Journal of Educational and Behavioral Statistics
Vol. XX, No. X, pp. 1–32
DOI: 10.3102/10769986221090609
Article reuse guidelines: sagepub.com/journals-permissions
© 2022 AERA. https://journals.sagepub.com/home/jeb

A Critical View on the NEAT Equating Design: Statistical Modeling and Identifiability Problems

Ernesto San Martín
Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI), Chile
Faculty of Mathematics, Pontificia Universidad Católica de Chile, Chile
Interdisciplinary Laboratory of Social Statistics, Chile
The Economics School of Louvain, Université Catholique de Louvain, Belgium

Jorge González
Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI), Chile
Faculty of Mathematics, Pontificia Universidad Católica de Chile, Chile
Interdisciplinary Laboratory of Social Statistics, Chile

The nonequivalent groups with anchor test (NEAT) design 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 partially observed score data formally lead to an identifiability problem, which has not been recognized as such in the equating literature and has been considered from different perspectives, all of them making different assumptions in order to estimate the unidentified score distributions. In this article, 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 show alternatives to traditional practices that have been proposed to identify the score distributions when conducting equating under the NEAT design.

Keywords: *statistical models; strong ignorability; partial identifiability*

1. Introduction

Test equating is conducted to adjust the score scales 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; Lord, 1950). The *equating problem* consists in mapping scores defined on one scale 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). Although different types of score linkages are distinguished in the literature (Holland & Dorans, 2006), with equating being a particular case of score linking, the same linking function can be used in each case (Kolen, 2007) leading to scores that are interpreted as either interchangeable (equating) or comparable (linking).

Because score differences can also arise 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 have been proposed in the literature, leading to what are called equating designs (see, e.g., González & Wiberg, 2017; Kolen & Brennan, 2014; von Davier et al., 2004). These designs differ in that either common persons or common items are used to perform the score transformation and in the conditions and assumptions made to collect the score data. In this article, we will focus the attention on the nonequivalent groups with anchor test design (NEAT).

The NEAT design (also known as the common item nonequivalent group design) is widely used in test equating. Under this design, two groups of test takers are administered different test forms that are intended to measure a common variable, with each test form containing a subset of common items. Because test takers from different groups respond to only one test form, *missing* score data emerge by design rendering some of the score distributions unavailable. The partially observed score data formally lead to an identifiability problem, which has not been recognized as such in the equating literature and has been considered from different perspectives, all of them making different assumptions in order to estimate the unidentified score distributions (see, e.g., Holland et al., 2008; Miyazaki et al., 2009; Sinharay & Holland, 2010).

In this article, we specify the statistical model underlying the NEAT design using a fully probabilistic approach. By doing so, we not only show that the notion of synthetic population is meaningless but also offer an alternative formal and useful conceptualization of it. One of the consequences of this conceptualization is that the population weights used in the definition of a synthetic population cannot be arbitrarily chosen. Moreover, using a partial identification approach (Manski, 2007; Tamer, 2010), we make explicit the devastating results implied by the identification problem that are hidden when the missing at random condition is assumed.

This article is organized as follows: In Section 2, current definitions and assumptions considered when conducting equating under the NEAT design are revised. Next, we criticize the aspects of the current view on NEAT equating, make explicit the statement of the problem to be discussed, and explain the strategy used in the article. In Section 3, the statistical model underlying the NEAT design is defined and the identifiability problem is described. In Section 4, we describe two approaches to tackle the identifiability problem and study the impact of the lack of identifiability in the actual equating using the theory of quantiles. An empirical illustration of the main findings is also presented at the end of this section. In the context of score linking, in Section 5, we discuss alternative identification restrictions leading to improve the previous results. This article finalizes in Section 6, summarizing the main points and discussing on how severe the identifiability problem can be in the context of test equating.

2. Modeling Problems Underlying the NEAT Design

In order to show what motivates our critique of the NEAT design, we first review the basic definitions and strategies given in the literature for estimating the corresponding equating transformation.

2.1. Equating Under the NEAT Design: A Summary of the Literature

Let X and Y be random variables representing the scores on tests forms X and Y, respectively. The equating problem is addressed by using an *equating function* $\varphi : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the sets of all possible scores on test forms X and Y (the score scales), so that for each raw score x, we can compute the corresponding $y = \varphi(x)$ (Braun & Holland, 1982; González & Wiberg, 2017). Following Lord (1950) and Angoff (1984), a function φ can be defined taking into account that “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). More specifically, the function $\varphi : \mathcal{X} \to \mathcal{Y}$ is defined as follows: For each $x \in \mathcal{X}$, we associate $y \doteq \varphi(x) \in \mathcal{Y}$, such that

$$F_Y[\varphi(x)] = F_X(x), \tag{2.1}$$

the symbol $\doteq$ means “defined as.” Although Equation 2.1 can always be solved using generalized inverses (Embrechts & Hofert, 2013), meaningful equating results are obtained only when scores are continuous (van der Linden, 2019). This is why all practical applications of test equating make use of what is called a continuization step (for details, see Braun & Holland, 1982; González & Wiberg, 2017; Kolen & Brennan, 2014).

Consider now a third random variable, A, which represents the scores obtained on an anchor test form A. Two definitions of the NEAT design appearing in the standard literature on equating are the following:

> In the Non-Equivalent groups with Anchor Test (NEAT) Design there are two populations, $\mathcal{P}$ and $\mathcal{Q}$, of test-takers and a sample of examinees from each. The sample from $\mathcal{P}$ takes test X, the sample from $\mathcal{Q}$ takes test Y, and both samples take a set of common items, the anchor test A. (von Davier et al., 2004, p. 32)
>
> For [the NEAT] design, two groups of examinees from different populations are each administered different test forms that have a set of items in common. (Kolen & Brennan, 2014, p. 103)

According to Braun and Holland (1982), the function φ should be defined with respect to a single population. This is a challenging task when performing equating under the NEAT design as, in fact, two populations, $\mathcal{P}$ and $\mathcal{Q}$, are involved. Consequently, $\mathcal{P}$ and $\mathcal{Q}$ "must be combined to obtain a single population for defining [the] equating function" (Kolen & Brennan, 2014, p. 103), which leads to introduce a *synthetic population*, also called *target population*, denoted by $\mathcal{T}$. This is done by "conceptualizing a larger population that has $\mathcal{P}$ and $\mathcal{Q}$ as two mutually and exhaustive strata" (Braun & Holland, 1982, p. 21), a "combined group, some of whom would be taking Form X, other taking Form Y, and all taking Form A, the anchor test" (Angoff, 1987, p. 296), and "a mixture of both $\mathcal{P}$ and $\mathcal{Q}$" (von Davier et al., 2004, p. 34). Each of these statements is assumed to be represented by the following expression:

$$\mathcal{T} = \omega\mathcal{P} + (1 - \omega)\mathcal{Q}, \tag{2.2}$$

where the populations $\mathcal{P}$ and $\mathcal{Q}$ are weighted by $\omega \in [0, 1]$ and $1 - \omega$, respectively. However, $\mathcal{P}$ and $\mathcal{Q}$ are just labels denoting the two populations, and thus, expression 2.2 would represent a "weighted average of labels" which, of course, does not make any sense. A consequence of using this expression is that the target score distributions are arbitrarily chosen, as it will be seen in Section 3.

The equating transformation is then typically computed after defining the score distributions of X, Y, and A on the synthetic population $\mathcal{T}$. Thus, for instance, for a specific and arbitrary choice of ω, the distribution $F_{X\mathcal{T}}(x)$, which has also been denoted in the literature as $P(X \leq x|\mathcal{T})$ (Sinharay & Holland, 2010; von Davier et al., 2004, p. 7), is obtained by first computing the distribution of raw scores X for each strata $\mathcal{P}$ and $\mathcal{Q}$ and then forming weighted averages of these distributions using ω and $1 - \omega$, namely

$$F_{X\mathcal{T}}(x) \doteq \omega F_{X\mathcal{P}}(x) + (1 - \omega)F_{X\mathcal{Q}}(x). \tag{2.3}$$

The score distributions $F_{X\mathcal{P}}(x)$ and $F_{X\mathcal{Q}}(x)$ have been denoted in the literature as $P(X \leq x|\mathcal{P})$ and $P(X \leq x|\mathcal{Q})$, respectively, with similar notation used for $F_{Y\mathcal{T}}(y)$ and $F_{A\mathcal{T}}(a)$. From the data collected under the NEAT design, the distributions $F_{X\mathcal{P}}(x), F_{Y\mathcal{Q}}(y), F_{A\mathcal{P}}(a)$, and $F_{A\mathcal{Q}}(a)$ may be computed. Nevertheless, the distributions $F_{X\mathcal{Q}}(x)$ and $F_{Y\mathcal{P}}(y)$ are not directly estimable. They are typically obtained by conditioning on the anchor score A and using the following assumptions:

$$(i) \quad F_{X\mathcal{P}|A}(x|a) = F_{X\mathcal{Q}|A}(x|a); \qquad (ii) \quad F_{Y\mathcal{P}|A}(y|a) = F_{Y\mathcal{Q}|A}(y|a), \tag{2.4}$$

where $F_{X\mathcal{P}|A}$ is the distribution function of X conditionally on A in $\mathcal{P}$; and the other distributions have a similar meaning. Under these assumptions, the target score distributions are given by

$$\begin{aligned} &(i) \quad F_{X\mathcal{T}}(x) = \omega F_{X\mathcal{P}}(x) + (1 - \omega)\sum_{a\in\mathcal{A}} F_{X\mathcal{P}|A}(x|a)P(A = a|\mathcal{Q}), \\ &(ii) \quad F_{Y\mathcal{T}}(y) = \omega\sum_{a\in\mathcal{A}} F_{Y\mathcal{Q}|A}(y|a)P(A = a|\mathcal{P}) + (1 - \omega)F_{Y\mathcal{Q}}(y), \end{aligned} \tag{2.5}$$

where $\mathcal{A}$ is the set of possible anchor-test scores. These score distributions are then used to compute $\varphi(x)$ for a given $x \in \mathcal{X}$ using the equality $\varphi(x) = F_{Y\mathcal{T}}^{-1}[F_{X\mathcal{T}}(x)]$. This method is called frequency estimation equating (FEE; Angoff, 1984). Other methods used for equating under the NEAT design, such as the chained equipercentile equating and item response theory-observed score equating (IRT-OSE), together with the assumptions needed in each case to obtain the distributions that are not directly estimable are described in Sinharay and Holland (2010).

2.2. What We Question: A Statement of the Problem

From a modeling point of view, equating under the NEAT design depends on both the synthetic population (2.2) and the assumptions in (2.4), which are typically known as the *missing-at-random assumptions* (Liou & Cheng, 1995; Miyazaki et al., 2009; Sinharay & Holland, 2010). Bearing in mind that the equating function is defined in probabilistic terms, the following questions arise:

1. Taking into account that $\mathcal{P}$ and $\mathcal{Q}$ are not random variables but mere labels, under which conditions are $P(X \leq x|\mathcal{P})$ and $P(X \leq x|\mathcal{Q})$ properly understood as conditional probability functions?
2. If such conditions exist, would it be possible to define the concept of a synthetic population in probabilistic terms?
3. Is it possible to provide alternative assumptions to the ones given in (2.4) to solve the problem of missing score data underlying the NEAT design? If so, what would be the impact of such alternative assumptions on the NEAT design in specific contexts?

The answers to the previous questions depend on the answer to a more fundamental one: How can “an understanding of the way in which the data” collected under the NEAT design “are supposed to, or did in fact, originate” be obtained (Fisher, 1973, p. 8)? This question can be answered once the parameters of interest that “describe [the observations] exhaustively in respect of all qualities under discussion” (Fisher, 1922, p. 311) are made explicit. Nevertheless, the data generating process is not necessarily characterized by the parameters of interest. When this is the case, we face an identifiability problem. This is the problem we will unveil in the NEAT design.

3. Statistical Specification of the NEAT Design and the Inherent Identification Problem

One way to answer these questions is to follow González and Wiberg’s (2017) perspective, which consists in specifying the NEAT design as a *statistical model*. This means to make explicit the following three components:

1. The *sample space* $(M, \mathcal{M})$, where M contains the elementary events and $\mathcal{M}$ is a class containing the events of interest and their combinations through unions of sets as well as complement of a set—this is why $\mathcal{M}$ is a σ-field of subsets of M. The pair $(M, \mathcal{M})$ is technically called a measurable space.
2. The *sampling probabilities* P^γ, indexed by a parameter γ, where each P^γ is defined on $(M, \mathcal{M})$.
3. The *parameter space* Γ, which represents the set of all plausible values of γ.

For details, see San Martín et al. (2015) and references therein.

What are the practical advantages of this formalism? First, the population of interest is made explicit by listing it through the elements of the sample space. Second, features of the population under study are fully characterized by the parameters indexing the sampling probabilities. Third, this formalism allows us to evaluate whether certain characteristics relevant to researchers can indeed be represented as (a function of) parameters: When this is not the case, then we face an identification problem (Koopmans, 1949).

The specification of the statistical model underlying the NEAT design will be performed in a sequential way, which leads to decompose the joint probability distribution defined on all the pertinent random variables into specific conditional submodels. The sequential specification makes explicit the way in which the data are (supposed to be) generated (Wunsch et al., 2014). This sequential specification is critically based on the Law of Total Probability in the sense that it ensures the *existence* of the conditional distribution of a new random variable given the previously introduced variables. This corresponds to the formal definition of a conditional probability function as introduced by Kolmogorov (1950, §6). By doing so, it will be made explicit which are the parameters of the statistical model and the lack of identifiability of the parameters of interest.

3.1. The Sample Space in the NEAT Design

In order to construct both the random variables and the probability distributions underlying the NEAT design, it is necessary to make explicit the sample space M. In the case of the NEAT design, M corresponds to the union of the index sets labeling the two mutually exclusive groups of examinees who were *exposed* to the test forms X and Y, namely

$$M = \mathcal{P} \cup \mathcal{Q}, \quad \text{where} \quad \mathcal{P} = \{i_1, \ldots, i_{n_\mathcal{P}}\}, \quad \mathcal{Q} = \{j_1, \ldots, j_{n_\mathcal{Q}}\}, \quad \mathcal{P} \cap \mathcal{Q} = \varnothing,$$

where $n_\mathcal{P}$ and $n_\mathcal{Q}$ are the total number of examinees exposed to test forms X and Y, respectively. Thus, M corresponds to the population of interest underlying the NEAT design. Since it is a finite set, the class $\mathcal{M}$ of events of interest corresponds to the class of subsets of M.

TABLE 1.
Toy Data Set Example

M	X	Y	A	Z
i_1	1	—	2	1
i_2	4	—	1	1
i_3	3	—	2	1
j_1	—	1	3	0
j_2	—	5	3	0
j_3	—	5	1	0
j_4	—	2	0	0
j_5	—	0	2	0

3.2. The Random Variable Z

All the random variables underlying the NEAT design should be defined from M to a set of possible values. In addition to the variables representing the test scores, namely, $X, Y,$ and A, we construct a random variable representing the groups underlying the NEAT design. To do so, remember that, by definition, given two measurable spaces $(M, \mathcal{M})$ and $(N, \mathcal{N})$, a function $f : M \rightarrow N$ is a random variable if $f^{-1}(B) \in \mathcal{M}$ for all $B \in \mathcal{N}$ (Karr, 1993; Kolmogorov, 1950). In the finite discrete case, as it is in the NEAT design, this formal definition reduces to the following characterization: A mapping of the set M into a set of values is a random variable if and only if the mapping induces a partition of M. For instance, for the case when $\mathcal{P} = \{i_1, i_2, i_3\}$ and $\mathcal{Q} = \{j_1, j_2, j_3, j_4, j_5\}$, consider the toy data set shown in Table 1. If $A : M \rightarrow \mathcal{A} \doteq \{0, 1, 2, 3\}$ is a random variable representing the scores obtained in an anchor test of three items, then $A^{-1}\{0\} = \{j_4\}, A^{-1}\{1\} = \{i_2, j_3\}, A^{-1}\{2\} = \{i_1, i_3, j_5\}$ and $A^{-1}\{3\} = \{j_1, j_2\}$. These sets belong to $\mathcal{M}$ and constitute a partition of M.

This characterization is relevant for at least two reasons: First, different examinees with equal values on a random variable will be treated as equivalent and put in the same equivalence class (i.e., the members of the partition) of examinees. Second, a random variable is the class of events in $\mathcal{M}$ (in the finite case, subsets of M) that is induced by the possible values the random variable can take (Florens & Mouchart, 1982), which means that all information provided by a random variable is captured in the sample space $(M, \mathcal{M})$.

The sample space M has been defined through the partition $\{\mathcal{P}, \mathcal{Q}\}$, which implies that a binary random variable Z can naturally be defined as

$$Z = \begin{cases} 1, & \text{if the examenee belongs to group } \mathcal{P}, \\ 0, & \text{if the examenee belongs to group } \mathcal{Q}. \end{cases} \tag{3.1}$$

That is, the preimage $Z^{-1}\{1\}$ corresponds to the event $\mathcal{P}$ and the preimage $Z^{-1}\{0\}$ to the event $\mathcal{Q}$: This is the semantic meaning of Z.

A natural first approach to construct a probability function of Z would be thinking in terms of experimental units (the members of the sample space M) and a random mechanism that assigns numbers to the experimental units, as, for instance, scores or measurements (Lord & Novick, 1968, Sections 1.3 and 2.2). Under this perspective, the groups $\mathcal{P}$ and $\mathcal{Q}$ could be viewed as "realizations" or "occurrences" of such a mechanism. However, using Little and Rubin's (1994) terminology, the groups $\mathcal{P}$ and $\mathcal{Q}$ in the NEAT design are considered as *selected samples* and, consequently, subject to selection bias. Thus, this approach is not useful to reveal some random assignment mechanism that makes sense of Z in the NEAT design.

Nevertheless, it is possible to propose an alternative approach that suits better the characteristics of the NEAT design. To do so, we resort to Carnap's (1962, §12) distinction between what a researcher says about a term in his formulation (semantics) and what he actually does with that term in the corpus of his formulation (syntax): the sense of a certain term is revealed from its use in formal arguments. Taking into account that, from an axiomatic perspective, "random" and "realization" or "occurrence" are mere terms, not probabilistic concepts (see Kolmogorov [1950, p. 2] and Itô [1984, pp. 1, 3], respectively), an axiomatically well-defined probability function of Z can be defined as follows: When $M = \{m_1, \dots, m_k\}$ is a finite set, it is enough to take an arbitrary set of non-negative numbers $\{p_1, \dots, p_k\}$ with their sum equal to 1, such that $P\{m_i\} = p_i$ for $i = 1, \dots, k$ (Kolmogorov, 1950, p. 3). Consequently, a possible probability function of Z can accordingly be constructed by considering the relative sizes of the groups $\mathcal{P}$ and $\mathcal{Q}$ (Braun & Holland, 1982), namely

$$P(Z = 1) = \frac{n_{\mathcal{P}}}{n_{\mathcal{P}} + n_{\mathcal{Q}}}, \qquad P(Z = 0) = \frac{n_{\mathcal{Q}}}{n_{\mathcal{P}} + n_{\mathcal{Q}}}. \tag{3.2}$$

Under this approach, both $\mathcal{P}$ and $\mathcal{Q}$ are viewed as constituting the population of interest rather than realizations from some sampling mechanism (see Manski, 2013, p. 126). This is due to the fact that we focus our attention on the identification problem underlying the NEAT design, which will be the same regardless of the type of sampling process it is reasonable to assume (see also Manski & Nagin, 1998, specially p .109). It is in fact possible to specify a different probability function for Z, but its identifiability should be justified, as it is the case for the one defined in Equation 3.2. Furthermore, in the identification analysis that will follow, the role played by the random variable Z is only through the *probability function that is defined on it.*

For its semantic meaning, the partition of M induced by Z can be interpreted as a "sample by specification" (Lord & Novick, 1968, Section 2.5), thus avoiding the idea of a possible random assignment mechanism. In any case, we define a "random assignment mechanism" in terms of "absence of bias," which in turn is rigorously described in terms of conditional probabilities (for details, see Supplemental Appendix A). This definition is actually motivated by a semantic relationship between both notions (Bhide et al., 2018; Odgaard-Jensen et al., 2011; Stephenson & Imrie, 1998). To the best of our knowledge, no formal proof exists deriving "absence of bias" from "random assignment" precisely because this last notion is not a probabilistic concept. As it will be seen in what follows, the identification problem arises from the selection bias inherent to the NEAT design, which formally means that $P(X \leq x|A, Z = 1) \neq P(X \leq x|A, Z = 0)$ and $P(Y \leq y|A, Z = 1) \neq P(Y \leq y|A, Z = 0)$. Before discussing this aspect, let us follow through the sequential specification of the NEAT design.

3.3. The Anchor Random Variable A

Once Z has been specified, we introduce a random variable A defined from M into the set $\mathcal{A}$ of possible scores in the anchor test. Following the sequential specification strategy, the existence of the conditional probability of A given Z is ensured by the following representation: for all $a \in \mathcal{A}$

$$\begin{aligned} P(A = a) &= P(A = a|Z = 1)P(Z = 1) + P(A = a|Z = 0)P(Z = 0), \\ &= E[P(A = a|Z = 1)1_{\{Z=1\}} + P(A = a|Z = 0)1_{\{Z=0\}}], \\ &= E[P(A = a|Z)]. \end{aligned} \tag{3.3}$$

It should be remarked that the existence of the conditional probability function $P(A = a|Z)$ depends on Z through the partition on M induced by it. Because both numbers $P(A = a|Z = 1)$ and $P(A = a|Z = 0)$ can be computed from the data generated by the sampling process, we conclude that, for each $a \in \mathcal{A}$, $P(A = a|Z)$ are identified parameters and, consequently, the joint distribution $P(A = a, Z = z)$ for $(a, z) \in \mathcal{A} \times \{0, 1\}$ is identified. Note that these statements do not depend on specific *values* of the probabilities $P(Z = 1)$ and $P(Z = 0)$, but only on the fact that Z is well defined and can be endowed with a probability distribution. Finally, it can be verified that A and Z are not independent by construction. For instance, for $a \in \mathcal{A}$

$$P(Z = 1, A = a) \doteq P[Z^{-1}\{1\} \cap A^{-1}\{a\}] = P\{m \in \mathcal{P} : A(m) = a\} \neq P(Z = 1)P(A = a).$$

3.4. The Random Variables X and Y

Once (A, Z) has been specified, we introduce two random variables X and Y defined from M into $\mathcal{X}$ and $\mathcal{Y}$, respectively. As discussed previously, the equating function depends on two probability distributions, namely, $F_X(x) = P(X \leq x)$ for all $x \in \mathcal{X}$ and $F_Y(y) = P(Y \leq y)$ for all $y \in \mathcal{Y}$: These are actually the parameters of interest. However, these parameters are not identified because it is *not possible to construct the respective conditional distributions* $P(X \leq x|A, Z)$ and

$P(Y \leq y|A, Z)$. As a matter of fact, the marginal distribution $P(X \leq x)$ can only be obtained once $P(X \leq x|A)$ is obtained, which in turn is defined through the following representation: for all $x \in \mathcal{X}$

$$\begin{aligned} P(X \leq x|A) &= P(X \leq x|A, Z = 1)P(Z = 1|A) + P(X \leq x|A, Z = 0)P(Z = 0|A), \\ &= E[P(X \leq x|A, Z{=}\ 1)1_{\{Z=1\}} + P(X \leq x|A, Z{=}\ 0)1_{\{Z=0\}}|A], \\ &= E[P(X \leq x|A, Z)|A]. \end{aligned} \tag{3.4}$$

In this representation, the function $P(X \leq x|A, Z = 0)$ is not identified because it is impossible to determine from the sampling process the event $X^{-1}\{x\} \cap Z^{-1}\{0\} \subset M$, that is, it is not observed which examinees exposed to test form Y would have obtained a score equal to x in test form X. This makes explicit the underlying selection bias problem according to which $P(X \leq x|A, Z = 1) \neq P(X \leq x|A, Z = 0)$.

The nonidentifiability of $P(X \leq x)$ immediately follows: for all $x \in \mathcal{X}$

$$\begin{aligned} F_X(x) \doteq P(X \leq x) &= P(X \leq x|Z = 1)P(Z = 1) + P(X \leq x|Z = 0)P(Z = 0), \\ &= E[P(X \leq x|Z)], \\ &= E[E[P(X \leq x|A, Z)]|A], \\ &\overset{by\ (3.4)}{=} E[P(X \leq x|A)]. \end{aligned} \tag{3.5}$$

By similar arguments, in the following representation, for $y \in \mathcal{Y}$

$$\begin{aligned} P(Y \leq y|A) &= P(Y \leq y|A, Z = 1)P(Z = 1|A) + P(Y \leq y|A, Z = 0)P(Z = 0|A), \\ &= E[P(Y \leq y|A, Z{=}\ 1)1_{\{Z=1\}} + P(Y \leq y|A, Z{=}\ 0)1_{\{Z=0\}}|A], \\ &= E[P(Y \leq y|A, Z)|A]. \end{aligned} \tag{3.6}$$

The function $P(Y \leq y|A, Z = 1)$ is not identified because it is impossible to determine from the sampling process the event $Y^{-1}\{y\} \cap Z^{-1}\{1\} \subset M$, that is, it is not observed which examinees exposed to test form X would have obtained a score equal to y in test form Y. This makes explicit the underlying selection bias problem according to which $P(Y \leq y|A, Z = 1) \neq P(Y \leq y|A, Z = 0)$. Therefore, $P(Y \leq y)$ is not identified because, for all $y \in \mathcal{Y}$

$$\begin{aligned} F_Y(y) \doteq P(Y \leq y) &= P(Y \leq y|Z = 1)P(Z = 1) + P(Y \leq y|Z = 0)P(Z = 0), \\ &= E[P(Y \leq y|Z)], \\ &= E[E[P(Y \leq y|A, Z)]|A], \\ &\overset{by\ (3.6)}{=} E[P(Y \leq y|A)]. \end{aligned} \tag{3.7}$$

Summarizing, the statistical model in the NEAT design can be written as $\{(M, \mathcal{M}) : (F_X, F_Y) \in \mathcal{F}(M, \mathcal{M})\}$, where $\mathcal{F}(M, \mathcal{M})$ denotes the space of probability distributions defined on $(M, \mathcal{M})$. The type of data that can be analyzed under the NEAT design is characterized by the sequential specification developed above.

3.5. Consequences of the Model Specification

The previous specification deserves several comments. First, one might ask what is the role of the anchor test in the previous identification analysis? More precisely, why it is necessary to use the conditional distribution $P(X \leq x|A, Z)$ (see Equation 3.4) to establish the nonidentifiability of $P(X \leq x)$, instead of establishing it using $P(X \leq x|Z)$ directly (see Equation 3.5, last line)? In fact, the nonidentifiability of both $P(X \leq x)$ and $P(Y \leq y)$ follows after marginalizing Z. However, we will show that the anchor test A is useful for introducing identification restrictions leading to either point-identify or partially identify them. Furthermore, the sampling process characterizing the NEAT design identifies

$$P(A = a, Z = z), \qquad P(X \leq x|A, Z = 1), \qquad P(Y \leq y|A, Z = 0),$$

for all $(z, a, x, y) \in \{0, 1\} \times \mathcal{A} \times \mathcal{X} \times \mathcal{Y}$. These distributions cannot be obtained from

$$P(A = a, Z = z), \quad P(X \leq x|Z = 1), \quad P(Y \leq y|Z = 0),$$

for all $(z, a, x, y) \in \{0, 1\} \times \mathcal{A} \times \mathcal{X} \times \mathcal{Y}$: It is necessary to make explicit the dependency between X and A for $\{Z = 1\}$ and between Y and A for $\{Z = 0\}$.

Second, an advantage of this fully probabilistic specification of the NEAT design is to show that the term *synthetic population* is useless. As a matter of fact, the population of interest underlying the NEAT design is given by $M = \mathcal{P} \cup \mathcal{Q}$: Once this is made explicit, all the necessary random variables can explicitly be defined. Moreover, the score distributions $P(X \leq x)$ and $P(Y \leq y)$ are decomposed in a *unique* way through the representations 3.5 and 3.7. In these decompositions, $P(Z = 1)$ is an identified parameter of the statistical model that, once it is fixed (for instance, through Equation 3.2), *cannot be modified arbitrarily* as suggested by Braun and Holland (1982), Brennan and Kolen (1987), and Kolen and Brennan (2014, Section 4.5.2), among many others. Therefore, the "target score distribution" used in the equating literature does not always coincide with $P(X \leq x)$ as decomposed in Equation 3.5, except when $\omega = P(Z = 1)$. In other words, the existence of the target score distribution is *not* well established.

4. Approaches to Tackle the Identification Problem

In this section, we discuss two different approaches to tackle the identification problem underlying the NEAT design. The first one is based on a missing-at-random condition, widely used in the equating literature. In order to grasp how strong such a condition is, we consider a second approach to the identification problem based on the theory of partial identification. The theoretical results are illustrated with a well-known data set appearing in the literature (Kolen & Brennan, 2014).

4.1. Strong Ignorability Condition

In the equating literature, the lack of identifiability of the marginal distributions F_X and F_Y has been considered as a missing data problem (see, among many others, Bolsinova & Maris, 2016; Liou, 1998; Liou & Cheng, 1995; Sinharay & Holland, 2010). As we have seen in Section 2.1, the problem is typically solved assuming that the anchor scores are informative enough such that $F_{X|Z=1,A}(x) = F_{X|Z=0,A}(x)$ for all $x \in \mathcal{X}$ and $F_{Y|Z=1,A} = F_{Y|Z=0,A}$ for all $y \in \mathcal{Y}$, which corresponds to

$$(i) \quad X \perp\!\!\!\perp Z|A; \quad (ii) \quad Y \perp\!\!\!\perp Z|A, \tag{4.1}$$

see Braun and Holland (1982), Kolen and Brennan (2014), von Davier et al. (2004), and González and Wiberg (2017). Here, $U \perp\!\!\!\perp V|W$ denotes the conditional independence of U and V given W (for details, see Florens et al., 1990, Chapter 2). Note that condition 4.1 is equivalent to the assertion of absence of bias. For details, see Supplemental Appendix A.

In the econometric literature, this condition is known as the *switching condition* (Maddala, 1983), whereas in the causal inference literature, it is known as the *strong ignorability condition* (Rosenbaum & Rubin, 1983). Condition 4.1 is not empirically refutable (Manski, 2007), but only justified in a specific application. Such justification could lead to answer affirmatively to the following question: Are we ready to believe that, conditionally on A, the score distribution of the examinees taking test form X would be the same as if they were exposed to test form Y?

What is important to emphasize is that condition 4.1 is *an identification restriction* allowing to identify $F_{X|A}$ and $F_{Y|A}$ and, by extension, F_X and F_Y. As a matter of fact, under condition 4.1, decompositions 3.4 and 3.6 imply that $F_{X|A}(x) = F_{X|Z=1,A}(x)$ for all $x \in \mathcal{X}$ and that $F_{Y|A}(y) = F_{Y|Z=0,A}(y)$ for all $y \in \mathcal{Y}$; and therefore

$$\begin{aligned} (i) \quad F_X(x) &= \sum_{a \in \mathcal{A}} P(X \leq x | Z = 1, A = a) P(A = a) \quad \forall x \in \mathcal{X}, \\ (ii) \quad F_Y(y) &= \sum_{a \in \mathcal{A}} P(Y \leq y | Z = 0, A = a) P(A = a) \quad \forall y \in \mathcal{Y}. \end{aligned} \tag{4.2}$$

The equating function can thus be obtained using Equation 4.2, which is equivalent to Equation 2.5 provided that $\omega = P(Z = 1)$; for a proof, see Supplemental Appendix B. Thus, the strong ignorability condition 4.1 is simply an identification restriction that leads to being able to identify $F_{X|A}$ and $F_{Y|A}$, which in turn implies the identifiability of the parameters of interest F_X and F_Y. Now, the question is: How strong is the condition (4.1)? The remainder of this article is focused in answering this question.

4.2. Partial Identification Analysis

A natural starting point is to see what the data alone reveal about $F_{X|A}$ and $F_{Y|A}$. This can be done by means of a partial identification strategy, widely used in empirical research (see, e.g., Blundell et al., 2007; Gundersen & Kreider, 2009; Molinari, 2010; Pepper, 2000).

In comparison with the traditional concept of identifiability that gives a binary status for the parameters of interest in a statistical model, namely, either identified or not, partial identification is an approach that recognizes that identification is not an all-or-nothing concept and that models that do not identify parameters of interest can, and typically do, contain valuable information about these parameters (Tamer, 2010). Following Manski (2007), a parameter is partially identified if the sampling process and maintained assumptions reveal that the parameter lies in a set, its "identification region," that is smaller than the logical range of the parameter but larger than a single point. In this sense, the smaller (larger) the identification region is, the larger (smaller) the information we have about the parameter.

4.2.1. Partial identification of the parameters of interest. In order to make explicit what the data alone reveal about $F_{X|A}$, it is enough to use that the unidentified conditional probability distribution $F_{X|Z=0,A}$ lies between 0 and 1. Decomposition 3.4 implies, therefore, that the conditional probability $F_{X|A}$ lies in the interval

$$F_{X|Z=1,A}(x)P(Z=1|A) \leq F_{X|A}(x) \leq F_{X|Z=1,A}(x)P(Z=1|A) + P(Z=0|A), \tag{4.3}$$

for all $x \in \mathcal{X}$. Similarly, $F_{Y|Z=1,A}$ lies between 0 and 1 and, therefore, decomposition 3.6 implies that the conditional probability $F_{Y|A}$ lies in the interval

$$F_{Y|Z=0,A}(y)P(Z=0|A) \leq F_{Y|A}(y) \leq F_{Y|Z=1,A}(y)P(Z=0|A) + P(Z=1|A), \tag{4.4}$$

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

Theorem 4.1: In the NEAT design, the parameters of interest F_X and F_Y are partially identified by the following identification intervals:

$$\begin{aligned}&(i)\quad F_{X|Z=1}(x)P(Z=1) \leq F_X(x) \leq F_{X|Z=1}(x)P(Z=1) + P(Z=0),\ x \in \mathcal{X},\\ &(ii)\quad F_{Y|Z=0}(y)P(Z=0) \leq F_Y(y) \leq F_{Y|Z=0}(y)P(Z=0) + P(Z=1),\ y \in \mathcal{Y}.\end{aligned} \tag{4.5}$$

Let us comment on some important consequences of this theorem. First, the partial identification of F_X and F_Y can be directly obtained using decompositions

3.5 and 3.7. However, the partial identification intervals 4.3 and 4.4 will be useful in Section 5 to obtain a more informative identification region.

Second, the partial identification intervals (Equations 4.5.i and 4.5.ii) do not depend on any assumption; they accordingly provide all the plausible values of F_X and F_Y coherent with the data and the probabilities identified by them. In particular, the solution given in Equation 4.2 is a plausible one (see Figure 1). It should be stressed that other solutions of the type 2.5, as the ones suggested in the literature where $\omega \neq P(Z = 1)$, are *not* plausible solutions because *some of the components are not based on the identified parameters underlying the NEAT design.*

Third, the width of interval (Equation 4.5.i) is $P(Z = 0)$, whereas the width of interval (Equation 4.5.ii) is $P(Z = 1)$. Taking into account that $P(Z = 0) + P(Z = 1) = 1$, the longer one interval is, the shorter the other. This trade-off—which is valid even if $P(Z = 1)$ and $P(Z = 0)$ do not correspond to the relative weight sizes of both groups—shows how strong the ignorability condition is: $F_X(x)$ 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., $P(Z = 1)$), which under the ignorability condition collapses to a point, namely, $F_{X|Z=1}(x)$ (respect., $F_{Y|Z=0}(y)$). The degree of information that the ignorability condition hides can be precisely quantified: It corresponds to $P(Z = 1)$ and $P(Z = 0)$, which cannot be arbitrarily manipulated once the groups of examinees have responded the tests. Figure 1 shows this trade-off graphically.

Fourth, the larger the partial identification intervals, the lesser is the information on the unidentified score distributions used to compute the equating function. It could be argued that smaller identification intervals can be obtained simultaneously for both score distributions by conditioning on the anchor scores A. This is in fact one of the ways in which the strong ignorability condition is motivated, namely, that A is so informative that allows us to ignore Z. However, this is not the case. It is enough to compare the width of the conditional partial identification intervals (Equations 4.3 and 4.4) with the intervals (Equations 4.5.i and 4.5.ii), respectively. For each $a \in \mathcal{A}$, we consider the following cases:

$$\begin{aligned}
&\text{(i)} \quad P(Z = 1) > P(Z = 1|A = a) \quad \Leftrightarrow \quad P(Z = 0) < P(Z = 0|A = a),\\
&\text{(ii)} \quad P(Z = 1) < P(Z = 1|A = a) \quad \Leftrightarrow \quad P(Z = 0) > P(Z = 0|A = a)\,. \qquad (4.6)
\end{aligned}$$

It can be seen that after conditioning on A, the identification interval for F_Y is indeed smaller but the one for F_X turns to be larger in Case (i). Likewise in Case (ii), the identification interval for F_X is smaller but the one for F_Y is larger. Thus, it is not possible to reduce simultaneously the width of the identification intervals by conditioning on A. Note that these conditions are empirically testable. Figure 2 shows a graphical representation of these comparisons.

4.2.2. Partial identification of the quantiles. The impact of the lack of identifiability of F_X and F_Y, and the amount of information that can be obtained from a

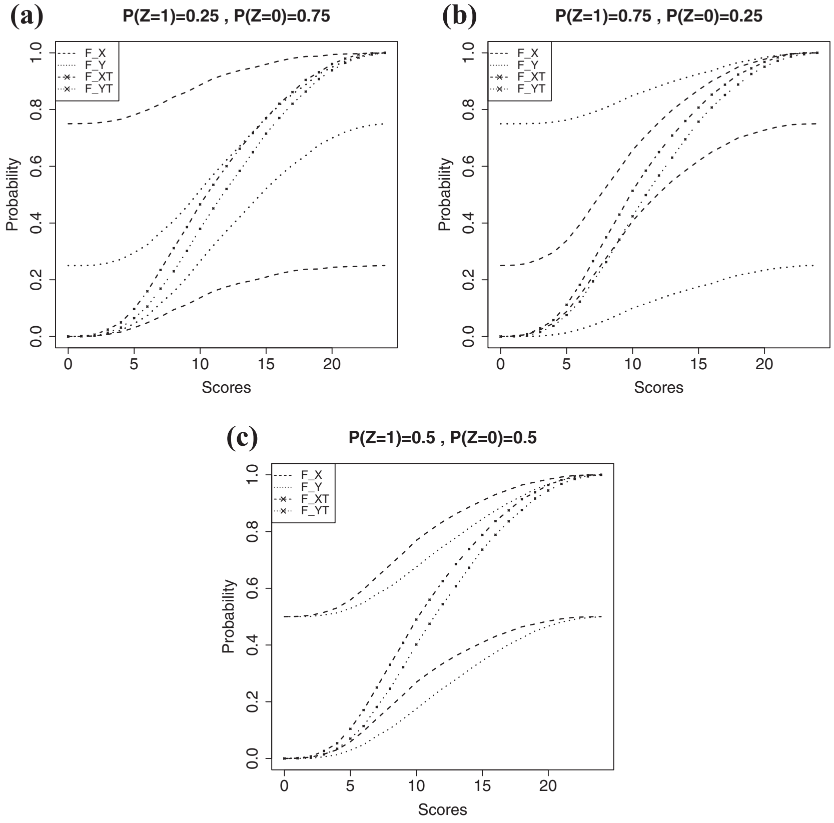

FIGURE 1. *Identifiability bounds for different relative sizes.*

partial identifiability approach on the actual equating, may be studied by analyzing the quantile functions involved in the process. In fact, the equating function actually equates the quantiles of the score distributions F_X and F_Y (see Section 2.1).

Let $\alpha \in (0, 1)$. Taking as a starting point the identifiability bounds in Equation 4.5, we define the following quantiles functions:

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

Note that $r_X(\alpha)$ and $s_X(\alpha)$ are identified, whereas $q_X(\alpha)$ is unidentified. The aim is thus to partially identify $q_X(\alpha)$ using both $r_X(\alpha)$ and $s_X(\alpha)$. To do that, we establish the following relationships using Equation 4.5.i: let $t < r_X(\alpha)$, it follows that $F_{X|Z=1}(t)P(Z = 1) + P(Z = 0) < \alpha$, which in turn implies that

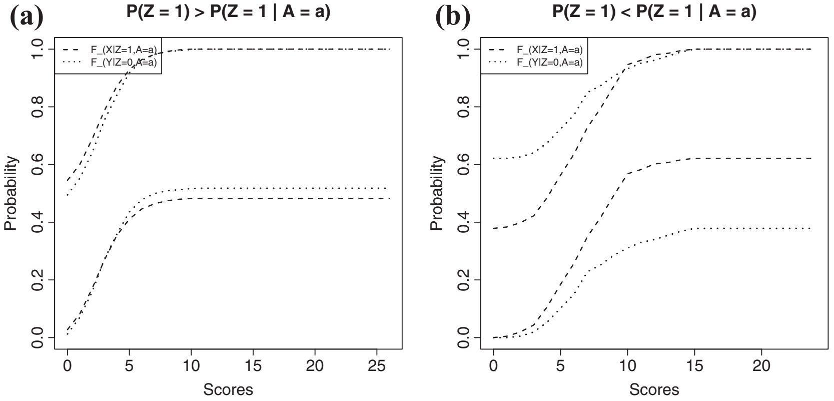

FIGURE 2. *Identifiability bounds for conditional score distributions.*

$F_X(t) < \alpha$ and, therefore, $q_X(\alpha) > t$. 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, let $t \geq s_X(\alpha)$, it follows that $F_{X|Z=1}(t)P(Z=1) > \alpha$, which implies $F_X(t) \geq \alpha$, which in turn implies that $q_X(\alpha) \leq t$. It follows that $q_X(\alpha) \leq s_X(\alpha)$ because $\alpha/P(Z=1) \geq \alpha$.

Similarly, for $\alpha \in [0, 1]$, we define the 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 Equation 4.5.ii, a similar argument leads to conclude that $r_Y(\alpha) \leq q_Y(\alpha) \leq s_Y(\alpha)$. Summarizing, we obtain the following theorem:

Theorem 4.2: In the NEAT design, the quantiles of the partially identified probability distributions F_X and F_Y are partially identified by the following intervals:

$$\text{(i)} \quad r_X(\alpha) \leq q_X(\alpha) \leq s_X(\alpha); \quad (ii) \quad r_Y(\alpha) \leq q_Y(\alpha) \leq s_Y(\alpha). \tag{4.7}$$

In what follows, we give some comments on Theorem 4.2 that highlight relevant aspects that can be learned from what the data are able to identify and to show how severe is the identification problem underlying the NEAT design. Although the analyses are valid for any identified specification of $P(Z = z)$, $z \in \{0, 1\}$, throughout the exposition, we assume the specification given in Equation 3.2.

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 groups $\mathcal{P}$ and $\mathcal{Q}$, we 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 $t_m^{X|Z=1} \doteq \min\{t : t \in S_{(X|Z=1)}\}$ and $t_M^{X|Z=1} \doteq \max\{t : t \in S_{(X|Z=1)}\}$, similarly 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

$$r_X(\alpha^*) = \inf\left\{t : F_{X|Z=1}(t) > \frac{\alpha - P(Z=0)}{P(Z=1)}\right\} = q_{X|Z=1}\left(\frac{\alpha - P(Z=0)}{P(Z=1)}\right),$$

$$s_X(\alpha^*) = \inf\left\{t : F_{X|Z=1}(t) > \frac{\alpha}{P(Z=1)}\right\} = q_{X|Z=1}\left(\frac{\alpha}{P(Z=1)}\right),$$

where $\alpha^* \in (0,1)$ is the percentage of the observations lying below t and it depends on the relative sizes of groups $\mathcal{P}$ and $\mathcal{Q}$. Similarly, the lower and upper bounds of $q_Y(\alpha)$ can be expressed as quantiles of the conditional distribution $F_{Y|Z=0}$, namely

$$r_Y(\alpha^*) = \inf\left\{t : F_{Y|Z=0}(t) > \frac{\alpha - P(Z=1)}{P(Z=0)}\right\} = q_{Y|Z=0}\left(\frac{\alpha - P(Z=1)}{P(Z=0)}\right),$$

$$s_Y(\alpha^*) = \inf\left\{t : F_{Y|Z=0}(t) > \frac{\alpha}{P(Z=0)}\right\} = q_{Y|Z=0}\left(\frac{\alpha}{P(Z=0)}\right).$$

Because these quantiles depend on the group's relative sizes, three cases can be distinguished, namely, $P(Z=1) < P(Z=0)$, $P(Z=1) > P(Z=0)$, and $P(Z=1) = P(Z=0)$. Tables 2 through 4 show the corresponding upper and lower bounds for each of these cases, respectively. The identification regions shown in these tables make explicit how severe is the nonuniqueness of the equated values due to the identification problem inherent to the NEAT design. Figure 3 complements the information given in these tables showing plots of the quantile functions in each case. We comment on the uniqueness issue by making reference to Table 2, focusing attention on the role of the relative sizes of groups $\mathcal{P}$ and $\mathcal{Q}$.

Interval 1: 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 $t \in \mathcal{W} = \mathcal{X} \cap \mathcal{Y}$. This implies that $S_{(Y|Z=0)} \supseteq S_{(X|Z=1)}$. Given that the quantile function respects the stochastic dominance (see, e.g., Stoye, 2010), it follows that

$$q_{Y|Z=0}\left(\frac{\alpha}{P(Z=0)}\right) \leq q_{X|Z=1}\left(\frac{\alpha}{P(Z=0)}\right). \tag{4.8}$$

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

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

TABLE 2.
Lower and Upper Bounds for $\mathrm{q_X}(\alpha)$ *and* $\mathrm{q_Y}(\alpha)$ *When* $\mathrm{P(Z = 1) < P(Z = 0)}$

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

TABLE 3.
Lower and Upper Bounds for $\mathrm{q_X}(\alpha)$ *and* $\mathrm{q_Y}(\alpha)$ *When* $\mathrm{P(Z=1) > P(Z=0)}$

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

TABLE 4.
Lower and Upper Bounds for $\mathrm{q_X}(\alpha)$ *and* $\mathrm{q_Y}(\alpha)$ *When* $\mathrm{P(Z = 1) = P(Z = 0)}$

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

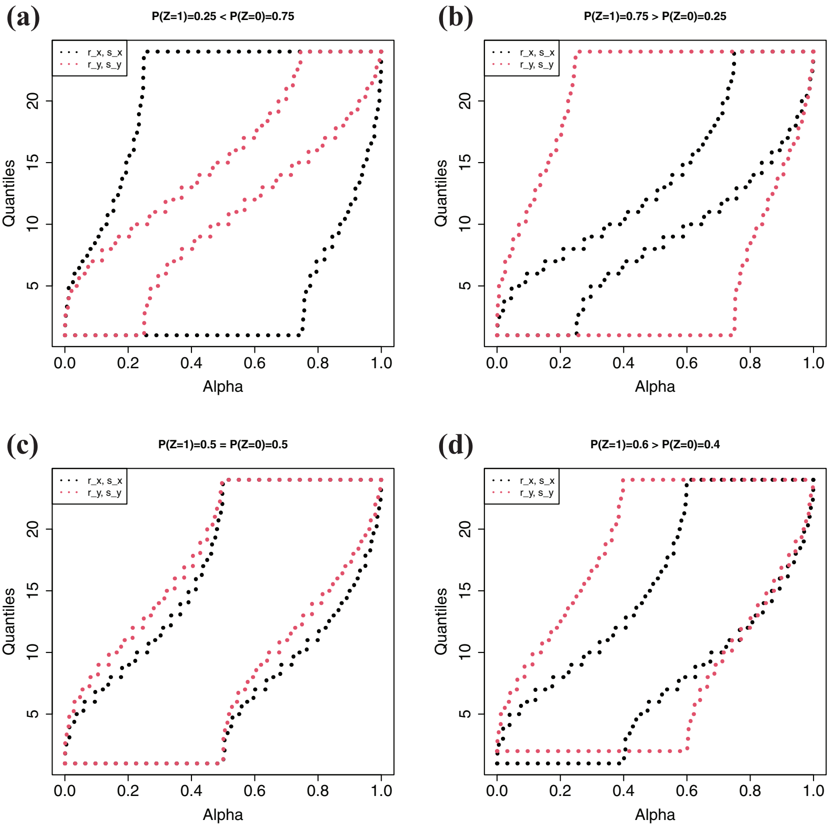

FIGURE 3. *Identifiability bounds of quantiles for different relative sizes.*

$$\leq q_{X|Z=1}\left(\frac{\alpha}{P(Z=1)}\right) = s_X(\alpha^*),$$

(see Figure 3a). If for simplicity we also assume that $t_m^{X|Z=1} = t_m^{Y|Z=0}$, equating a X-score to a Y-score would lead to shrunk values of the X-scores. Notice that in this case, no explicit order between $r_X(\alpha^*)$ and $r_Y(\alpha^*)$ can be established (see Figure 3d).

Remark 4.1: The stochastic dominance assumption is only meaningful in the context of equating where the test forms to be equated might be assembled as parallel as possible, which leads to score distributions with common supports. Nevertheless, the same argument follows if instead of assuming the stochastic dominance we assume the quantile inequality (Equation 4.8). By doing so, the conclusions are also valid in a less restricted linking context as can be seen in van der Linden (2019, p. 427).

Interval 2: If $\alpha \in (P(Z=1), P(Z=0))$, $q_X(\alpha^*)$ is partially identified on $(r_X(\alpha^*), s_X(\alpha^*)) = S_{(X|Z=1)}$, and thus, no bounded information is obtained to learn about it. On the other hand, $q_Y(\alpha^*)$ provides information characterized by the respective identification interval. As a consequence, Y-scores on the interval $(r_Y(\alpha^*), s_Y(\alpha^*))$ would be considered equivalent to all possible X-scores (see Figure 3a).

Interval 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 $t \in \mathcal{W} = \mathcal{X} \cap \mathcal{Y}$, which implies that $S_{(X|Z=1)} \supseteq S_{(Y|Z=0)}$. It follows that

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

But $P(Z=1) < P(Z=0)$, so $[\alpha - P(Z=1)]/P(Z=0) > [\alpha - P(Z=0)]/P(Z=1)$. It follows that

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

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

If for simplicity we assume that $t_M^{X|Z=1} = t_M^{Y|Z=0}$, equating an X-score to a Y-score would lead to X-scores defined on an expanded sample space. Note that in this case, no explicit order between $s_X(\alpha^*)$ and $s_Y(\alpha^*)$ can be established.

Similar comments apply for Table 3 (see, in particular, Figure 3b). Regarding Table 4, it can be noted that if $P(Z=1) = P(Z=0)$, the asymmetry observed (i.e., shrunk or expanded sample spaces) disappears, but the severity of the identification problem still persists: An X-score is not uniquely transformed to a Y-score (see Figure 3c). In other words, the results show that partial identification implies lack of equitability for examinees with scores inside the identification region.

4.3. Empirical Illustration

To illustrate Theorems 4.1 and 4.2 and complement the comments given in the previous sections, we use a data set appearing in Kolen and Brennan (2014). The data consist of two 36-item test forms for which 12 of the 36 items are common between both test forms (Items 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, and 36). Test Form X was administered to $n_{\mathcal{P}} = 1,655$ examinees, whereas Test Form Y was administered to $n_{\mathcal{Q}} = 1,638$ examinees, so that, for this data $P(Z = 1) = \frac{n_{\mathcal{P}}}{n_{\mathcal{P}}+n_{\mathcal{Q}}} = \frac{1,655}{1,655+1638} = 0.503$. Subsamples were used to illustrate cases when the relative sizes of both groups differed.

We derived the identification regions for both F_X and F_Y and for the associated quantiles $q_X(\alpha)$ and $q_Y(\alpha)$. Figure 1 shows the identifiability bounds for different values of the relative sizes and the corresponding target score distributions proposed in the literature that are obtained under the traditional practice of identifying using the ignorability condition for the case when $\omega = P(Z = 1)$. It can be seen that the target score distributions computed as in Equation 2.5 are indeed a plausible solution that lies between the identification bounds. Figure 1 also illustrates how the degree of information associated with F_X and F_Y due to the identification problem is quantified by the width of the identification intervals, which corresponds to the relative sizes of groups $\mathcal{P}$ and $\mathcal{Q}$ respectively. The degree of information on F_X is quantified by $P(Z = 0)$, while for F_Y is quantified by $P(Z = 1)$.

Figure 2 illustrates the fact that conditioning on A does not necessarily lead to improve the identification intervals simultaneously for F_X and F_Y (as it was commented after Theorem 4.1).

The identifiability bounds for the quantile functions are shown in Figure 3. This figure corroborates the findings related to the asymmetry of sample spaces derived from Theorem 4.2. For instance, if we consider Figure 3a, we can conclude that transforming Y-scores to X-scores increases the lack of information inherent to the Y-scale; if we do the converse transformation, the degree of information inherent to the X-scale decreases (see Figure 3b). This result could be considered as a criterion to choose which direction the equating transformation should be performed. Nevertheless, it also shows that the degree of information can*not* be increased arbitrarily because it depends on the relative sizes of populations $\mathcal{P}$ and $\mathcal{Q}$. 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 3c), but the lack of information by design persists: If a score on $\mathcal{X}$ is transformed to the $\mathcal{Y}$ scale using φ, no *unique* equivalent Y-score is obtained.

5. Improved Partial Identification Intervals

In this section, we show how the partial identification interval can be improved in the sense that their widths decrease. This can be done by introducing alternative assumptions. We motivate them in the context of a real data set example.

TABLE 5.
Anchor Scores for Different Quantiles of the Distributions in $\mathcal{P}$ and $\mathcal{Q}$

Group	Size	$q(0.25)$	$q(0.5)$	$q(0.75)$	Min.	Max.
$\mathcal{P}$	15.048	10	17	28	0	53
$\mathcal{Q}$	13.771	11	18	30	0	53

5.1. Motivation

The Chilean *Prueba de Selección Universitaria* test (PSU, by their initials in Spanish) is a university entrance test composed of four different sections: Language, Mathematics, History, and Sciences (for details, see Horn et al., 2014). For the Sciences section, there are three different test forms, each of them composed of 80 items and sharing 54 of the 80 items in common. More specifically, an examinee can choose among a Sciences–Physics, Sciences–Biology, and Sciences–Chemistry test form, where each form contains 26 subject-specific items and $18 \times 3 = 54$ common items, 18 of each subject. A particular feature of the sciences test is that a unique *score on sciences* should be reported, no matter what of the three forms was chosen or, in other words, no matter to which of the group the examinee belongs to. Thus, to report a score on sciences, a linking procedure under a NEAT design is needed.

For illustrative purposes, let us focus the attention on the following two groups: Group $\mathcal{P}$ corresponds to the examinees who choose the Sciences–Chemistry form and Group $\mathcal{Q}$ corresponds to the examinees who choose the Sciences–Physics form. Following the notation of the main text, Form X corresponds to Chemistry and Form Y to Physics. Table 5 shows the descriptive summaries of the anchor scores. Both show that the distribution of anchor scores in groups $\mathcal{P}$ and $\mathcal{Q}$ is similar. Furthermore, for Group $\mathcal{P}$, the correlation between the X-scores and the anchor scores is 0.9, whereas that for Group $\mathcal{Q}$, the correlation between the Y-scores and the anchor scores is 0.86.

5.2. Partial Identification of the Parameters of Interest

5.2.1. Modeling self-selection. The fact that each examinee can choose which specific form to take 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 better more likely than if they had chosen form Y and those who take the Form Y do so because they believe they will score better more likely than if they had chosen Form X. This type of strategy is indeed used in the context of the Sciences PSU test, as some students are better trained in a subject than in other. As can be seen, these assumptions are intended to formalize the idea that an examinee chooses one form of testing or another because they want to maximize their score.

Given that the test forms include a set of common items, the assumption that examinees choose form X aiming to score better more likely had they chosen form Y can be represented probabilistically by at least the following three set of conditions:

$$P(X > t|Z = 1, A = a) > P(Y > t|Z = 1, A = a), \quad \forall (t, a) \in \mathcal{W} \times \mathcal{A}, \tag{5.1}$$

$$P(Y > t|Z = 0, A = a) > P(X > t|Z = 0, A = a), \quad \forall (t, a) \in \mathcal{W} \times \mathcal{A}, \tag{5.2}$$

$$P(X > t|Z = 1, A \in \mathcal{A}_1) > P(Y > t|Z = 1, A \in \mathcal{A}_1), \quad \forall t \in \mathcal{W},\ \mathcal{A}_1 \subsetneqq \mathcal{A}, \tag{5.3}$$

$$P(Y > t|Z = 0, A \in \mathcal{A}_1) > P(X > t|Z = 0, A \in \mathcal{A}_1), \quad \forall t \in \mathcal{W},\ \mathcal{A}_1 \subsetneqq \mathcal{A}, \tag{5.4}$$

$$P(X > t|Z = 1) > P(Y > t|Z = 1), \quad \forall t \in \mathcal{W}, \tag{5.5}$$

$$P(Y > t|Z = 0) > P(X > t|Z = 0), \quad \forall t \in \mathcal{W}. \tag{5.6}$$

Conditions 5.1 and 5.2 depend on all the A-scores, whereas Equations 5.3 and 5.4 depend on some specific A-scores values, for instance, $\mathcal{A}_1 = \{a \in \mathcal{A} : a \geq a_1\}$. Here, we further assume that examinees who obtain an anchor score $a \in \mathcal{A}_1$ are supposed to maximize their specific scores by choosing either Form X or Form Y. Consequently, for the remaining examinees obtaining an A-score in $\mathcal{A}_1^c$, the complement of the set $\mathcal{A}$, no self-selection assumptions are considered. Finally, conditions 5.5 and 5.6 do not depend on the anchor. It should be noted that all these conditions, like the strong ignorability (Equation 4.1), are not empirically testable because they all depend on unidentified parameters. However, they all have an impact on the identification intervals in the sense that the stronger are the conditions, the shorter are the identification intervals obtained, which is in line with the so-called *mean monotonicity assumptions* used in econometrics (see Manski, 2007).

5.2.2. Impact of the self-selection assumption on the identifiability of F_X and F_Y. The following theorem summarizes the impact of the self-selection assumptions on the identifiability of F_X and F_Y:

Theorem 5.1: In the NEAT design, under the assumptions 5.1 and 5.2, the parameters of interest F_X and F_Y are partially identified by the following intervals: for all $t \in \mathcal{W}$

$$F_{X|Z=1}(t)P(Z = 1) + F_{Y|Z=0}(t)P(Z = 0) \leq F_X(t) \leq F_{X|Z=1}(t)P(Z = 1) + P(Z = 0), \tag{5.7}$$

$$F_{X|Z=1}(t)P(Z = 1) + F_{Y|Z=0}(t)P(Z = 0) \leq F_Y(t) \leq F_{Y|Z=0}(t)P(Z = 0) + P(Z = 1). \tag{5.8}$$

Under assumptions 5.3 and 5.4, the parameters of interest F_X and F_Y are partially identified by the following intervals: for all $t \in \mathcal{W}$

TABLE 6.
Widths of Identification Intervals

Identification Interval	Width of the Interval
(4.5.i)	$P(Z = 0)$
(4.5.ii)	$P(Z = 1)$
(5.7)	$P(Y > t, Z = 0)$
(5.8)	$P(X > t, Z = 1)$
(5.9)	$P(Z = 0, A \in \mathcal{A}_1^c) + P(Y > t, Z = 0, A \in \mathcal{A}_1)$
(5.10)	$P(Z = 1, A \in \mathcal{A}_1^c) + P(X > t, Z = 1, A \in \mathcal{A}_1)$

$$\begin{aligned} &F_{X|Z=1}(t)P(Z = 1) + F_{Y|Z=0,A\in\mathcal{A}_1}(t)P(Z = 0, A \in \mathcal{A}_1) \\ &\leq F_X(t) \leq F_{X|Z=1}(t)P(Z = 1) + P(Z = 0), \end{aligned} \tag{5.9}$$

$$\begin{aligned} &F_{X|Z=1,A\in\mathcal{A}_1}(t)P(Z = 1, A \in \mathcal{A}_1) + F_{Y|Z=0}(t)P(Z = 0) \\ &\leq F_Y(t) \leq F_{Y|Z=0}(t)P(Z = 0) + P(Z = 1). \end{aligned} \tag{5.10}$$

Finally, under assumptions 5.5 and 5.6, the parameters of interest F_X and F_Y are partially identified by the intervals 5.7 and 5.8, respectively.

For a proof, see Supplemental Appendix C.

The upper bounds for the three sets of intervals are the same as the ones derived in Theorem 4.1. Furthermore, intervals 5.7 and 5.8 have a common lower bound despite that these identification intervals follow from two different assumptions, namely, Equations 5.1 and 5.2, or 5.5 and 5.6. In order to see how the identification intervals in Theorem 5.1 improve those in Theorem 4.1, Table 6 summarizes the corresponding width of the identification intervals. The identification interval (Equation 5.7) improves the interval (Equation 5.9), which in turn improves the interval (Equation 4.5.i) because

$$\begin{aligned} P(Y > t, Z = 0) &= P(Y > t, Z = 0, A \in \mathcal{A}_1) + P(Y > t, Z = 0, A \in \mathcal{A}_1^c) \\ &\leq P(Y > t, Z = 0, A \in \mathcal{A}_1) + P(Z = 0, A \in \mathcal{A}_1^c) \\ &\leq P(Z = 0, A \in \mathcal{A}_1) + P(Z = 0, A \in \mathcal{A}_1^c) \\ &= P(Z = 0). \end{aligned}$$

Similarly, the identification interval (Equation 5.8) improves the interval (Equation 5.10), which in turn improves the interval (Equation 4.5.ii). Note finally that the width of the intervals (Equations 5.7, 5.8, 5.9, and 5.10) depends on t and, consequently, their width is not constant.

These results show that the stronger an assumption is, the smaller the intervals, thus alleviating the severity of the identification problem; or, using Manski's (2007) jargon, this is an example of the *Law of Decreasing Credibility*: "The credibility of inference decreases with the strength of the assumptions maintained" (p. 1).

In the Supplementary Material, we briefly discuss the impact of the self-selection assumptions on the partial identification of the quantiles of both F_X and F_Y.

5.2.3. Illustration. We come back to our motivating example to illustrate the results presented in this section. The identification intervals for F_X and F_Y are shown graphically in Figure 4. The results when no particular assumptions are made are shown in Figure 4a and are used here as a reference for comparison. The improvement of the identification intervals is graphically illustrated in Figure 4b, for $\mathcal{A}_1 = \{a \in \mathcal{A} : a \geq 30\}$, and Figure 4c, when no conditioning on A is made. It can be seen that compared to Figure 4a, both identification intervals are narrower. The differences in the width of the intervals and the dependency on t is also reflected in these figures. Note that the width of the partial identification intervals decreases as the scores increases which seems to be relevant in the context of a selection university process.

Summarizing, under assumptions (5.3) and (5.4), and (5.5) and (5.6), the partial identification intervals of the parameters of interest are better than the corresponding intervals derived in Sections 4.2.1 and 4.2.2. The example shows that if self-selection assumptions are introduced, we learn something additional from the evidence (examinees' scores).

6. Discussion

The objective of statistical modeling is to specify the probability distribution that generates the *observables*. This modeling process corresponds to a combination of evidence (the observables, the data) with the 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, p. 2064). The content of this article 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 two type of assumptions used to tackle the identification problem inherent to the NEAT design: strong ignorability (discussed in Section 4.1) that leads to identify the score distributions and partial identifiability (discussed in Sections 4.2.1 and 4.2.2), where no assumption regarding the unobserved quantities is needed because the aim is to learn what the data alone (i.e., without assumptions) can inform us.

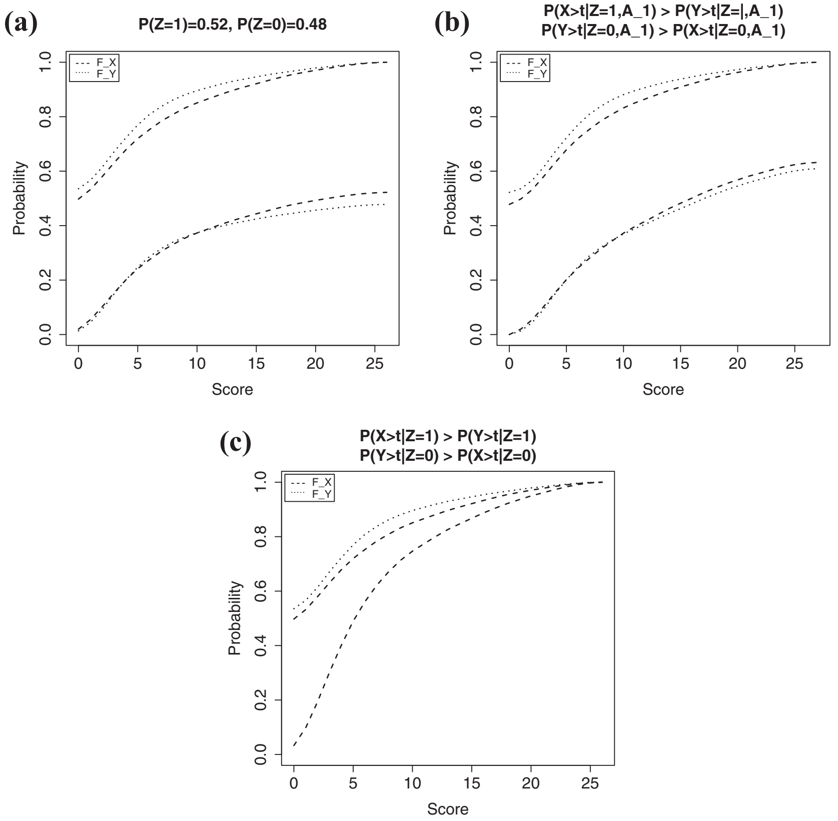

FIGURE 4. *Identifiability bounds for the score distributions of the Chilean* Prueba de Selección Universitaria *Sciences test. (a) Under no assumptions. (b) Under a self-selection process conditioning on* $\mathcal{A}_1 = \{a \in \mathcal{A} : a \geq 30\}$. *(c) Under a self-selection process without conditioning on* A *(for the lower bounds, the point and segmented curves are superimposed in c).*

The modeling process consisted in specifying the statistical model (developed in Section 3), that is, to specify (i) a set of probability distributions generating the observations, (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 thus, we want to know if such characteristics can be expressed as functional of the sampling probabilities: 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 modeling strategy allows us to make explicit their meaning as well as their lack of identifiability. We suppose that the researcher knows that those distributions are necessary to define the equating function (see Section 2.1), and that it is possible to *explicitly show* why those parameters cannot be derived from the statistical model, which is done through the decompositions (3.5) and (3.7): 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 (3.5) and (3.7) are a key step that surprisingly have not been used in the equating literature, even though it has been recognized that the lack of identifiability of $F_{X|A,Z=0}$ and $F_{Y|A,Z=1}$ can be thought as a missing data problem (see Holland et al., 2008; Sinharay & Holland, 2010). The missing data problem and, in more general terms, the selection problem become palatable after using the Law of Total Probability: It allows us to correctly relate F_X and F_Y with the identified conditional probability distributions. Consequences of this key step are twofold: On the one hand, the notion of a target *synthetic population* is meaningless; on the other hand, 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 gain when we combine the evidence (examinees' scores) with assumptions on the unobserved quantities? If we are ready to believe the strong ignorability condition, then $F_{X|A}$ and $F_{Y|A}$ become 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, $P(Z=1)$ and $P(Z=0)$ do not play any role.

If no assumption regarding the unobserved quantities is made, we have shown the severity of the lack of identifiability of F_X and F_Y. The partial identification intervals show what we learn from the evidence in the absence of assumptions regarding the unidentified parameters $F_{X|Z=0}$ and $F_{Y|Z=1}$. In particular, $P(Z=1)$ and $P(Z=0)$ quantify the *lack of information by design* that underlies F_X and F_Y: It involves a trade-off in the sense that the bigger $P(Z=1)$ (respect., $P(Z=0)$) is, the less we know about F_Y (respect., F_X).

The lack of information on the unidentified parameters can somehow be reduced by obtaining improved (shorter) identification intervals as those discussed in Section 5. The researcher's assumptions reflect what they think about the performance of the examinees taking the test forms. When these assumptions are combined with the observations, it is possible to assess their impact in the sense that it will become explicit to what extent the inference depends largely on such assumptions and not just on the observations. A concrete example of this modeling strategy is developed in Section 5.2.

In this article, we have used the equipercentile equating transformation (see Equation 2.1) that is more appealing for the FEE method (Angoff, 1984), also called postratification (von Davier et al., 2004). Other equating transformations based on IRT are also not exempted of the identifiability problem. For instance, in OSE, different constraints must be imposed to solve the identifiability problem (Sinharay & Holland, 2010; van der Linden & Barrett, 2016). Also, in local equating (van der Linden, 2019), which is based on the assumption that X, Y, and A all measure the same ability, θ, the parameters of interest are the conditional distributions $F_{X|\theta}(x)$ and $F_{Y|\theta}(y)$. These parameters are identified when the items in the two forms and the anchor are jointly calibrated with appropriate identification restriction under the same response model or when all parameters are linked afterward through the anchor in the case of separate calibration of the two forms.

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 modeling process like the one developed in this article is relevant because it makes explicit to what extent scientific knowledge is highly dependent on those assumptions. Psychometrics needs to travel these avenues in order to be honest (Pielke, 2007):

> Scientific honesty demands that the specification of a model be based on prior knowledge of the phenomenon studied and possibly on criteria of simplicity, but not on the desire for identifiability of characteristics that the researcher happens to be interested in. (Koopmans & Reiersol, 1950, p. 169f)

Acknowledgments

The first author acknowledges the partial financial support of the FONDECYT Project No. 1181261. The second author acknowledges the partial financial support of the FONDECYT Project No. 1201129. The revisions of this work were supported by the Agencia Nacional de Investigacion y Desarrollo (ANID) Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI); (NCS2021072). 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. The authors thank the Editor, Associate Editor, and two anonymous Referees for their valuable comments, suggestions, and stimulating questions, which helped us to improve this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Fondo Nacional de Desarrollo Científico y Tecnológico (1181261 and 1201129).

References

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service, Princeton.

Angoff, W. H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement*, *11*(3), 291–300.

Bhide, A., Shah, P. S., & Acharya, G. (2018). A simplified guide to randomized controlled trials. *Acta Obstetricia et Gynecologica Scandinavica*, *97*(4), 380–387.

Blundell, R., Gosling, A., Ichimura, H., & Meghir, C. (2007). Changes in the distribution of male and female wages accounting for employment 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. B. Rubin (Eds.), *Test equating* (pp. 9–49). Academic Press.

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

Carnap, R. (1962). *Logical foundations of probability*. The University of Chicago Press.

Embrechts, P., & Hofert, M. (2013). A note on generalized inverses. *Mathematical Methods of Operations Research*, *77*(3), 423–432.

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. (1973). *Statistical Methods for Research Workers*. Hafner Publishing, New York, USA.

Florens, J. P., & Mouchart, M. (1982). A note on noncausality. *Econometrica*, *50*(3), 583–591.

Florens, J. P., Mouchart, M., & Rolin, J.-M. (1990). *Elements of Bayesian statistics*. Marcel Dekker.

González, J., & Wiberg, M. (2017). *Applying test equating methods using R*. Springer.

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

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. *Educational Measurement*, *4*, 187–220.

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.

Horn, C., Santelices, M.V., & Avendaño, X.C. (2014). Modeling the impacts of national and institutional financial aid opportunities on persistence at an elite Chilean university. *Higher Education*, *68*(3), 471–488.

Itô, K. (1984). *An introduction to probability theory*. Cambridge University Press.

Karr, A. F. (1993). *Probability*. Springer.

Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). Springer.

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

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. Chelsea Publishing Company.

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

Koopmans, T. C., & Reiersol, O. (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics*, *21*(2), 165–181.

Liou, M. (1998). Establishing score comparability in heterogeneous populations. *Statistica Sinica*, 669–690.

Liou, M., & Cheng, P. E. (1995). Equipercentile equating via data–imputation technique. *Psychometrika*, *60*, 119–136.

Little, R. J., & Rubin, D. B. (1994). Test equating from biased samples, with application to the armed services vocational aptitude battery. *Journal of Educational and Behavioral Statistics*, 309–335.

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

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

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

Manski, C. F. (2007). *Identification for prediction and decision*. Harvard University Press.

Manski, C. F. (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. F., & Nagin, D. S. (1998). Bounding disagreements about treatment effects: A case study of sentencing and recidivism. *Sociological Methodology*, *28*(1), 99–137.

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.

Odgaard-Jensen, J., Vist, G. E., Timmer, A., Kunz, R., Akl, E. A., Schünemann, H. … Oxman, A. D. (2011). Randomisation to protect against selection bias in healthcare trials. *Cochrane Database of Systematic Reviews*, *2011*(4), MR000012. https://doi.org/10.1002/14651858.MR000012.pub3

Pepper, J. (2000). The intergenerational transmission of welfare receipt: A nonparametric bounds analysis. *The Review of Economics and Statistics*, *82*, 472–488.

Pielke, R. A., Jr, (2007). *The honest broker: Making sense of science in policy and politics*. Cambridge University Press.

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., González, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *80*, 450–467.

Sinharay, S., & Holland, P. W. (2010). The missing data assumptions of the NEAT design and their implications for test equating. *Psychometrika*, *75*, 309–327.

Stephenson, J., & Imrie, J. (1998). Why do we need randomised controlled trials to assess behavioural interventions? *BMJ*, *316*(7131), 611–613.

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

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, *2*(1), 167–195.

van der Linden, W. J. (2019). Lord's equity theorem revisited. *Journal of Educational and Behavioral Statistics*, *44*(4), 415–430.

van der Linden, W. J., & Barrett, M. D. (2016). Linking item response model parameters. *Psychometrika*, *81*(3), 650–673.

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

Wunsch, G., Mouchart, M., & Russo, F. (2014). Functions and mechanisms in structural-modelling explanations. *Journal for General Philosophy of Science*, *45*(1), 187–208.

Authors

ERNESTO SAN MARTÍN is a professor at the Faculty of Mathematics, Pontificia Universidad Católica de Chile, Chile, an invited professor at the Economics School of Louvain, Catholic University of Louvain, Belgium, and the director of the Interdisciplinary Laboratory of Social Statistics, Chile. His research interests are identification problems in psychometrics and econometrics models, causal inference in social sciences, and school effectiveness and value-added models.

JORGE GONZALEZ is an associate professor at the Faculty of Mathematics, Pontificia Universidad Católica de Chile, Chile, and the deputy director of the Interdisciplinary Laboratory of Social Statistics, Chile. His research interests are psychometrics and educational measurement, with particular emphasis in test equating.

Manuscript received April 14, 2021
Revision received November 25, 2021
Accepted January 11, 2022

A Critical View on the NEAT Equating Design: Statistical Modelling and Identifiability Problems

Supplementary Material

A A definition of random assignment in terms of bias

A way of conceiving the "realization of score data in test equating" is motivated by the assignment of tests forms to "randomly selected" students. A clear example in test equating is the assignment of test forms to randomly selected test takers under the equivalent groups design, a scheme that can be associated with what is called a Randomized Controlled Trial. But, what is actually the meaning of random selection/random assignment/random allocation, when terms such as "random" and "realization" are mere denominations? In what follows we define "random assignment" in terms of "absence of bias", which in turn is formally described in terms of conditional probabilities.

Stephenson and Imrie (1998) point out that a randomized control trial (RCT) is the best way of measuring the efficacy of intervention "because of its ability to minimise bias and avoid false conclusions. Random assignment of individuals to different treatment groups is the best way of achieving a balance between groups for the known and unknown factors that influence outcome" (p.611). Odgaard-Jensen et al. (2011) are even more explicit in affirming that bias is reduced thanks to random assignment:

> Randomised trials use the play of chance to assign participants to comparison groups. The unpredictability of the process, if not subverted, should prevent systematic differences between comparison

> groups (selection bias). Differences due to chance will still occur and these are minimised by randomising a sufficiently large number of people.

Bhide, Shah, and Acharya (2018) contrasts a RCT with an observational study: the evidence based on RCTs is the highest one because “evidence based on observational data is prone to bias”. *Bias* is understood as “the systematic tendency of any factors associated with the design, conduct, analysis, evaluation and interpretation of the results of a study to make the estimate of the effect of a treatment or intervention deviate from its true value”. The presence of bias in an observational study is described in the following terms: “If two or more groups are being compared in an observational study, there are often systematic differences between the groups, so much so that the outcome of the groups may be different because of these differences rather than actual exposure or intervention” (p.381). The way to overcome this serious drawback is through a random assignment:

> The only way to eliminate these differences is to allocate each individual to one or the other intervention at random. Therefore, the probability of any individual receiving one intervention or the other is decided solely by chance (p.381).

All these statements provide us with an intuition, but they fail to be formal in probabilistic terms because *random* is not a probabilistic concept, just a designation, as we mentioned in Section 3.2. However, they provide us with a term that can indeed be defined in probabilistic terms: *bias*. We propose a *structural definition* of bias, that is, one that makes explicit the concept of bias and, therefore, is applicable to many concrete situations.

This definition arise from the Law of Total Probability because it allows us to make explicit the counterfactual aspect underlying bias selection. More specifically, let M be the sample space representing the population of interest. Let V be a random variable defined on M representing the outcome of interest (for instance, the consequence of a treatment), and C a random vector defined on M describing characteristics of the statistical units. The target is to learn about $P(V \leq v \mid C)$ from $P(V \leq v \mid C, Z = 1)$, where Z denotes the random assignment variable, namely

$$Z = \begin{cases} 1, & \text{if a statistical unit in } M \text{ is selected;} \\ 0, & \text{if a statistical unit in } M \text{ is not selected.} \end{cases}$$

Since Z induces a partition on M, it is natural to decompose $P(V \leq v \mid C)$ through the Law of Total Probability, namely

$$P(V \leq v \mid C) = P(V \leq v \mid C, Z = 1)P(Z = 1 \mid C) + P(V \leq v \mid C, Z = 0)P(Z = 0 \mid C);$$

here $P(V \leq v \mid C, Z = 0)$ corresponds to the conditional probability of the outcome that the statistical units would have experienced if they had been assigned to the treatment. Consequently, to learn about $P(V \leq v \mid C)$, it is possible to ignore this last conditional probability only if

$$P(V \leq v \mid C, Z = 1) = P(V \leq v \mid C, Z = 0), \tag{A.1}$$

or equivalently, if $V \perp\!\!\!\perp Z \mid C$, which means *absence of bias*.

Additionally, if the statistical units in $Z^{-1}\{1\}$ were chosen without taking into account their characteristics captured by C, then

$$Z \perp\!\!\!\perp C. \tag{A.2}$$

A RCT can thus accordingly be defined through the structural conditions (A.1) and (A.2). Furthermore, *bias* should be defined as the logical negation of (A.1). Let us remark that we call these conditions *structural* because we emphasize the “form” or “anatomy” of the concept, which need to be revealed for specific problems. Moreover, up to the best of our knowledge, there not exist a formal proof deriving conditions (A.1) and (A.2) from “random assignment mechanism”: if the proof existed, these conditions would be a consequence; if not, they would be a formal definition.

B Proof of the equivalence between (2.5) and (4.2)

If $\omega = P(Z=1)$, then

$$\begin{aligned}
F_X(x) &= F_{X|Z=1}(x)P(Z=1) + F_{X|Z=0}(x)P(Z=0)\\
&= F_{X|Z=1}(x)P(Z=1) + \sum_{a\in\mathcal{A}} P(X \le x \mid A=a, Z=0)P(A=a \mid Z=0)P(Z=0), \quad \text{namely (2.5)}\\
&= F_{X|Z=1}(x)P(Z=1) + \sum_{a\in\mathcal{A}} P(X \le x \mid A=a, Z=1)P(A=a \mid Z=0)P(Z=0) \quad \text{by (4.1)}\\
&= \sum_{a\in\mathcal{A}} P(X \le x \mid Z=1, A=a)P(A=a \mid Z=1)P(Z=1) +\\
&\quad \sum_{a\in\mathcal{A}} P(X \le x \mid Z=1, A=a))P(A=a \mid Z=0)P(Z=0)\\
&= \sum_{a\in\mathcal{A}} P(X \le x \mid Z=1, A=a)P(A=a, Z=1) + \sum_{a\in\mathcal{A}} P(X \le x \mid Z=1, A=a)P(A=a, Z=0)\\
&= \sum_{a\in\mathcal{A}} P(X \le x \mid Z=1, A=a)P(A=a),
\end{aligned}$$

which is precisely (4.2.i). Similar arguments can be used to obtain (4.2.ii).

C Proof of Theorem 5.1

Proof of (5.7): Using (5.2) to see that $F_{Y|A,Z=0} < F_{X|A,Z=0} \le 1$, the result follows directly from decomposition (3.5) after marginalizing with respect to A. The proof of (5.8) is obtained similarly using (5.1) and (3.7).

Proof of (5.9): Note first that

$$\begin{aligned}
&\text{(i)} \quad F_X(t) = F_{X|A\in\mathcal{A}_1}(t)P(A \in \mathcal{A}_1) + F_{X|A\in\mathcal{A}_1^c}(t)P(A \in \mathcal{A}_1^c);\\
&\text{(ii)} \quad F_Y(t) = F_{Y|A\in\mathcal{A}_1}(t)P(A \in \mathcal{A}_1) + F_{Y|A\in\mathcal{A}_1^c}(t)P(A \in \mathcal{A}_1^c),
\end{aligned} \qquad \text{(C.1)}$$

where

$$\begin{aligned}
&\text{(i)} \quad F_{X|A\in\mathcal{A}_1}(t) = F_{X|A\in\mathcal{A}_1,Z=1}(t)P(Z=1 \mid A\in\mathcal{A}_1) + F_{X|A\in\mathcal{A}_1,Z=0}(t)P(Z=0 \mid A\in\mathcal{A}_1);\\
&\text{(ii)} \quad F_{X|A\in\mathcal{A}_1^c}(t) = F_{X|A\in\mathcal{A}_1^c,Z=1}(t)P(Z=1 \mid A\in\mathcal{A}_1^c) + F_{X|A\in\mathcal{A}_1^c,Z=0}(t)P(Z=0 \mid A\in\mathcal{A}_1^c);\\
&\text{(iii)} \quad F_{Y|A\in\mathcal{A}_1}(t) = F_{Y|A\in\mathcal{A}_1,Z=1}(t)P(Z=1 \mid A\in\mathcal{A}_1) + F_{Y|A\in\mathcal{A}_1,Z=0}(t)P(Z=0 \mid A\in\mathcal{A}_1);\\
&\text{(iv)} \quad F_{Y|A\in\mathcal{A}_1^c}(t) = F_{Y|A\in\mathcal{A}_1^c,Z=1}(t)P(Z=1 \mid A\in\mathcal{A}_1^c) + F_{Y|A\in\mathcal{A}_1^c,Z=0}(t)P(Z=0 \mid A\in\mathcal{A}_1^c),
\end{aligned} \tag{C.2}$$

for all $t \in \mathcal{W}$. In (C.2.i), $F_{X|A\in\mathcal{A}_1,Z=0}$ is unidentified. Using (5.3), the corresponding lower bound of $F_{X|A\in\mathcal{A}_1}$ is given by

$$F_{X|Z=1,A\in\mathcal{A}_1}(t)P(Z=1 \mid A\in\mathcal{A}_1) + F_{Y|Z=0,A\in\mathcal{A}_1}(t)P(Z=0 \mid A\in\mathcal{A}_1). \tag{C.3}$$

Furthermore, $F_{X|A\in\mathcal{A}_1,Z=0} \leq 1$ and, therefore, the corresponding upper bound of $F_{X|A\in\mathcal{A}_1}$ is given by

$$F_{X|Z=1,A\in\mathcal{A}_1}(t)P(Z=1 \mid A\in\mathcal{A}_1) + P(Z=0 \mid A\in\mathcal{A}_1). \tag{C.4}$$

Consider now (C.2.ii). Taking into account that no self-selection condition is assumed for those examinees scoring $a \in \mathcal{A}_1^c$, it follows that the lower bound of $F_{X|A\in\mathcal{A}_1^c}$ is given by

$$F_{X|Z=1,A\in\mathcal{A}_1^c}(t)P(Z=1 \mid A\in\mathcal{A}_1^c); \tag{C.5}$$

and the corresponding upper bound by

$$F_{X|Z=1,A\in\mathcal{A}_1^c}(t)P(Z=1 \mid A\in\mathcal{A}_1^c) + P(Z=0 \mid A\in\mathcal{A}_1^c). \tag{C.6}$$

Combining (C.3) and (C.2.i), and (C.5) and (C.2.ii), the lower identification bound for F_X in (C.1.) is obtained as

$$\begin{aligned}
F_X(t) &\geq \left\{F_{X|Z=1,A\in\mathcal{A}_1}(t)P(Z=1 \mid A\in\mathcal{A}_1) + F_{Y|Z=0,A\in\mathcal{A}_1}(t)P(Z=0 \mid A\in\mathcal{A}_1)\right\} P(A\in\mathcal{A}_1)\ + \\
&\quad F_{X|Z=1,A\in\mathcal{A}_1^c}(t)P(Z=1 \mid A\in\mathcal{A}_1^c)\, P(A\in\mathcal{A}_1^c) \\
&= P(X\leq t, Z=1, A\in\mathcal{A}_1) + P(Y\leq t, Z=0, A\in\mathcal{A}_1) + P(X\leq t, Z=1, A\in\mathcal{A}_1^c) \\
&= F_{X|Z=1}(t)P(Z=1) + F_{Y|Z=0,A\in\mathcal{A}_1}(t)P(Z=0, A\in\mathcal{A}_1).
\end{aligned}$$

Similarly, combining (C.4) and (C.2.i), and (C.6) and (C.2.ii), the upper bound for F_X in (C.1.) is obtained as

$$\begin{aligned}
F_X(t) &\leq \left\{F_{X|Z=1,A\in\mathcal{A}_1}(t)P(Z=1 \mid A\in\mathcal{A}_1) + P(Z=0 \mid A\in\mathcal{A}_1)\right\} P(A\in\mathcal{A}_1)\ + \\
&\quad \left\{F_{X|Z=1,A\in\mathcal{A}_1^c}(t)P(Z=1 \mid A\in\mathcal{A}_1^c) + P(Z=0 \mid A\in\mathcal{A}_1^c)\right\} P(A\in\mathcal{A}_1^c) \\
&= P(X\leq t, Z=1, A\in\mathcal{A}_1) + P(Z=0, A\in\mathcal{A}_1) + P(X\leq t, Z=1, A\in\mathcal{A}_1^c) + \\
&\quad P(Z=0, A\in\mathcal{A}_1^c) \\
&= F_{X|Z=1}(t)P(Z=1) + P(Z=0).
\end{aligned}$$

The partial identification intervals for F_Y are obtained using similar arguments.

■

D Partial identification of the quantiles for the self-selection case

Let us discuss the partial identification of the quantiles under assumptions (A.5) and (A.6). 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 in Section 4.2.2, the following theorem follows:

Theorem D.1 *In the NEAT design, under the assumptions(A.5) and (A.6), the quantiles of the partially identified probability distributions F_X and F_Y are partially identified by the following intervals:*

$$\begin{aligned} &\textit{(i)} \quad u_X(\alpha) \leq q_X(\alpha) \leq v(\alpha); \\ &\textit{(ii)} \quad u_Y(\alpha) \leq q_Y(\alpha) \leq v(\alpha). \end{aligned} \tag{D.7}$$

One of the conclusions of Theorem 4.2 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 assumptions (A.5) and (A.6), this situation is improved for the upper bound. As a matter of fact, the identification intervals (D.7.i) and (D.7.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

$$\delta_* = \inf\left\{\delta \in (0,1)\ :\ \exists\, \epsilon \in (0,1)\ \text{ s.t. } P(Z=1)\,\delta + P(Z=0)\,\epsilon = \alpha,\ \ F^{-1}_{X|Z=1}(\delta) \geq F^{-1}_{Y|Z=0}(\epsilon)\right\} \tag{D.8}$$

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

$$\epsilon_* = \frac{\alpha - P(Z=1)\delta_*}{P(Z=0)}. \tag{D.9}$$

Then

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

This equality shows that $v(\alpha)$ will be informative, as can be seen in Tables 1, 2 and 3.

Table 1: Lower and upper bounds for $q_X(\alpha)$ and $q_Y(\alpha)$ when $P(Z=1) < P(Z=0)$ under assumptions (A.5) and (A.6); δ_* and ϵ_* are defined by (D.8) and (D.9), respectively

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

Table 2: Lower and upper bounds for $q_X(\alpha)$ and $q_Y(\alpha)$ when $P(Z=1) > P(Z=0)$ under assumptions (A.5) and (A.6); δ_* and ϵ_* are defined by (D.8) and (D.9), respectively

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

References

Bernard, C., & Vanduffel, S. (2015). Quantile of a Mixture with Application to Model Risk Assessment.

Table 3: Lower and upper bounds for $q_X(\alpha)$ and $q_Y(\alpha)$ when $P(Z = 1) = P(Z = 0)$ under assumptions (A.5) and (A.6); δ_* and ϵ_* are defined by (D.8) and (D.9), respectively

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

Dependence Modeling, *3*, 172–181.

Bhide, A., Shah, P. S., & Acharya, G. (2018). A simplified guide to randomized controlled trials. *Acta Obstetricia et Gynecologica Scandinavica*, *97*(4), 380–387.

Odgaard-Jensen, J., Vist, G. E., Timmer, A., Kunz, R., Akl, E. A., Schünemann, H., … Oxman, A. D. (2011). Randomisation to protect against selection bias in healthcare trials. *Cochrane Database of Systematic Reviews*(4).

Stephenson, J., & Imrie, J. (1998). Why do we need randomised controlled trials to assess behavioural interventions? *BMJ*, *316*(7131), 611–613.