An Alternative View on the NEAT Design in Test Equating

Jorge González and Ernesto San Martín

Abstract Assuming a "synthetic population" and imposing strong assumption to estimate score distributions has been the traditional practice when performing equating under the nonequivalent groups with anchor tests design (NEAT). In this paper, we use the concept of partial identification of probability distributions to offer an alternative to this traditional practice in NEAT equating. Under this approach, the score probability distributions used to obtain the equating transformation are bounded on a region where they are identified by the data. The advantages of this approach are twofold: first, there is no need to define a synthetic population and, second, no particular assumptions are needed to obtain bounds for the score probability distributions that are used to build the equating transformation. The results show that the uncertainty about the score probability distributions, reflected on the width of the bounds, can be very large, and can thus have a big impact on equating.

Keywords Test equating · NEAT design · Partial identifiability · Ignorability condition

1 Introduction

Test equating is used to make scores from different test forms comparable. An equating transformation function is used to map the scores on one scale into their equivalents on the other. Before this score transformation takes place, it is necessary to control for test takers ability differences, and different data collection designs have been described in the equating literature for such purpose (von Davier et al. 2004, Chap. 2; Kolen and Brennan 2014, Sect. 1.4 and González and Wiberg 2017, Sect. 1.3.1). These equating designs differ in that either common persons or common items are

J. González (✉) · E. San Martín
Faculty of Mathematics, Pontificia Universidad Católica de Chile,
Av. Vicuña Mackenna, 4860 Macul, Santiago, Chile
e-mail: jorge.gonzalez@mat.uc.cl

E. San Martín
e-mail: esanmart@mat.uc.cl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_10

used to perform the score transformation. In this paper we will focus the attention on the nonequivalent groups with anchor test design (NEAT).

The NEAT design is widely used in test equating. Under this design, two groups of test takers are administered separate test forms with each test form containing a common subset of items. Because test takers from different populations are assigned only one test form, missing score data emerge by design rendering some of the score distributions unavailable. The equating literature has treated this problem from different perspectives all of them making different assumptions in order to estimate the missing score distributions. In this paper, we offer an alternative view that is free of these types of assumptions to obtain the score distributions under a NEAT design.

We first argue that, rather than viewing the problem as one of missing data, there is an inherent identifiability problem underlying the NEAT design. Then, we further argue that the typical assumptions on the equality of conditional distributions are nothing more that identifiability restrictions. Because these assumptions might be too strong, and, moreover, are not empirically testable, we offer an alternative that does not make use of any assumption and show that the non identified score distributions are actually partially identified, deriving bounds for them on the partially identified region.

The rest of this paper is organized as follows. We first briefly revisit the current view on the NEAT design, including the definition of synthetic population and the assumptions commonly made to estimate score distributions. Then we introduce our view on the NEAT design as an identifiability problem and derive bounds where the non identified score distributions are partially identified. An illustration using an hypothetical data example appearing in the equating literature is presented. The paper ends with final remarks and ideas for future work.

2 NEAT Equating: The Current and an Alternative View

2.1 Notation and Preliminaries

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be the random variables representing test scores from tests forms X, Y. As mentioned before, the equating function $\varphi : \mathcal{X} \mapsto \mathcal{Y}$ defined as $\varphi(x) = F_Y^{-1}(F_X(x))$ maps the scores on the scale $\mathcal{X}$ into their equivalents on the $\mathcal{Y}$ scale (González and Wiberg 2017). This definition is established for φ defined on a common population where the equating is to be performed (Braun and Holland 1982). Accordingly, the score cumulative distribution functions used to build the equating transformation, should also be defined on a common population that will be denoted as T.

When single groups (SG), equivalent groups (EG) or counter balanced groups (CB) equating designs are considered, defining φ on a common population does not constitute a problem as samples of test takers are in fact taken from the same population. However, this is not the case for the NEAT design where samples of test takers come from two different populations, called here P and Q. As a consequence,

score distributions of X and Y are defined in both P and Q and we denote these distributions here as $F_{XP}(x)$, $F_{YP}(y)$, $F_{XQ}(x)$, and $F_{YQ}(x)$, respectively.

2.2 NEAT Equating: The Current View

To solve the problem of defining the equating transformation on a common population, the equating literature has resorted in what is called a *synthetic population* (Braun and Holland 1982). This definition conceptualizes a common population as a weighted combination of P and Q in the form

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

where w is a weight such that $0 \leq w \leq 1$. Using this definition, the corresponding score distributions used to build the equating transformation are obtained as

$$\begin{aligned} F_{XT}(x) &= wF_{XP}(x) + (1 - w)F_{XQ}(x) \\ F_{YT}(y) &= wF_{YP}(y) + (1 - w)F_{YQ}(y). \end{aligned} \tag{2}$$

A typical representation of the NEAT equating design is shown in Table 1. From the table, it can be seen that because test takers in P are only administered test X and those in Q are only administered Y, the corresponding score distributions F_{XQ} and F_{YP} needed to obtain F_{XT} and F_{YT} in (2) are said to be *missing*. Additional assumptions are thus needed to estimate them, and here is where the anchor test, A, has played a fundamental role. Most commonly, it is assumed that the conditional score distributions of X and Y given A are the same in both population: $F_{XP}(x \mid a) = F_{XQ}(x \mid a)$ and $F_{YP}(y \mid a) = F_{YQ}(y \mid a)$, with $A \in \mathcal{A}$. Using these assumptions, and the fact that marginal distributions of A are indeed observed in both populations, the score distributions of X and Y in T are obtained by marginalizing the joint distributions over A. The obtained score distributions are then used to build $\varphi(x) = F_{YT}^{-1}(F_{XT}(x))$.

2.3 NEAT Equating: An Alternative View

Rather than facing missing score distribution, what happens in reality is that the sampling process underlying the NEAT design does not give information on F_{YP} and F_{XQ}, and thus the target score distributions $F_{XT}(x)$ and $F_{YT}(y)$ are not identified.

Table 1 Schematic representation of the NEAT design

Population	Sample	X	Y	A
P	1	✓		✓
Q	2		✓	✓

Note X and Y are test forms. A is an anchor test

Moreover, the assumptions on equality of conditional score distributions are actually identification restrictions.

To introduce these ideas better, let us briefly revisit the definition of identifiability. If θ is a parameter indexing a family of distributions $\{f(x \mid \theta) : \theta \in \Theta\}$, then θ is said to be identified if distinct values of it lead to distinct probability distributions (Casella and Berger 2002). Equivalently, if the probability distribution can be uniquely determined by θ, then θ is identified. If the probability distribution cannot be uniquely determined (i.e., the model is not identified), putting certain restrictions on the parameter space can make the model identifiable.

In what follows, we show that the score distributions needed to build the equating transformation are identified on a bounded region. No assumptions or restrictions are needed for the derivation of these bounds.

2.3.1 Conditional Score Distributions with No Assumptions

Although the marginal score distributions are of main interest to build the equating transformation, we start analyzing the conditional score distributions as they are typically used in NEAT equating.

Let Z be a binary variable such that

$$Z = \begin{cases} 1, \text{ if test taker is administered X in } P; \\ 0, \text{ if test taker is administered Y in } Q. \end{cases} \tag{3}$$

Then, by the law of total probability (Kolmogorov 1950), it follows that

$$\begin{aligned} \text{(a) } P(X \leq x \mid A) =& P(X \leq x \mid A, Z = 1)P(Z = 1 \mid A) + \\ & P(X \leq x \mid A, Z = 0)P(Z = 0 \mid A), \\ \text{(b) } P(Y \leq y \mid A) =& P(Y \leq y \mid A, Z = 1)P(Z = 1 \mid A) + \\ & P(Y \leq y \mid A, Z = 0)P(Z = 0 \mid A). \end{aligned} \tag{4}$$

The statistical model underlying the NEAT design is accordingly parameterized by the parameters $\{P(X \leq x \mid A = a), P(Y \leq y \mid A = a)\}$. In order to show that these parameters are not identified, consider the following comments on (4):

1. $P(X \leq x \mid A = a, Z = 1)$ is the conditional score probability of X given A for a test taker who actually answered form X (i.e., sampled from P) and scored $A = a$ on the anchor test.
2. $P(Z = 1 \mid A = a)$ corresponds to the proportion of test takers who were administered form X (or equivalently, proportion of people sampled from P) and scored $A = a$ on the anchor test.
3. $P(Z = 0 \mid A = a)$ corresponds to the proportion of test takers who were administered form Y (or equivalently, proportion of people sampled from Q) and scored $A = a$ on the anchor test.

4. $P(X \leq x \mid A = a, Z = 0)$ is the conditional score probability of X for a test taker who was actually administered form Y (or sampled from Q).

Consequently, $P(X \leq x \mid A = a)$ corresponds to the probability of scoring x on test form X *as if* all test takers with a score $A = a$ were administered test form X. However, this conditional probability is *not identified*. As a matter of fact, the data generating process that underlies the NEAT design only identifies $P(X \leq x \mid A = a, Z = 1)$ and $P(Z = z \mid A)$ for $z \in \{0, 1\}$. However, it does not provide any information about $P(X \leq x \mid A = a, Z = 0)$ and therefore the sampling process only reveals that

$$\begin{aligned} P(X \leq x \mid A = a) =& P(X \leq x \mid A = a, Z = 1)P(Z = 1 \mid A = a) + \\ & \gamma\, P(Z = 0 \mid A) \end{aligned}$$

for some *unknown* probability distribution γ. Therefore, $P(X \leq x \mid A = a)$ cannot be uniquely determined because γ can not be uniquely chosen. Consequently, $P(X \leq x \mid A = a)$ is not identified. Similar conclusions can be drawn for $P(Y \leq y \mid A)$.

In practice, $P(X \leq x \mid A)$ and $P(Y \leq y \mid A)$ are identified under an hypothesis of strong ignorability (e.g., Rosenbaum and Rubin 1983), namely

$$\begin{aligned} P(X \leq x \mid A, Z = 1) =& P(X \leq x \mid A, Z = 0) = P(X \leq x \mid A), \\ P(Y \leq y \mid A, Z = 1) =& P(Y \leq y \mid A, Z = 0) = P(Y \leq y \mid A), \end{aligned} \tag{5}$$

which, in the context of the current application can compactly be defined as

$$(X, Y) \perp\!\!\!\perp Z \mid A. \tag{6}$$

As a matter of fact, the strong ignorability condition essentially tells us that γ is not unknown, but it coincides with $P(X \leq x \mid A = a, Z = 1)$. This implies that $P(X \leq x \mid A = a)$ is uniquely determined, and thus identified. It is necessary to emphasize that the strong ignorability condition cannot empirically be refuted and, therefore, it should be justified in the context of an application.

2.3.2 Partially Identified Probability Distributions

The strong ignorability condition can be avoided if we find a region where the score probabilities are actually identified. In this section we show that such region indeed exists. As a matter of fact, because $P(X \leq x \mid A, Z = 0)$ is bounded between 0 and 1, from (4) it can easily be verified that

$$L_x \leq P(X \leq x \mid A) \leq U_x, \tag{7}$$

where

$$\begin{aligned} L_x =& P(X \leq x \mid A, Z = 1)P(Z = 1 \mid A) \\ U_x =& P(X \leq x \mid A, Z = 1)P(Z = 1 \mid A) + P(Z = 0 \mid A) \end{aligned} \tag{8}$$

Analogously for Y it can be verified that

$$L_y \leq P(Y \leq y \mid A) \leq U_y, \tag{9}$$

where

$$\begin{aligned} L_y =& P(Y \leq y \mid A, Z = 0)P(Z = 0 \mid A) \\ U_y =& P(Y \leq y \mid A, Z = 0)P(Z = 0 \mid A) + P(Z = 1 \mid A) \end{aligned} \tag{10}$$

Thus, the conditional score distributions are partially identified (Tamer 2010) on regions defined by the derived bounds. Note that the length of the intervals for $P(X \leq x \mid A)$ and $P(Y \leq y \mid A)$ are $P(Z = 0 \mid A)$ and $P(Z = 1 \mid A)$, respectively, and as mentioned before they correspond to the proportion of test takers in P and Q, respectively, for a given score A.

2.3.3 Marginal Distributions with No Assumptions

The equating transformation φ is built from marginal score distributions defined on a common population. It is thus of interest to examine if the preceding arguments are also valid when the conditional distributions are marginalized over the anchor scores. It is easy to see that marginalizing over A in (4) we obtain

$$P(X \leq x) = P(X \leq x \mid Z = 1)P(Z = 1) + P(X \leq x \mid Z = 0)P(Z = 0). \tag{11}$$

Note that the identifiability problem still remains in the marginal score distribution as $P(X \leq x \mid Z = 0)$ is non identified. However, because this probability is bounded between 0 and 1, we can show similarly as before that $P(X \leq x)$ can also be bounded. In fact,

$$L_x \leq P(X \leq x) \leq U_x, \tag{12}$$

where

$$\begin{aligned} L_x =& P(X \leq x \mid Z = 1)P(Z = 1) \\ U_x =& P(X \leq x \mid Z = 1)P(Z = 1) + P(Z = 0) \end{aligned} \tag{13}$$

Note that using the definition in (3), Eq. (11) can be rewritten as

$$F_X(x) = wF_{XP}(x) + (1 - w)F_{XQ}(x) \tag{14}$$

with $w = P(Z = 1)$. Interestingly, the right hand sides of Eqs. (2) and (14) are *visually* identical. This result would indicate that the weights in the definition of a synthetic population are actually related to the proportion of test takers in the populations and thus should not be arbitrarily chosen. Moreover, w corresponds to the length of the interval where the score distribution is partially identified. Analogous results as the ones shown in (11), (12), (13), and (14) can be derived for $P(Y \leq y)$.

A natural question at this stage is how $F_X(x)$ and $F_Y(y)$ compare to $F_{XT}(x)$ and $F_{YT}(y)$, respectively. Such comparison is not possible because the formers distributions are not identified and thus non observable. We have shown that they are however partially identified on a bounded region so that it is possible to evaluate the behavior of the bounds and how it relates to the target distributions traditionally obtained in NEAT equating using the definition of synthetic population and the ignorability condition. This is done in the following section.

Table 2 Bivariate score frequencies (X, A) and (Y, A)

X	A	Frequency	Y	A	Frequency
0	0	4	0	0	4
0	1	4	0	1	3
0	2	2	0	2	1
0	3	0	0	3	0
1	0	4	1	0	7
1	1	8	1	1	5
1	2	2	1	2	7
1	3	1	1	3	1
2	0	6	2	0	3
2	1	12	2	1	5
2	2	5	2	2	12
2	3	2	2	3	2
3	0	3	3	0	3
3	1	12	3	1	4
3	2	5	3	2	13
3	3	5	3	3	5
4	0	2	4	0	2
4	1	3	4	1	2
4	2	4	4	2	5
4	3	6	4	3	6
5	0	1	5	0	1
5	1	1	5	1	1
5	2	2	5	2	2
5	3	6	5	3	6

3 Illustrations

3.1 Data

We use data from an hypothetical example shown in Kolen and Brennan (Kolen and Brennan (2014), Sect. 5.1.3). In this example forms X and Y each contain 5 items and 3 common items. The data in Kolen and Brennan (2014) are originally displayed as joint probabilities $f_{XP}(x,a) = P(X = x, A = a)$ and $f_{YQ}(y,a) = P(Y = y, A = a)$ and we use this information to create raw data as displayed in Table 2. The table shows bivariate score frequencies for each test form. From the table, it can be seen that, for instance, 8 test takes scored $X = 1$ and $A = 1$, whereas 13 scored $Y = 3$ and $A = 2$, etc. For the information in the table (frequency), it follows that the sample size considered is 100 for both populations.

3.2 Results

Figure 1 shows a graphical representation of the bounds derived in (8) for the case when $A = 2$. From the figure, it can be seen that the bounds for the conditional distribution of X given A are wider than the ones for the conditional distribution of Y given A, when $A = 2$. Note, however, that this situation could change for other values of the anchor score. Moreover, the curves are *parallel* in the sense that the length of the intervals are constant for all values of scores on the scale, for a given value of A.

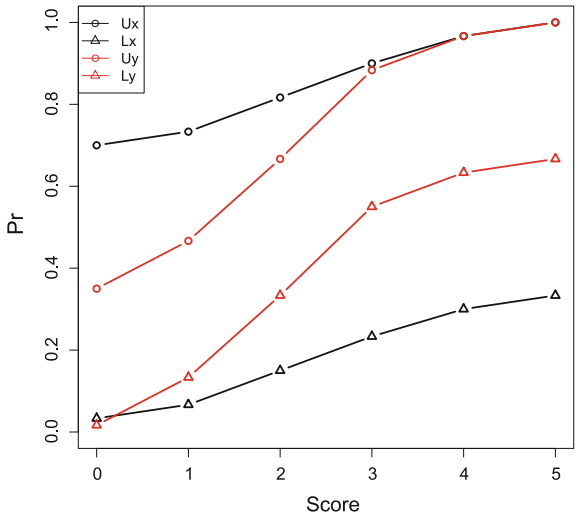

Fig. 1 Bounds for conditional score distributions $P(X \leq x \mid A = 2)$ and $P(Y \leq y \mid A = 2)$

Table 3 Target cumulative distributions for Forms X and Y scores, and derived bounds

Score	F_{XT}	$[L_x, U_x]$	F_{YT}	$[L_y, U_y]$
0	0.100	[0.050; 0.550]	0.105	[0.040; 0.540]
1	0.250	[0.125; 0.625]	0.320	[0.140; 0.640]
2	0.500	[0.250; 0.750]	0.530	[0.250; 0.750]
3	0.750	[0.375; 0.875]	0.755	[0.375; 0.875]
4	0.900	[0.450; 0.950]	0.900	[0.450; 0.950]
5	1.000	[0.500; 1.000]	1.000	[0.500; 1.000]

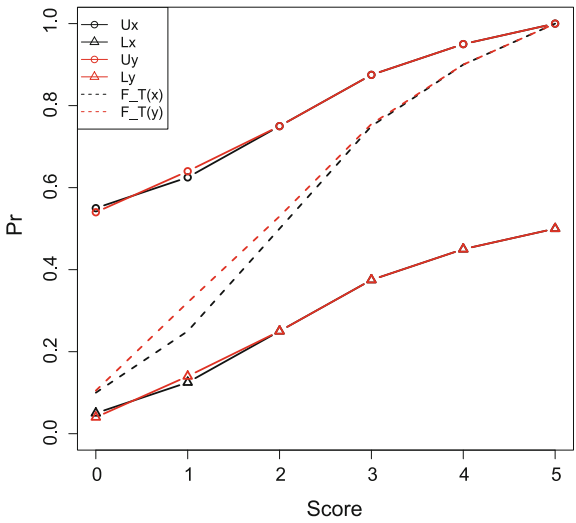

Fig. 2 Bounds for $F_X(x) = P(X \leq x)$ and $F_Y(y) = P(Y \leq y)$, and target score distributions $F_{XT}(x)$ and $F_{YT}(y)$ for the case $w = 1$

This is due to the fact that, as seen at the end of Sect. 2.3.2, the length of the intervals are defined by $P(X \leq x \mid A)$ and $P(Y \leq y \mid A)$.

Next, we calculated the bounds derived in Sect. 2.3.3 for each of the marginal score distributions. Because the real value of F_X and F_Y is unknown, we use the derived target cumulative distribution functions F_{XT} and F_{YT} as reference for comparison. The latter where obtained assuming that $w = 1$. Table 3 shows the target cumulative distributions and the corresponding bounds where the marginal score distributions are partially identified. Figure 2 shows a graphical representation of these results.

From Table 3 and Fig. 2, it can be seen that all the values of F_{XT} and F_{YT} lie in the intervals $[L_x, U_x]$ and $[L_y, U_y]$, respectively, as expected. Note also that the intervals have length equal to 0.5. This is because the sample sizes in both populations is exactly the same (100 in this case), so that $P(Z = 1) = P(Z = 0) = \frac{100}{200} = 0.5$ (see comments on Sect. 2.3.3 below Eq. (14)).

4 Concluding Remarks

In this paper, we have argued that there is an inherent identification problem underlying the NEAT equating design. The assumption on the equality of conditional score distributions, typically made in NEAT equating and called here an ignorability condition, has been shown to actually be an identification restriction. We offered an alternative to the ignorability condition and proposed to work with partially identified probability distributions.

The derived bounds on the partially identified region showed that there is huge uncertainty about the probability distributions that are to be used for equating. The actual impact of this method on equating is currently being investigated by the authors.

The exposition focused on poststratification equating under the NEAT design. However, the identifiability problem also arises for the case when chained equipercentile equating (e.g., Kolen and Brennan 2014) is used to equate score data collected under the NEAT design. In fact, different assumptions are needed to identify the target score distributions used to build the equating transformation (see, e.g., von Davier et al. 2004, Sect. 2.4.1). The derivation of bounds where the score distributions are partially identified for the case of chained equating is currently being investigated by the authors.

Acknowledgements Jorge González acknowledges partial support of grant Fondecyt 1150233. Ernesto San Martín acknowledges partial support of grant Fondecyt 1141030.

References

Braun, H., & Holland, P. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. Holland, D. Rubin (Eds.), *Test Equating*, (Vol. 1, pp. 9–49). Academic Press.

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

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

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

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

Kolmogorov AN. (1950). Foundations of the theory of probability. Chelsea Publishing Co.

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.

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