X|Z=0}(t)\,P(Z=0) && (3.6)\\
&= E\left[F_{X|Z=1}(t)\,1\!\!1_{\{Z=1\}} + F_{X|Z=0}(t)\,1\!\!1_{\{Z=0\}}\right] \\
&= E\left[E\left(1\!\!1_{\{X\leq t\}} \mid Z\right)\right] && (3.7)\\
&= E\left[F_{X|Z}(t)\right],
\end{aligned}$$

where (3.6) follows from the Law of Total Probability and (3.7) follows from the definition of the conditional expectation $E\left(1\!\!1_{\{X\leq t\}} \mid Z\right)$. For details, see Rao (2005). Furthermore, using the strong ignorability condition, the target score probability distribution F_{XT} is given by

$$\omega\, F_{X|Z=1}(t) + (1-\omega)\, E\left[F_{X|A,Z=1}(t) \mid Z=0\right], \quad \omega \in [0,1].$$

When F_X can be represented by a conditional probability distribution, then it can be interpreted as a probability distribution defined on the population of interest which in this case is just both populations $\mathcal{P}$ and $\mathcal{Q}$ considered simultaneously.

3. The width of these intervals depend on the relative size of populations $\mathcal{P}$ and $\mathcal{Q}$: the width of interval (3.4.i) is $P(Z=0)$, whereas the width of interval (3.4.ii) is $P(Z=1)$. Taking into account that $P(Z=0)+P(Z=1)=1$, the longer an interval is, the shorter the other: it is a trade-off between both populations relative sizes, except for the case $P(Z=0)=P(Z=1)$. In particular, this shows how strong is the ignorability condition: $F_X(x)$ with $x \in \mathcal{X}$ (respect., $F_Y(y)$ with $y \in \mathcal{Y}$) belongs to an identification interval of width $P(Z=0)$ (respect., $P(Z=1)$); under the ignorability condition, this interval collapses to a point, namely $F_{X|Z=1}(x)$ (respect., $F_{Y|Z=0}(y)$). The inherent uncertainty that the ignorability condition allows to hide is precisely

the relative sizes of the populations $\mathcal{P}$ and $\mathcal{Q}$, which can not be arbitrarily manipulated. Figure 2 shows this trade-off graphically.

4. It is clear that the partial identification intervals (3.2) and (3.3) are similar in form to the intervals (3.4.i) and (3.4.ii). Moreover, it might seem that the latter can be deduced *without using the anchor items*: it is enough to directly use the decompositions

$$\begin{aligned} F_X(x) &= F_{X\mid Z=1}(x)P(Z=1) + F_{X\mid Z=0}(x)P(Z=0), \quad x \in \mathcal{X}, \\ F_Y(y) &= F_{Y\mid Z=1}(y)P(Z=1) + F_{Y\mid Z=0}(y)P(Z=0), \quad y \in \mathcal{Y}. \end{aligned}$$

and the fact that $F_{X\mid Z=1}$ and $F_{Y\mid Z=0}$ are identified. However, taking into account the statistical model underlying the NEAT design, these distributions are not directly identified, but also through the equalities

$$F_{X\mid Z=1}(x) = E[F_{X\mid Z=1,A}(x) \mid A], \;\; x \in \mathcal{X}; \quad F_{Y\mid Z=0}(x) = E[F_{Y\mid Z=0,A}(y) \mid A], \;\; y \in \mathcal{Y},$$

given that $F_{X\mid Z=1,A}$ and $F_{Y\mid Z=0,A}$ are identified parameters of the NEAT design. It could be replied that, looking at the data, it is plain that $F_{X\mid Z=1}$ and $F_{Y\mid Z=0}$ are identified without requiring the information provided by A. If this was the case, then it would be necessary to assume that $Z \perp\!\!\!\perp A$. Nevertheless, such condition does not correspond to any characteristic of the NEAT design.

5. The larger are the partial identification intervals, the more severe is the identification problem. It could be argued that such severity could be decreased by conditioning on the anchor scores A. However, this is not the case. It is enough to compare the width of the conditional partial identification intervals (3.2) and (3.3) with the intervals (3.4.i) and (3.4.ii), respectively. For each $a \in \mathcal{A}$, three possibilities can be considered:

$$\begin{array}{rlcl} \text{(i)} & P(Z=1) > P(Z=1 \mid A=a) & \iff & P(Z=0) < P(Z=0 \mid A=a) \\ \text{(ii)} & P(Z=1) < P(Z=1 \mid A=a) & \iff & P(Z=0) > P(Z=0 \mid A=a) \\ \text{(iii)} & P(Z=1) = P(Z=1 \mid A=a) & \iff & P(Z=0) = P(Z=0 \mid A=a). \end{array} \tag{3.8}$$

These conditions are empirically testable. Moreover, in cases (3.8.i) and (3.8.ii), the longer an interval is, the shorter the other. Figure 3 shows a graphical representation of these comparisons. Note that these inequalities can be interpreted in terms of covariance because, for instance, (3.8.i) is equivalent to

$$E[\mathbb{1}_{\{Z=1\}}\mathbb{1}_{\{A=a\}}] - E[\mathbb{1}_{\{Z=1\}}]\,E[\mathbb{1}_{\{A=a\}}] = cov(\mathbb{1}_{\{Z=1\}}, \mathbb{1}_{\{A=a\}}) < 0.$$

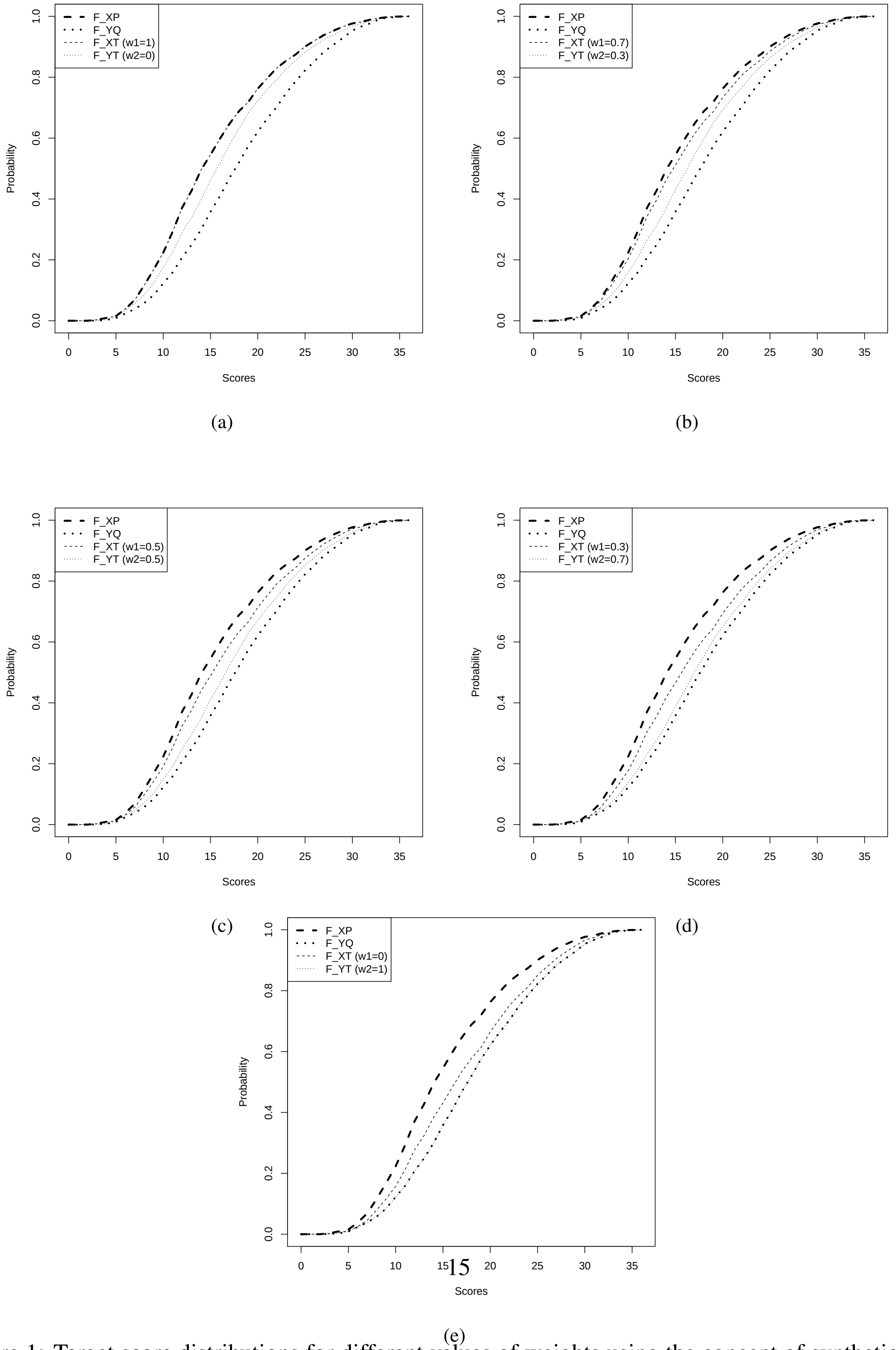
In particular, this suggests to empirically describe $P(Z = 1 \mid A)$ as a function of A, namely

$$P(Z = 1 \mid A) = \sum_{a \in \mathcal{A}} P(Z = 1 \mid A = a)\, \mathbb{1}_{\{A=a\}}.$$

So far, we have been general in considering the complete score distributions that lead to the equipercentile equating function. To finalize this section, let us for completeness mention that there exist linear versions of the equating function that also suffer from the inherent identifiability problem under the NEAT design. To estimate the unidentified parameters in a linear equating setting, Angoff (1984) used what he called the three principal assumptions of univariate selection theory (Angoff, 1984, page 110), namely that the intercept, the slope and the variance of the regression of X on $(A, Z = 1)$ are the same as the corresponding intercept, slope and variance of the regression of X on $(A, Z = 0)$. He referred to Gulliksen (1950, Chapter 11), who in turn justified these assumptions by assuming that the joint distribution of X and A (respectively, Y and A) is fully known before administering test form $\mathtt{X}$ to population $\mathcal{P}$ (respectively, the test form $\mathtt{Y}$ to population $\mathcal{Q}$), which seems to be a very strong assumption; for details, see Alarcón-Bustamante, San Martín, and González (in press, and the references therein) and, in the context of equating, Braun and Holland (1982, Theorem 4). However these three assumptions can be derived from the strong ignorability condition in the case that the score distributions are completely known up to the first two moments.

3.2.2 Partial identification of the quantiles

As discussed in Section 2.3, the equating function actually equates the quantiles of the score distributions F_X and F_Y. As a matter of fact, when X is a continuous random variable then F_X^{-1} is the corresponding quantile function. Thus, it is of interest to see how the partial identifiability of the score probability distributions described in the previous sections extrapolates to the quantiles. In this section we will show what is the impact of the partial identification of F_X and F_Y on the corresponding quantiles.



(a)

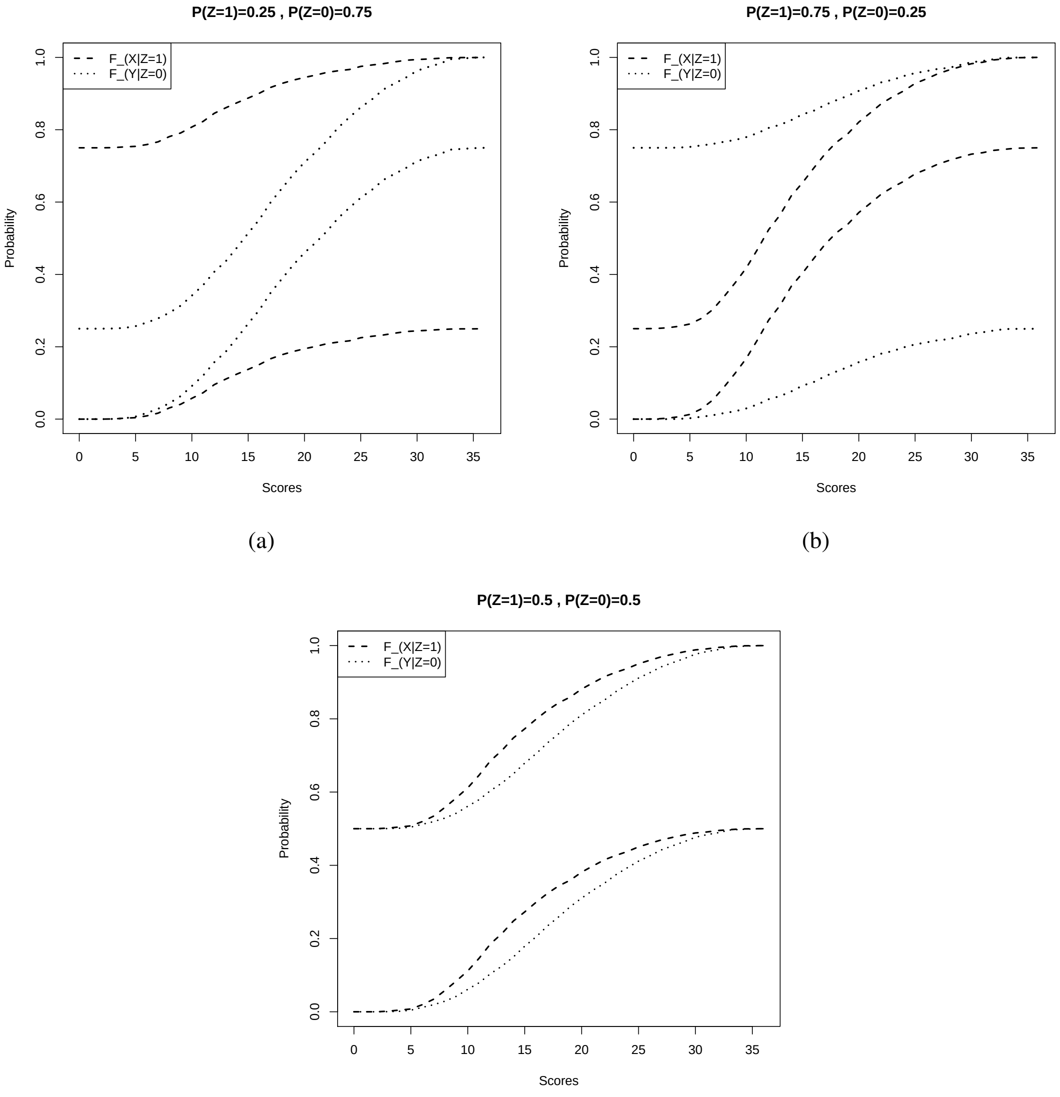
(b)

(c)

(d)

(e)

Figure 1: Target score distributions for different values of weights using the concept of synthetic population.

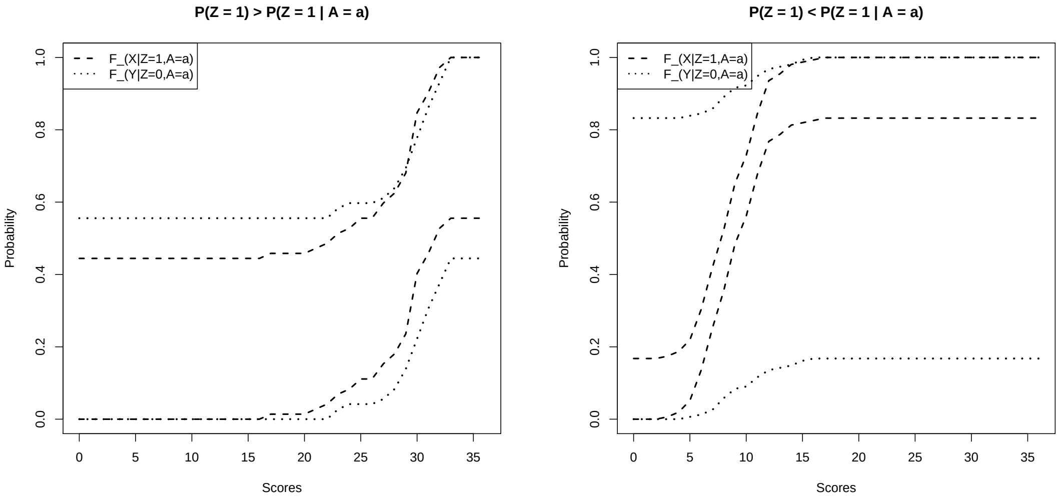


(a)

(b)

(c)

Figure 2: Identifiability bounds for different relative sizes



(a)

(b)

Figure 3: Identifiability bounds for conditional score distributions

Let $\alpha \in (0, 1)$. Define the following quantiles:

$$\begin{aligned} q_X(\alpha) &\doteq \inf\{t : F_X(t) > \alpha\}; \\ r_X(\alpha) &\doteq \inf\{t : F_{X|Z=1}(t)P(Z = 1) + P(Z = 0) > \alpha\}; \\ s_X(\alpha) &\doteq \inf\{t : F_{X|Z=1}(t)P(Z = 1) > \alpha\}. \end{aligned}$$

Note that $r_X(\alpha)$ and $s_X(\alpha)$ are identified, whereas $q_X(\alpha)$ is unidentified. The problem is to partially identify $q_X(\alpha)$ using both $r_X(\alpha)$ and $s_X(\alpha)$. To do that, we establish the following relationships using (3.4.i):

$$\begin{aligned} t < r_X(\alpha) &\implies F_{X|Z=1}(t)P(Z = 1) + P(Z = 0) < \alpha \\ &\implies F_X(t) < \alpha \\ &\implies q_X(\alpha) > t. \end{aligned}$$

It follows that $r_X(\alpha) < q_X(\alpha)$: if not, take $t = q_X(\alpha)$ and conclude that $F_X[q_X(\alpha)] < \alpha$, which is a contradiction with the definition of $q_X(\alpha)$. On the other hand,

$$\begin{aligned} t \geq s_X(\alpha) &\implies F_{X|Z=1}(t)P(Z=1) > \alpha \\ &\implies F_X(t) \geq \alpha \\ &\implies q_X(\alpha) \leq t. \end{aligned}$$

It follows that $q_X(\alpha) \leq s_X(\alpha)$ because $\alpha/P(Z=1) \geq \alpha$.

Similarly, for $\alpha \in [0,1]$, define the following quantiles:

$$\begin{aligned} q_Y(\alpha) &\doteq \inf\{t : F_Y(t) > \alpha\}; \\ r_Y(\alpha) &\doteq \inf\{t : F_{Y|Z=0}(t)P(Z=0) + P(Z=1) > \alpha\}; \\ s_Y(\alpha) &\doteq \inf\{t : F_{Y|Z=0}(t)P(Z=0) > \alpha\}. \end{aligned}$$

By using (3.4.ii) we conclude that $r_Y(\alpha) \leq q_Y(\alpha) \leq s_Y(\alpha)$.

Summarizing, we obtain the following theorem:

Theorem 3.2 *In the NEAT design, the quantiles of the partially identified probability distributions F_X and F_Y are partially identified by the following intervals:*

$$\begin{aligned} &\textit{(i)} \quad r_X(\alpha) \leq q_X(\alpha) \leq s_X(\alpha); \\ &\textit{(ii)} \quad r_Y(\alpha) \leq q_Y(\alpha) \leq s_Y(\alpha). \end{aligned} \tag{3.9}$$

In the following section we discuss on the relevant aspects that can be learnt from what the data allows to identify, showing how severe is the identification problem underlying the NEAT design.

3.2.3 Consequences of the partial identifiability of the quantiles

In order to describe both the lower and upper bounds of the quantiles $q_X(\alpha)$ and $q_Y(\alpha)$ as a function of the relative sizes of populations $\mathcal{P}$ and $\mathcal{Q}$, let us introduce additional notation: let $S_{(X|Z=1)}$ and $S_{(Y|Z=0)}$ be the supports of the conditional distributions $F_{X|Z=1}$ and $F_{Y|Z=0}$, respectively. Let

$$t_m^{X|Z=1} \doteq \min\{t : t \in S_{(X|Z=1)}\}, \qquad t_M^{X|Z=1} \doteq \max\{t : t \in S_{(X|Z=1)}\}.$$

Similar definitions apply for, $t_m^{Y|Z=0}$ and $t_M^{Y|Z=0}$.

The lower and upper bounds of $q_X(\alpha)$ can be expressed as quantiles of the conditional distribution $F_{X|Z=1}$, namely

$$\begin{aligned} r_X(\alpha) &= \inf\left\{t : F_{X|Z=1}(t) > \frac{\alpha - P(Z=0)}{P(Z=1)}\right\} = q_{X|Z=1}\left(\frac{\alpha - P(Z=0)}{P(Z=1)}\right), \\ s_X(\alpha) &= \inf\left\{t : F_{X|Z=1}(t) > \frac{\alpha}{P(Z=1)}\right\} = q_{X|Z=1}\left(\frac{\alpha}{P(Z=1)}\right). \end{aligned}$$

Similarly, the lower and upper bounds of $q_Y(\alpha)$ can be expressed as quantiles of the conditional distribution $F_{Y|Z=0}$, namely

$$\begin{aligned} r_Y(\alpha) &= \inf\left\{t : F_{Y|Z=0}(t) > \frac{\alpha - P(Z=1)}{P(Z=0)}\right\} = q_{Y|Z=0}\left(\frac{\alpha - P(Z=1)}{P(Z=0)}\right), \\ s_Y(\alpha) &= \inf\left\{t : F_{Y|Z=0}(t) > \frac{\alpha}{P(Z=0)}\right\} = q_{Y|Z=0}\left(\frac{\alpha}{P(Z=0)}\right). \end{aligned}$$

Three cases can be distinguished:

1. $P(Z=1) < P(Z=0)$. The respective upper and lower bounds are detailed in Table 1.
2. $P(Z=1) > P(Z=0)$. The respective upper and lower bounds are detailed in Table 2.
3. $P(Z=1) = P(Z=0)$. The respective upper and lower bounds are detailed in Table 3.

Table 1, 2 and 3 make explicit how severe is the non-uniqueness of the equating transformation due to the partial observability inherent to the NEAT design. Figure 4 complements these tables showing plots of the quantile functions en each case. Let us comment on this problem by making reference to Table 1, focusing the attention on the role of the relative sizes of populations $\mathcal{P}$ and $\mathcal{Q}$ in the partially identified quantiles.

1.1. If $\alpha \in (\,P(Z=1), P(Z=0)\,)$, $q_X(\alpha)$ is uninformative because $(r_X(\alpha), s_X(\alpha)) = S_{(X|Z=1)}$, whereas $q_Y(\alpha)$ provides information characterized by the respective identification interval. This means that all the Y-scores in the interval $(r_Y(\alpha), s_Y(\alpha))$ are equivalent to all possible X-scores; see Figure 4(a).

1.2. Let $\alpha \in [0, P(Z=1)]$. Suppose that $F_{X|Z=1}$ stochastically dominates $F_{Y|Z=0}$, that is, $F_{X|Z=1}(t) \leq F_{Y|Z=0}(t)$ for all $t \in \mathcal{W} \supseteqq \mathcal{X} \cup \mathcal{Y}$ (for details on stochastic dominance, see Shaked & Shanthikumar, 2007). Note that this implies that $S_{(Y|Z=0)} \supseteqq S_{(X|Z=1)}$. Given that the quantile function respects the stochastic dominance (see, e.g., Stoye, 2010), it follows that

$$q_{Y|Z=0}\left(\frac{\alpha}{P(Z=0)}\right) \leq q_{X|Z=1}\left(\frac{\alpha}{P(Z=0)}\right).$$

Because $P(Z=1) < P(Z=0)$, it follows that $\alpha/P(Z=0) < \alpha/P(Z=1)$ and

$$\begin{aligned} s_Y(\alpha) = q_{Y|Z=0}\left(\frac{\alpha}{P(Z=0)}\right) &\leq q_{X|Z=1}\left(\frac{\alpha}{P(Z=0)}\right) \\ &\leq q_{X|Z=1}\left(\frac{\alpha}{P(Z=1)}\right) = s_X(\alpha); \end{aligned}$$

see Figure 4(a). Thus, if moreover we assume for simplicity that $t_m^{X|Z=1} = t_m^{Y|Z=0}$, to equate a X-score to a Y-score means to compress or shrink the X-scores. Let us mention that in this case there not exist an explicit order between $r_X(\alpha)$ and $r_Y(\alpha)$; see Figure 4(d).

1.3. Let $\alpha \in [P(Z=0), 1]$. Suppose now that $F_{Y|Z=0}$ stochastically dominates $F_{X|Z=1}$, that is, $F_{Y|Z=0}(t) \leq F_{X|Z=1}(t)$ for all $t \in \mathcal{W} \supseteqq \mathcal{X} \cup \mathcal{Y}$, which implies that $S_{(X|Z=1)} \supseteqq S_{(Y|Z=0)}$. It follows that

$$q_{X|Z=1}\left(\frac{\alpha - P(Z=0)}{P(Z=1)}\right) \leq q_{Y|Z=0}\left(\frac{\alpha - P(Z=0)}{P(Z=1)}\right).$$

Because $P(Z=1) < P(Z=0)$ implies that

$$\frac{\alpha - P(Z=1)}{P(Z=0)} > \frac{\alpha - P(Z=0)}{P(Z=1)},$$

it follows that

$$\begin{aligned} r_X(\alpha) = q_{X|Z=1}\left(\frac{\alpha - P(Z=0)}{P(Z=1)}\right) &\leq q_{Y|Z=0}\left(\frac{\alpha - P(Z=0)}{P(Z=1)}\right) \\ &\leq q_{Y|Z=0}\left(\frac{\alpha - P(Z=1)}{P(Z=0)}\right) = r_Y(\alpha). \end{aligned}$$

Thus, if moreover we assume for simplicity that $t_M^{X|Z=1} = t_M^{Y|Z=0}$, to equate a X-score to a Y-score means to expand the X-scores. Let us mention that in this case there not exist an explicit order between $s_X(\alpha)$ and $s_Y(\alpha)$.

Similar comments can be done for Table 2; see in particular Figure 4(b). Regarding Table 3, it can be noted that if $F_{X|Z=1} = F_{Y|Z=0}$, the severity of the identification problem still persists: an X-score is not

Table 1: Lower and upper bounds for $q_X(\alpha)$ and $q_Y(\alpha)$ when $P(Z=1) < P(Z=0)$

α	$r_X(\alpha)$	$s_X(\alpha)$	$r_Y(\alpha)$	$s_Y(\alpha)$
$[\,0, P(Z=1)\,]$	$t_m^{X\|Z=1}$	$q_{X\|Z=1}\left(\frac{\alpha}{P(Z=1)}\right)$	$t_m^{Y\|Z=0}$	$q_{Y\|Z=0}\left(\frac{\alpha}{P(Z=0)}\right)$
$(\,P(Z=1), P(Z=0)\,)$	$t_m^{X\|Z=1}$	$t_M^{X\|Z=1}$	$q_{Y\|Z=0}\left(\frac{\alpha-P(Z=1)}{P(Z=0)}\right)$	$q_{Y\|Z=0}\left(\frac{\alpha}{P(Z=0)}\right)$
$[\,P(Z=0), 1\,]$	$q_{X\|Z=1}\left(\frac{\alpha-P(Z=0)}{P(Z=1)}\right)$	$t_M^{X\|Z=1}$	$q_{Y\|Z=0}\left(\frac{\alpha-P(Z=1)}{P(Z=0)}\right)$	$t_M^{Y\|Z=0}$

Table 2: Lower and upper bounds for $q_X(\alpha)$ and $q_Y(\alpha)$ when $P(Z=1) > P(Z=0)$

α	$r_X(\alpha)$	$s_X(\alpha)$	$r_Y(\alpha)$	$s_Y(\alpha)$
$[\,0, P(Z=0)\,]$	$t_m^{X\|Z=1}$	$q_{X\|Z=1}\left(\frac{\alpha}{P(Z=1)}\right)$	$t_m^{Y\|Z=0}$	$q_{Y\|Z=0}\left(\frac{\alpha}{P(Z=0)}\right)$
$(\,P(Z=0), P(Z=1)\,)$	$q_{X\|Z=1}\left(\frac{\alpha-P(Z=0)}{P(Z=1)}\right)$	$q_{X\|Z=1}\left(\frac{\alpha}{P(Z=1)}\right)$	$t_m^{Y\|Z=0}$	$t_M^{Y\|Z=0}$
$[\,P(Z=1), 1\,]$	$q_{X\|Z=1}\left(\frac{\alpha-P(Z=0)}{P(Z=1)}\right)$	$t_M^{X\|Z=1}$	$q_{Y\|Z=0}\left(\frac{\alpha-P(Z=1)}{P(Z=0)}\right)$	$t_M^{Y\|Z=0}$

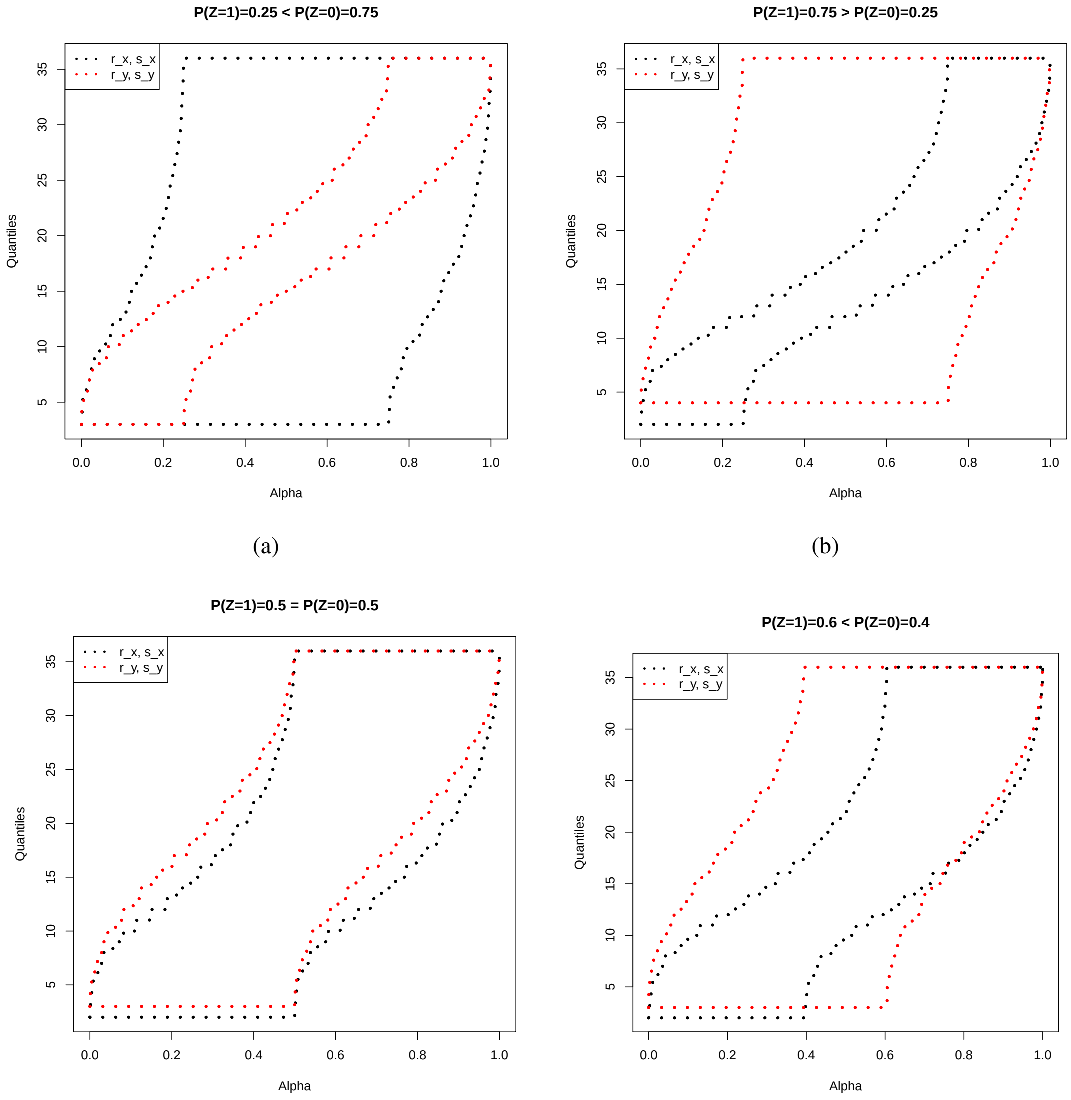
uniquely transformed to a Y-score; see Figure 4(c).

3.3 Consequences of the proposed solution

Following Lord (1980), Dorans and Holland (2000), von Davier et al. (2004), Kolen and Brennan (2014) and González and Wiberg (2017, Section 1.2.6), for test scores to be considered interchangeable, the elements in $\mathcal{X}$ and $\mathcal{Y}$ must be *similar in nature* –requirement that is necessary to perform the equating, as it was pointed out above in Section 2.3. The extent of such similarity is summarized in the following requirements that are needed for the mapping to be validly called an equating: i) *same construct*: the test forms being equated should measure the same construct; ii) *reliability*: the test forms should be equally reliable; iii) *symmetry*: the equating transformation to map $\mathcal{Y}$ into $\mathcal{X}$ should be the inverse φ^{-1} of φ as

Table 3: Lower and upper bounds for $q_X(\alpha)$ and $q_Y(\alpha)$ when $P(Z=1) = P(Z=0)$

α	$r_X(\alpha)$	$s_X(\alpha)$	$r_Y(\alpha)$	$s_Y(\alpha)$
$[\,0, \frac{1}{2}\,)$	$t_m^{X\|Z=1}$	$q_{X\|Z=1}\,(2\alpha)$	$t_m^{Y\|Z=0}$	$q_{Y\|Z=0}\,(2\alpha)$
$[\,\frac{1}{2}, 1\,]$	$q_{X\|Z=1}\,(2\alpha-1)$	$t_M^{X\|Z=1}$	$q_{Y\|Z=0}\,(2\alpha-1)$	$t_M^{Y\|Z=0}$



(a)

(b)

(c)

(d)

Figure 4: Identifiability bounds of quantiles for different relative sizes

defined by (2.2); iv) *equity*: if X and Y have been equated, then the administered form should not be a concern for test takers; and v) *group invariance*: the equating function φ should be invariant if score data from different groups in the population are used for estimation.

The concept of equity is of special relevance in the NEAT design context: it should not matters to examinees which test form they are administered, provided that both forms are equated. Equated scores ensure that both `X`-scores and `Y`-scores are interchangeably, that is, for any $x \in \mathcal{X}$ it corresponds a unique $y \in \mathcal{Y}$ and conversely. Nevertheless, the lack of identifiability of both F_X and F_Y, inherent to the NEAT design, jeopardizes equity by at least two reasons:

1. If we take a `X`-score and we transform it by using φ to the `Y`-scale, there is not a *unique* equivalent `Y`-score. Thus, for instance, if $P(Z = 1) < P(Z = 0)$, an α-quantile for $\alpha \in (P(Z = 1),\ P(Z = 0))$ in the `Y`-scale is equivalent to all the `X`-scores. From an equity perspective, this is rather an unfair situation because the `Y`-score of an examinee can be transformed to the worst `X`-score, to the better `X`-score, or to any other `X`-score: how should one of them be chosen in order to ensure equity? It is a challenge to find a reasonable criterion leading to choose *one* score.

2. The severity of the identification problem is quantified by the relative size of population $\mathcal{P}$ and $\mathcal{Q}$. More specifically, the severity for F_X is quantified by $P(Z = 0)$, whereas the severity for F_Y is quantified by $P(Z = 1)$; see Theorem 4.1. Thus, for instance, if we consider Figure 2(a), we can conclude that if we transform Y-scores to X-scores, we increase the uncertainty inherent to the Y-scale; if we do the converse transformation, we decrease the uncertainty inherent to the X-scale (see Figure 2(b)). This result can be considered as a criterion to choose in which direction the equating transformation should be performed. Nevertheless, it also shows that the uncertainty can *not* be decreased arbitrarily because it depends on the relative sizes of populations $\mathcal{P}$ and $\mathcal{Q}$: it is actually a matter of *uncertainty by design*. Moreover, it is palatable that the lack of symmetry is due to the fact that $P(Z = 0) \neq P(Z = 1)$. When both relative sizes are equal, symmetry seems to be recovered (see Figure 2(c)), but the uncertainty by design persists.

A way to overcome the inherent identification problem is to introduce additional assumptions leading to decrease the width of the partial identification intervals. These assumptions reflect what the modeler thinks about the behavior of examinees when they take the test forms. When these assumptions are combined with the observations, it will be possible to assess their impact in the sense that it will illustrate to what extent the inference depends largely on such assumptions and not just the observations. This is discussed in the next section.

4 NEAT design under a self-selection process

4.1 Partial identification of the parameters of interest

One of the practical aspects of the NEAT design is the possibility that each examinee can decide when to take a form of a test. This raises a problem, namely how to model a process of self-selection in the personal choice of the test form. A plausible assumption to represent a “rational choice” of a test form is to assume that those who choose to take the form X do so because they believe they will score more than t more likely than if they had chosen form Y, which can be written as

$$P(X > t \mid Z = 1) > P(Y > t \mid Z = 1). \tag{4.1}$$

Similarly, those who take the form Y do so because they believe they will score more than t more likely than if they had chosen form X, which can be written as

$$P(Y > t \mid Z = 0) > P(X > t \mid Z = 0). \tag{4.2}$$

Two remarks are pertinent for the assumptions (4.1) and (4.2):

1. These assumptions means that the observed behavior correspond to the better choice examinees have made.
2. These assumptions use the fact that both $\mathcal{X}$ and $\mathcal{Y}$ are of the same nature, as pointed out in Section 2.3. In the case of modelling a self-selection process, this is important because the decision is taken with respect to objects of a similar nature, the only difference being for instance the time at which an examinee takes a test form.
3. These conditions implicitly assume that the equating problem in itself is invisible to examinees: they suppose that a score equal to t is similar for both forms. It can accordingly be said that examinees take a decision with partial information.

Combining (2.5) and (2.6) with the assumptions (4.1) and (4.2), we obtain the following theorem:

Theorem 4.1 *In the NEAT design, under the assumptions of self-selection (4.1) and (4.2), the parame-*

ters of interest F_X and F_Y are partially identified by the following intervals: for all t $\in \mathcal{W} \supset \mathcal{X} \cup \mathcal{Y}$,

$$\text{(i)} \qquad F_{X|Z=1}(t)P(Z=1) + F_{Y|Z=0}(t)P(Z=0) \leq F_X(t) \leq F_{X|Z=1}(t)P(Z=1) + P(Z=0), \tag{4.3}$$

$$\text{(ii)} \qquad F_{X|Z=1}(t)P(Z=1) + F_{Y|Z=0}(t)P(Z=0) \leq F_Y(t) \leq F_{Y|Z=0}(t)P(Z=0) + P(Z=1).$$

This theorem deserves some comments:

1. The upper bounds for both intervals are the same as the ones derived for the case of no self-selection (see Theorem 3.1).
2. Both intervals have a common lower bound. This bound corresponds to the actual proportion of examinees who scored at most t in test form X or Y. This means that, if all students had chosen the form X or Y under the optimistic self-selection assumptions (4.1) and (4.2), the probability to score at most t is at least equal to such proportion.
3. The identification intervals (4.3) improve the lower bounds of the identification intervals (3.4). More specifically:
   (a) The width of the interval (4.3.i) is equal to

   $$P(Y > t, Z = 0) = P(Y > t \mid Z = 0)\, P(Z = 0) \quad t \in \mathcal{W}; \tag{4.4}$$

   that is, the proportion of examinees that take the form Y and scored higher than t. Given that $P(Y > t \mid Z = 0) \in [0, 1]$, then the resulting interval is narrower than the interval shown in (3.4.i). Therefore, the optimistic self-selection assumptions allow to improve the partial identification of F_X. Finally, the width (4.4) is not constant, but it is a function of t.

   (b) The width of the interval (4.3.ii) is equal to

   $$P(X > t, Z = 1) = P(X > t \mid Z = 1)\, P(Z = 1) \quad t \in \mathcal{W}; \tag{4.5}$$

   that is, the proportion of examinees that take the form X and scored higher than t. Given that $P(X > t \mid Z = 1) \in (0, 1)$, then the resulting interval is narrower that the interval shown in (3.4.ii). Therefore, the optimistic self-selection assumptions allow to improve the partial identification of F_X. Finally, the width (4.5) is not constant, but it is a function of t.

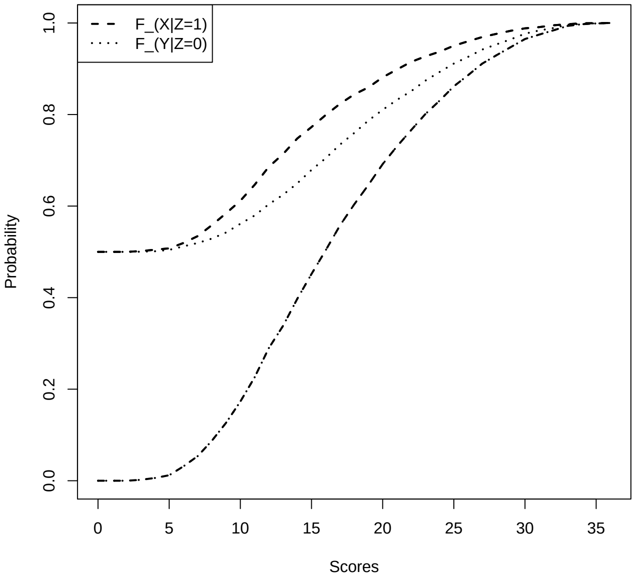


Figure 5: Identifiability bounds under a self-selection process (for the lower bounds, the point and segmented curves are superimposed)

The improvement of the identification intervals is graphically illustrated in Figure 5 for the case when the relative sizes of the populations are the same. When compared to Figure 2(c), it can be seen that both identification intervals are narrower. Note that the identification bounds are more accurate in the upper part of the score scale, which is coherent with the assumption that one would score better in X than in Y with larger probability.

4. Let $t \in \mathcal{W}$ be fixed. If the width of the interval (4.3.i) is less than the width of the interval (4.3.ii), namely
$$P(Y > t, Z = 0) < P(X > t, Z = 1),$$
then the interval (4.3.i) is contained in the interval (4.3.ii), and conversely.

4.2 Partial identification of the quantiles

Let us discuss the partial identification of the quantiles under the optimistic self-selection assumption. Let $\alpha \in [0,1]$ and define

$$\begin{aligned} u_X(\alpha) &\doteq \inf\{t : F_{X|Z=1}(t)P(Z=1) + P(Z=0) \geq \alpha\}; \\ u_Y(\alpha) &\doteq \inf\{t : F_{Y|Z=0}(t)P(Z=0) + P(Z=1) \geq \alpha\}; \\ v(\alpha) &\doteq \inf\{t : F_{X|Z=1}(t)P(Z=1) + F_{Y|Z=0}(t)P(Z=0) \geq \alpha\}. \end{aligned}$$

Using arguments similar to those used in Section 3.2.2, the following theorem follows:

Theorem 4.2 *In the NEAT design, under the optimistic self-selection assumptions (4.1) and (4.2), the quantiles of the partially identified probability distributions F_X and F_Y are partially identified by the following intervals:*

$$\begin{aligned} &\textit{(i)} \quad u_X(\alpha) \leq q_X(\alpha) \leq v(\alpha); \\ &\textit{(ii)} \quad u_Y(\alpha) \leq q_Y(\alpha) \leq v(\alpha). \end{aligned} \tag{4.6}$$

One of the conclusions of Theorem 3.8 was that the partial identified quantiles are not always informative. For instance, when $P(Z=1) < P(Z=0)$, the upper bound $s_X(\alpha)$ of $q_X(\alpha)$ is equal to $t_M^{X|Z=1}$ for all $\alpha > P(Z=1)$; see Table 1; or when $P(Z=1) > P(Z=0)$, the upper bound $s_Y(\alpha)$ of $q_Y(\alpha)$ is equal to $t_M^{Y|Z=0}$ for all $\alpha > P(Z=0)$; see Table 2. Under the optimistic self-selection assumption, this situation is improved for the upper bound. As a matter of fact, the identification intervals (4.3.i) and (4.3.ii) of $q_X(\alpha)$ and $q_Y(\alpha)$, respectively, depend on the quantile $v(\alpha)$, which in turn corresponds to the quantile of the distribution of getting an score equal to t either in form X or in form Y. This distribution corresponds to a mixture. Following Bernard and Vanduffel (2015), it is possible to express the quantile of the mixture in terms of $F^{-1}_{X|Z=1}(\alpha)$ and $F^{-1}_{Y|Z=0}(\alpha)$: let $\alpha \in [0,1]$ and define $\delta_* \in [0,1]$ by

$$\delta_* = \inf\left\{\delta \in (0,1) \ : \ \exists\, \epsilon \in (0,1) \ \text{ s.t. } P(Z=1)\,\delta + P(Z=0)\,\epsilon = \alpha, \ \ F^{-1}_{X|Z=1}(\delta) \geq F^{-1}_{Y|Z=0}(\epsilon)\right\} \tag{4.7}$$

and $\epsilon_* \in [0,1]$ by

$$\epsilon_* = \frac{\alpha - P(Z=1)\delta_*}{P(Z=0)}. \tag{4.8}$$

Then

$$v(\alpha) = \max\left\{F^{-1}_{X|Z=1}(\delta_*), F^{-1}_{Y|Z=0}(\epsilon_*)\right\}.$$

Table 4: Lower and upper bounds for $q_X(\alpha)$ and $q_Y(\alpha)$ when $P(Z = 1) < P(Z = 0)$ under the optimistic self-selection assumptions; δ_* and ϵ_* are defined by (4.7) and (4.8), respectively

α	$u_X(\alpha)$	$v(\alpha)$	$u_Y(\alpha)$	$v(\alpha)$
$[\,0, P(Z=1)\,]$	$t_m^{X\mid Z=1}$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$	$t_m^{Y\mid Z=0}$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$
$(\,P(Z=1), P(Z=0)\,)$	$t_m^{X\mid Z=1}$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$	$q_{Y\mid Z=0}\left(\frac{\alpha-P(Z=1)}{P(Z=0)}\right)$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$
$[\,P(Z=0), 1\,]$	$q_{X\mid Z=1}\left(\frac{\alpha-P(Z=0)}{P(Z=1)}\right)$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$	$q_{Y\mid Z=0}\left(\frac{\alpha-P(Z=1)}{P(Z=0)}\right)$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$

Table 5: Lower and upper bounds for $q_X(\alpha)$ and $q_Y(\alpha)$ when $P(Z = 1) > P(Z = 0)$ under the optimistic self-selection assumptions; δ_* and ϵ_* are defined by (4.7) and (4.8), respectively

α	$u_X(\alpha)$	$v(\alpha)$	$u_Y(\alpha)$	$v(\alpha)$
$[\,0, P(Z=0)\,]$	$t_m^{X\mid Z=1}$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$	$t_m^{Y\mid Z=0}$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$
$(\,P(Z=0), P(Z=1)\,)$	$q_{X\mid Z=1}\left(\frac{\alpha-P(Z=0)}{P(Z=1)}\right)$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$	$t_m^{Y\mid Z=0}$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$
$[\,P(Z=1), 1\,]$	$q_{X\mid Z=1}\left(\frac{\alpha-P(Z=0)}{P(Z=1)}\right)$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$	$q_{Y\mid Z=0}\left(\frac{\alpha-P(Z=1)}{P(Z=0)}\right)$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$

This equality shows that $v(\alpha)$ will be informative, as can be seen in Tables 4, 5 and 6.

Summarizing, under the optimistic self-selection assumptions (4.1) and (4.2), the partial identification intervals of the parameters of interest are better than the corresponding intervals derived in Sections 3.2.1 and 3.2.2: they are better because their width is lesser and they are more informative for the upper bounds of the identification intervals of the quantiles.

5 Discussion

The objective of statistical modelling is to specify the probability distribution that generates the *observables*. This modelling process corresponds to a combination of evidence (the observables, the data) with

Table 6: Lower and upper bounds for $q_X(\alpha)$ and $q_Y(\alpha)$ when $P(Z = 1) = P(Z = 0)$ under the optimistic self-selection assumptions; δ_* and ϵ_* are defined by (4.7) and (4.8), respectively

α	$u_X(\alpha)$	$v(\alpha)$	$u_Y(\alpha)$	$v(\alpha)$
$\left[0, \frac{1}{2}\right)$	$t_m^{X\mid Z=1}$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$	$t_m^{Y\mid Z=0}$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$
$\left[\frac{1}{2}, 1\right]$	$q_{X\mid Z=1}\,(2\alpha-1)$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$	$q_{Y\mid Z=0}\,(2\alpha-1)$	$\max\left\{F^{-1}_{X\mid Z=1}(\delta_*), F^{-1}_{Y\mid Z=0}(\epsilon_*)\right\}$

researcher's ideas on the explanation or formation of the phenomenon studied, which in turn are assumptions about unobserved quantities. "Knowledge is the set of conclusions that one draws by combining evidence with assumptions about unobserved quantities" (Manski, 2013). The content of this paper is precisely in this line and tries to make explicit the *knowledge* we obtain by combining the scores provided by examinees under the NEAT design (the evidence provided by the phenomenon under analysis) with three type of assumptions leading to solve the identification problem inherent to the NEAT design: strong ignorability (discussed in Section 3.1), no assumption regarding the unobserved quantities (discussed in Sections 3.2.1 and 3.2.2) and the optimistic self-selection assumption (discussed in Section 4).

The modelling process consisted in specifying the statistical model (developed in Section 2.4.1), that is, to specify (i) a set of probability distributions generating the observations (we call them *sampling probabilities*), (ii) making explicit their parameters, and (iii) pointing out the parameters of interest. Following Fisher (1922), the parameters of the sampling probabilities are characteristics of the observations under study. However, the researcher focuses the attention on additional characteristic of interest (represented by parameters of interest) and, therefore, the question is to know if such characteristics can be expressed as functional of the sampling probabilities or not: when this is not possible, we face an identification problem.

When conducting equating under the NEAT design, the parameters of interest are F_X and F_Y: our modelling strategy allows us to make explicit their meaning as well as their lack of identifiability. We suppose the researcher knows that those distributions are necessary to define the equating function (see Section 2.3), and that it is possible to *explicitly show* why those parameters can not be derived from the statistical model, which is done through the decompositions (2.7) and (2.8): the lack of identifiability of F_X and F_Y is due to the lack of identifiability of $F_{X|Z=0}$ and $F_{Y|Z=1}$ –which in turn follows from the lack of identifiability of $F_{X|A,Z=0}$ and $F_{Y|A,Z=1}$.

Decompositions (2.7) and (2.8) are a key step that surprisingly has not been followed in the equating literature, even though it is recognized that the lack of indentifiability of $F_{X|A,Z=0}$ and $F_{Y|A,Z=1}$ can be considered as a missing data problem; see Holland et al. (2008) and Sinharay and Holland (2010). The missing data problem and, more in general, the selection problem becomes palatable after using the Law of Total Probability: it allows us to correctly relates F_X and F_Y with the identified conditional probability distributions, some of them being unidentified. Consequences of this key step are twofold: on the one hand, to show that the notion of *target population* is meaningless; on the other hand, to show that the target distributions are arbitrary and therefore difficult to interpret with respect to the statistical model underlying the NEAT design.

What is the knowledge we get when we combine the evidence (examinees' scores) with assumptions on the unobserved quantities? If we are ready to believe in the strong ignorability condition, then $F_{X|A}$ and $F_{Y|A}$ becomes point identified and, in fact, are equal to $F_{X|A,Z=1}$ and $F_{Y|A,Z=0}$. From these distributions, we obtain F_X and F_Y, and the equating function can be computed. Note that in this case the relative sizes of populations $\mathcal{P}$ and $\mathcal{Q}$ do not play any role. Under the strong ignorability condition, equity is ensured, which means that the price to pay for getting equity is to accept strong ignorability.

If no assumption regarding the unobserved quantities is made, we realize how severe is the lack of identifability of F_X and F_Y. The partial identification intervals show what we learn from the evidence in absence of assumptions regarding the unidentified parameters $F_{X|Z=0}$ and $F_{Y|Z=1}$. In particular, the *uncertainty by design* emerges and the role of the relative sizes of $\mathcal{P}$ and $\mathcal{Q}$ in such uncertainty becomes explicit: it involves a trade-off in the sense that the larger $P(Z = 1)$ (respect., $P(Z = 0)$ is, the less precise F_Y (respect., F_X) is. The consequence of these findings is that equity is lost, which means that equity is *not a property inherent to the NEAT design*.

If an optimistic self-selection assumption is introduced, we learn something additional from the evidence (examinees' scores) because the width of the partial identification intervals decreases. What we learn is that the identification bounds are more accurate in the upper part of the score scale, which is coherent with the assumption that one would score better in X than in Y with larger probability. Unfortunately, equity is not recovered.

Scientific knowledge is obtained when evidence is combined with assumptions about unobserved quantities. Such assumptions become relevant when an identification problem is present.A constructive modelling process like the one developed in this article is relevant because it makes it explicit to what extent scientific knowledge is highly dependent on those assumptions. Psychometrics needs to travel these avenues in order to be honest (Pielke Jr, 2007).

Acknowledgements The first author acknowledges the partial financial support of the FONDECYT Project No. 1181261. Preliminary versions of this research were presented by the first author at the Brazilian Conference CONBRATRI VI, Juiz de Fora, Brasil (August, 2018); by the second author at the IMPS 2017, Zurich, Switzerland; at the Iowa Equating Summit, 2017, IA; and at the IMPS 2019, Santiago, Chile.

References

Alarcón-Bustamante, E., San Martín, E., & González, J. (in press). Predictive validity under partial observability. In M. Wiberg, D. Molenaar, J. González, U. Bockenholt, & J.-S. Kim (Eds.), *Quantitative psychology* (pp. xxx–xxx). Springer.

Angoff, W. H. (1984). *Scales, Norms, and Equivalent Scores*. New Jersey: Educational Testing Service, Princeton.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Bernard, C., & Vanduffel, S. (2015). Quantile of a Mixture with Application to Model Risk Assessment. *Dependence Modeling*, *3*, 172–181.

Blundell, R., Gosling, A., Ichimura, H., & Meghir, C. (2007). Changes in the Distribution of Male and Female Wages Accounting for Emplyement Composition Using Bounds. *Econometrica*, *75*, 323–363.

Bolsinova, M., & Maris, G. (2016). Can irt solve the missing data problem in test equating? *Frontiers in Psychology*, *6*, 1956.

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

Brennan, R. L., & Kolen, M. J. (1987). A reply to angoff. *Applied Psychological Measurement*, *11*(3), 301-306.

Cochran, W. G., & Chambers, S. (1965). The Planning of Observational Studies of Human Populations. *Journal of the Royal Statistical Society, Series A*, *128*(2), 234–266.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical Statistics*. Chapman and Hall/CRC.

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

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A*, *222*, 309–368.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1973). *Statistical Methods for Research Workers*. Hafner Publishhing, New York, USA.

González, J., & Wiberg, M. (2017). *Applying Test Equating Methods Using R*. New Yok: Springer.

Gulliksen, H. (1950). *Theory of Mental Tests*. John Wiley and Sons, Inc.

Gundersen, C., & Kreider, B. (2008). Food Stamps and Food Insecurity: What Can Be Learned in the Presence of Nonclassical Measurement Error. *The Journal of Human Resources*, *43*, 352–382.

Gundersen, C., & Kreider, B. (2009). Bounding the effects of food insecurity on children's health outcomes. *Journal of Health Economics*, *28*, 971–983.

Gundersen, C., Kreider, B., & Pepper, J. (2012). The impact of the National School Lunch Program on child health: A nonparametric bounds analysis. *Journal of Econometrics*, *166*, 79–91.

Holland, P. W., Dorans, N. J., & Petersen, S. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics 26. Psychometrics* (pp. 169–204). Amsterdam: North Holland.

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the neat design. *Journal of Educational Measurement*, *45*(1), 17-43.

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

Kolen, M. J., & Hendrickson, . B. (2011). Scalinf, Norms, and Equating. In C. Secolsky & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 161–177). London: Routledge.

Kolmogorov, A. N. (1956). *Foundations of the Theoty of Probability*. New Yok: Chelsea Publishing Company.

Kreider, B., & Pepper, J. (2007). Disability and employment: reevaluating the evidence in light of reporting errors. *Journal of the American Statistical Association*, *102*, 432–441.

Liou, M., & Cheng, P. E. (1995). Equipercentile Equating Via Data-Imputation Technique. *Psychometrika*, *60*, 119-136.

Lord, F. M. (1950). *Notes on Comparable Scales for Test Scores* (Tech. Rep.). Educational Testing Service.

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

Maddala, G. (1983). *Qualitative and limited dependent variable models in econometrics*. Cambridge: Cambridge University Press.

Manski, C. (1995). *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

Manski, C. (2003). *Partial identification of probability distributions*. New York: Springer.

Manski, C. (2007). *Identification for Prediction and Decision*. Cambridge: Harvard University Press.

Manski, C. (2013). Diagnostic testing and treatment under ambiguity: using decision analysis to inform clinical practice. *Proceedings of the National Academy of Sciences*, *110*, 2064–2069.

Manski, C., & Pepper, J. (2013). Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections. *Journal of Quantitative Crimonology*, *29*, 123–141.

McCullagh, P. (2002). What is a Statistical Model? *Annals of Statistics*, *30*, 1225–1267.

Miyazaki, K., Hoshino, T., Mayekawa, S.-i., & Shigemasu, K. (2009). A new concurrent calibration method for nonequivalent group design under nonrandom assignment. *Psychometrika*, *74*(1), 1.

Molinari, F. (2010). Missing Treatments. *Journal of Business & Economic Statistics*, *28*, 82–95.

Pepper, J. (2000). The Intergenerational Transmission of Welfare Receipt: A Nonparametric Bounds Analysis. *The Review of Economics and Statistics*, *82*, 472–488.

Pielke Jr, R. A. (2007). *The Honest Broker: Making Sense of Science in Policy and Politics*. Cambridge University Press.

Rao, M. M. (2005). *Conditional Measures and Applications*. Chapman and Hall/CRC.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41–55.

San Martín, E. (2016). Identification of Item Response Theory Models. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory. Volumen Two: Statistical Tools* (pp. 127–150). London: CRC Press.

San Martín, E. (2018). Identifiability of Structural Characteristics: How Relevant is it for the Bayesian Approach? *Brazilian Journal of Probability and Statistics*, *32*, 346–373.

San Martín, E., González, J., & Tuerlinckx, F. (2015). On the Unidentifiability of the Fixed-effects 3PL Model. *Psychometrika*, *80*, 450–467.

Shaked, M., & Shanthikumar, J. G. (2007). *Stochastic Orders*. New York: Springer.

Sinharay, S., & Holland, P. W. (2010). The Missing Data Assumptions of the NEAT Design and Their Implications for Test Equating. *Psychometrika*, *75*, 309–327.

Stoye, J. (2010). Partial identification of spread parameters. *Quantitative Economics*, *1*, 223–257.

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