TPM®

A HEURISTIC APPROACH TO LINK SOME UNLINKABLE TESTS

RENATO MICELI
UNIVERSITY OF TORINO
UNIVERSITY OF VALLE D'AOSTA

DAVIDE MARENGO
MICHELE SETTANNI
UNIVERSITY OF TORINO

The ability to place different tests on a common metric, that is, test equating, is essential for many purposes in psychological and educational research. When traditional linking designs are not viable, however, the ability to perform equating is compromised. To solve this problem, authors have proposed the use of alternative equating procedures based on the identification of pseudocommom items or collateral data. Here we propose a heuristic allowing for the estimation of the intercept linking parameter using as the sole input the Rasch item difficulties of the tests. Using simulated data, the heuristic showed high accuracy and efficiency in estimating the intercept linking parameter (MAE = 0.15 logit; RMSE = 0.21 logit; R^2= .87). Test length and true-parameter size had a small-to-moderate impact on the accuracy of parameter estimation. Results are discussed in light of implications for practical purposes, as well as future improvements of the proposed heuristic.

Key words: Test equating; Test linking; Simulation; Item response theory; Rasch model.

Correspondence concerning this article should be addressed to Davide Marengo, Department of Psychology, University of Torino, Via Verdi 10, 10124 Torino (TO), Italy. Email: davide.marengo@unito.it

In this study, we propose a new heuristic allowing for the estimation of the coefficients needed to link test administrations for which information required to establish this link is not available. In particular, first, we review the existing approaches to horizontal linking and highlight their limits in the case of tests sharing no common data. Then, in Section "A Heuristic Approach to Link Tests when no Linking Data is Available," we present the heuristic and describe the main concepts of how it functions. In Section "Evaluation of the Performance of the Heuristic," using simulated test response data, we evaluate the functioning of this heuristic in estimating linking coefficients allowing for the equating of test administrations for which no linking information — such as that provided by common items, common person, or collateral information — is available. In Section "Results," results are discussed.

EQUATING OF TEST FORMS WITH LITTLE TO NO DATA: LINKING DESIGNS AND LIMITS

The ability to place scores of multiple test forms on a common metric is essential for a variety of important practical and theoretical purposes in educational research, for example, to allow the examination of test fairness and measurement invariance (Dorans, 2004), the comparison of the ability of groups, and the investigation of trends of educational achievement (Kolen & Brennan, 2004). Test equating describes a

group of methods that enable researchers and test practitioners to achieve these aims, and to obtain the additive and/or multiplicative constants needed to rescale scores from different tests on the same metric (e.g., Angoff, 1971; Lord, 1980; Kolen & Brennan, 2004; Wright & Stone, 1979).

When equating is performed between two alternative versions of the same test (e.g., equating of parallel forms), it is commonly referred to as horizontal equating, so as to distinguish it from vertical equating, which is usually performed on tests varying in difficulty levels. In the case of horizontal linking of educational tests, which will be the focus of the present study, one of the most widely used equating designs is the nonequivalent groups with anchor test design (NEAT; Kolen & Brennan, 2004). In this design, two tests forms are administered to different populations of examinees and a set of common items — generally referred to as the *anchor test* — is included in both test forms. The anchor test, then, provides the internal statistical link required to put the scores on the two test forms on a common metric scale. An alternative administration approach is the "common item equating to a calibrated item pool" design (Kolen & Brennan, 2004; equating-to-a-pool design, for short). In this design the test forms include a set of items selected from a common item pool consisting of previously calibrated items. The item pool, thus, serves the role of a common external anchor (i.e., the statistical link) for the two forms. Alternative equating designs also exist which are based on the assumption that the groups taking different tests are extracted from the same population (equivalent groups equating; Kolen & Brennan, 2004), or nonequivalent groups can be equated using sample matching techniques based on collateral information about individual characteristics (e.g., Powers & Kolen, 2012; Wright & Dorans, 1993). Use of these approaches may not be appropriate for the linking of high-stakes tests — i.e., when test scores are used for gatekeeping purposes, such as to determine students' admission to the next cycle of education (Nagy, 2000). In case of high-stakes testing, consecutive test administration for the same grade level generally do not include common items, as it may result in unwanted exposure of items and breaches of security. Further, tests are generally administered to different populations (e.g., students at the same grade level taking high-stakes tests in different years, or in different countries), rendering the use of linking approaches leveraging examinees' characteristics, such as random-equivalent groups equating (Kolen & Brennan, 2004), or common-person linking (Masters, 1985), not feasible. To deal with this problematic lack of available information, previous studies have proposed tentative solutions, mostly leveraging pseudocommon items (e.g., Fisher, 1997; Kaspersen, Pepin, & Sikko, 2017), collateral data about both items (Hsu, Wu, Yu, & Lee, 2002; Mislevy, Sheehan, & Wingersky, 1992) and persons (Kim, Livingston, & Lewis, 2011), or post-test anchor data (e.g., Marengo, Miceli, Rosato, & Settanni, 2018). However, because they require tests that are almost identical in content, presume the availability of collateral data, or additional test administrations, these methods are not suitable for general application. Further, equating approaches based on sample matching are also not generally feasible due to the by-design lack of collateral information about examinees linked with privacy regulations of educational tests.

Still, in all these situations, the possibility to establish a link between test forms would be beneficial for different purposes, such as the need to compare the ability of groups, to determine the stability of testing conditions across test administrations (Miceli, Marengo, Molinengo, & Settanni, 2015), as well as the possibility to obtain a calibrated item bank (Miceli & Molinengo, 2005). For this reason, there remains a need for methods which allow accurate equating in situations when no data is available to perform the linking of the tests. In the absence of information allowing the linking of alternative test forms, researchers and practitioners find themselves in a situation in which they can either restrain from using test data, or assume the tests to be randomly equal, which we may refer to as *naïve linking*, both representing suboptimal solutions to the linking problem.

In the following sections, we present a new heuristic which permits the linking of test administrations sharing no common item or persons, as well as any other collateral information thus allowing the equating of the tests. Finally, we provide information about accuracy of the heuristic on simulated response data.

A Heuristic Approach to Link Tests when no Linking Data is Available

Let **X** and **Y** be two vectors of item difficulties of order *i* and *j* obtained by applying the Rasch model for dichotomous responses on response data collected administering two tests, each respectively including *i* and *j* items measuring the same latent ability, to two nonequivalent groups of examinees. The tests are assumed to have no items in common. Further, difficulty estimates are calibrated independently for **X** and **Y**, and in both vectors the average difficulty is fixed to 0.

Although it is reasonable to assume that both vectors can be considered random samples extracted from the same universe including all possible items for that specific domain (Bejar, 1984; Briggs & Wilson, 2007), it is certain that the values of the two vectors, when centered, are not directly comparable between them. This problem can be solved, and is generally expressed in terms of affine transformations; so, for example, to put the values of the vector **Y** in the same metric as the vector **X** we will need to work with an equation of the type: $y_i^* = a + b \cdot y_i$, and to know the values of the parameters *a* and *b* we resort to the usual linking strategies. For example, in the case in which the two tests include "common item," the *b* parameter, also known as the slope parameter, is expected to be positive and approximately equal to a unit ($b \cong +1$), so that the transformation will consist essentially of a translation of the same amount of the value of the *a* parameter, which we will refer to as the intercept linking parameter (ILP). In the absence of appropriate linking strategies the parameters *a* and *b* remain unknown, and any comparison between the two vectors (**X** and **Y**) is precluded. In turn, two tests are "connected" when the values of the two vectors (**X** and **Y**) estimated difficulties are comparable; in item response theory (IRT), if you can relate to the difficulty of the items of two tests in a common metric scale, you can always get new estimates of the skills of individuals by anchoring these estimates to fix the new scale.

Given that the connection between the two tests is analytically impractical, the usual linking strategies (e.g., common-item or common-person design, or use of collateral information), may allow to perform the equating of the tests. However, when these strategies are, for different reasons, unavailable, every further attempt appears doomed to failure.

The heuristic hereinafter described is grounded in the idea that in these situations, it may be possible to proceed in a similar way to what happens in the presence of the common items, that is by assuming the slope parameter is fixed at +1, and that researchers are only required to estimate the ILP. Further, it is assumed that the estimation of the ILP, which is performed automatically by the computer, represents only the first phase required to establish the link between the tests. The second and final phase, which is performed manually by human operators, involves the examination of the semantic content of the items and is aimed at validating the solution proposed by the computer. More precisely, these two phases are: 1) the heuristic supplies a limited set of possible solutions, each of them in the form of subsets of items that can be coupled to work as common items, as well as the associated ILP needed to perform the linking; 2) based on the examination of the content of coupled items, the human operator determines the acceptability of the solutions provided by the computer.

The heuristic proposed here concerns only the first phase, and provides a limited number of possible solutions to make the connection between the two tests using only the two vectors **X** and **Y** of estimated

difficulties. As will be seen below, the operations that the heuristic needs to perform are conceptually simple, but computationally time consuming, making the use of the computer essential. These operations can be implemented as a sequence of algorithms. For the purpose of the present study, the algorithms were written using IML® in the SAS package.

In order for the heuristic proposed here to provide useful indications for the connection between two tests, it is necessary to assume that: (a) the two tests to be connected measure the same construct, even though they are made up of different items (i.e., tests that are intended for use in the same setting to assess the same trait or ability, e.g., INVALSI math tests for the same grade in different years); (b) item difficulty estimates for both tests are estimated using the Rasch model for dichotomous responses, constraining the difficulty of the items to have mean of 0; (c) in both tests, response data show good fit to the Rasch model (i.e., absence of variations in item discrimination, trait unidimensionality), and items are well distributed along the difficulty continuum.

Bearing in mind that, in this paper, we will always consider the situation in which we want to place the values of the vector **Y** in the reference context of **X** (and that, of course, all the above applies also in the opposite situation), the illustration of the operations to be completed can be divided into the three following steps:

Step 1. In the first step, we define the set of all the possible ILPs, which, under the given conditions, can be obtained from the two vectors (**X** and **Y**) of item difficulties. First, we identify all the possible couples of items from the two vectors **X** and **Y** and we compute Euclidean distances between the difficulties[1] of all possible couples of items. Next, we identify the groups of three or more item couples which include items showing the same distance in difficulty in both tests, that is those item couples whose item difficulties draw a regression line with slope equal to +1 and a coefficient of determination equal to 1. Because of this, we are able to focus only on linking operations involving translations[2], thus requiring only information about the ILP. In this way, given the two vectors **X** and **Y**, a bundle of parallel lines is obtained in which each line has a different intercept value indicating the point in which the line crosses the *x*-axis. These intercept values correspond to all the possible ILPs[3], and each of these values can be used to connect the estimates of the difficulty of the two tests. The couples of items that define each line are those subsets of items that "behave" just as you would expect if they were common to the two tests (that is, pairs of items that maintain the same difference in difficulty in the two tests). As an example, Figure 1 shows a series of lines resulting from five distinct item-couples sets (selected from a larger set of 73 item-couple sets) emerging from the two vectors of item difficulties, each including 30 items. In the figure, bolded points indicate item couples, which vary in number across lines, ranging from 3 to 5 item couples. For each line, the relative ILP is determined by the point where the line crosses the *x*-axis.

Step 2. Based on the estimates of the ILP obtained in the previous step using the coupled items, connections are established between the **X** and **Y** vectors. Performed operations consist in using the estimated ILP to link the vectors **X** and **Y** using a translation transformation. These operations are widely known, and coincide with those used in the "common item" situation (see e.g., Wright & Stone, 1979, p. 112 et seq.). The only difference is that we have to deal with several different ILPs, among which we are still not able to make a choice.

Step 3. We now look to define which estimated ILP represents an adequate estimate of the true ILP value linking vectors **X** and **Y**, and thus can be preferred to reconnect the two vectors. One way to proceed is to determine whether some property of the known vectors of difficulties (**X** and **Y**) is able to guide the choice. Having chosen to operate only and exclusively with translations, it is reasonable to direct

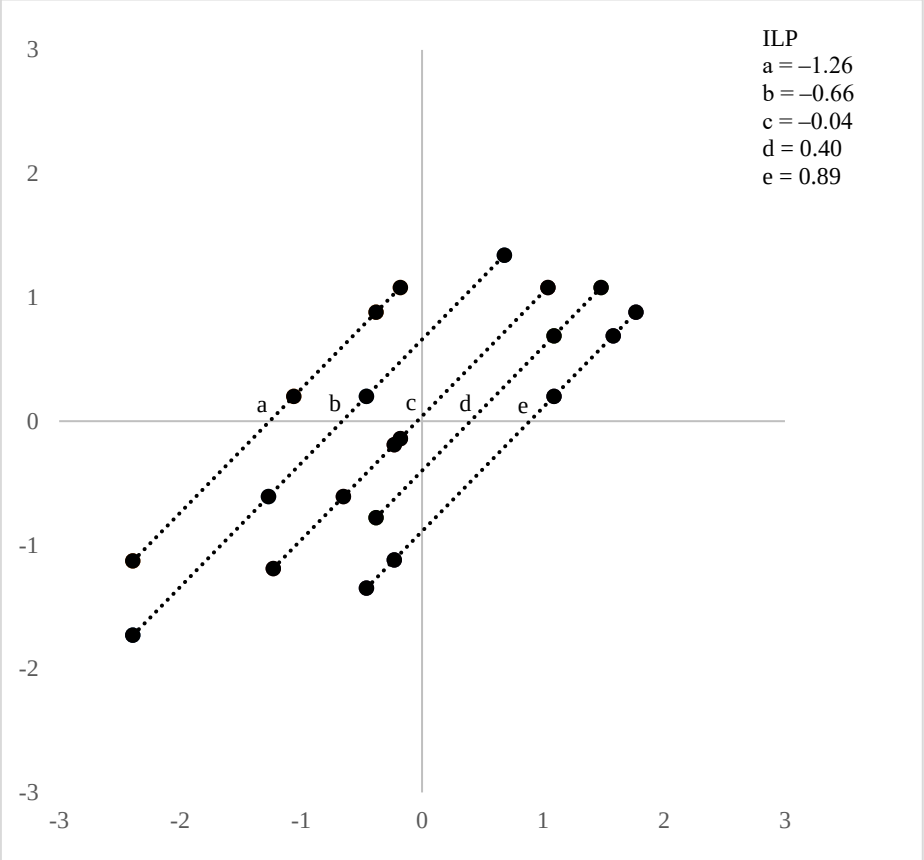

FIGURE 1
Five (out of 73) item-couples sets and relative estimated intercept linking parameter (ILP) resulting from two vectors of item difficulties (True ILP = –0.69 logits; 30 items per vector).
Note. [a-e] Letters indicate ILP values resulting from lines crossing the *x*-axis.

attention to those statistics that are invariant for such transformations and, among the different possibilities, attention is here directed to the range statistic. Remembering that all the available connected vectors are substantially the result of translations performed starting from the two known vectors, it follows that no range calculated on the vector resulting from the reconnection the two **X** and **Y** vectors may be lower than the maximum range of the two separate vectors. Of course, this does not mean that, in theory, the connected unknown vector cannot have a range lower than the maximum between the range of **X** and that of **Y**, but that none of the available vectors could coincide exactly with the unknown vector. On the other hand, intuitively, it seems reasonable to assume that the range of the unknown vector will be close enough to the calculated maximum range between the two known vectors.

Furthermore, it is also possible to make an "indirect" estimate of the range of the unknown vector, using for this purpose the vector (**X** or **Y**) which possesses, between the two known vectors, the maximum range. This estimate of the range of the unknown vector[4] provides a value greater than the maximum range calculated between the two known vectors. Thus, the maximum range (between these two known vectors) and the estimated range of the unknown connected vector provide two values that also delimit an interval in which the range of the vectors may be present. It should not be forgotten that each connected vector (obtained in Step 2) has a certain range, but above all that it was generated by a different ILP; therefore, saying

that in the resulting interval a certain number of connected vectors is present is equivalent to claiming that in this interval there exist several ILPs. It is therefore possible to reach a limited number of alternative solutions to the linking problem, each associated with a specific ILP connecting the two tests. In order to select the optimal ILP value among those present in the interval, the item couples corresponding to each ILP value should be examined and compared based on considerations related to their actual characteristics, including item format, content and assessed domain (Marengo, Miceli, & Settanni, 2016; Miceli et al., 2015).

Evaluation of the Performance of the Heuristic

The evaluation of the performance of the present heuristic was conducted by implementing a simulation design. Under this design, the input of the heuristic are vectors of the difficulties **X** and **Y** which are obtained by calibrating the Rasch model on two simulated matrices of dichotomous item responses. These matrices are generated using the SAS® Package, while difficulty and ability estimates were obtained by calibrating response data using the Winsteps Rasch Measurement software (ver 3.68.2; Linacre, 2009). All item response matrices, as well as the difficulty and ability estimates, are obtained with reference to the Rasch model for dichotomous responses. More specifically, item difficulty values are generated from a uniform distribution (U ± 2.5), and then randomly assigned to the two tests to be connected (Test A and Test B). For each test, 1,500 person ability values are generated following the normal distribution with random mean and variances (the means were obtained in the ± 0.75 logit interval, and variances varied between 1 and 3).

In the simulation design, we manipulate the following variables: (1) Number of items (*n*) in the two tests (20, 30, 40, 50 items); (2) the size of the true ILP parameter (A: $0 \leq |ILP| \leq 0.25$; B: $0.25 \leq |ILP| \leq 0.50$; C: $0.50 \leq |ILP| \leq 0.75$; D: $0.75 \leq |ILP| \leq 1.00$). Overall, 1,600 simulated cases were generated, 4 (number of item conditions) × 4 (ILP size conditions) × 100 (replications).[5] It is worthy to note that by using a simulation design, it is not possible to perform a selection of ILP values which is based on the examination of the content or domain assessed by coupled items. Hence, in order to be able to provide a preliminary evaluation of the performance of the heuristic, for each replication, after having identified the interval of ILP values, among the possible linking solutions we selected the one with the ILP value closest to the average ILP values in the interval.

Statistics Used to Evaluate the Performance of the Heuristic

The heuristic[6] provides a solution consisting of: (1) a set of item couples, and associated item difficulties, selected from Tests A and B; (2) an estimate of the true ILP obtained from the selected item couples; (3) a vector of item difficulties consisting of the combination of Tests A and B as reconnected using the ILP estimate. Further, the vector of item difficulties obtained reconnecting Tests A and B using the heuristic can be used in Winsteps to perform the anchored estimation of abilities of (simulated) examinees taking Test B, which are then placed in the same metric as Test A. Hence, the performance of the heuristic is evaluated concerning the following three areas: evaluation of accuracy and efficiency in estimating the true ILP; evaluation of the accuracy in recovering the true vector of item difficulties; evaluation of the accuracy in recovering the true vector of examinees' abilities.

The statistics used to evaluate the performance of the heuristic in each of these areas are presented in the Appendix (see "Statistics Used to Evaluate the Performance of the Heuristic"). To our knowledge in similar situations, no standard exists allowing a clear-cut decision in evaluating accuracy in estimating linking parameters. Hence, it is reasonable to expect that, on average, the error in recovering the true ILP should not exceed 0.25 logit (MAE and RMSE), and the heuristic should be as efficient as possible (Nash-Sutcliffe coefficient > 0), and that in the majority of situations the solution provided by the heuristic would be more accurate than that obtainable by using a naïve linking approach (i.e., assuming true ILP = 0). Concerning the recovering of true abilities and difficulties, we expect the association between obtained estimates and true values to be strong (e.g., $r \geq + 0.90$).

RESULTS

Accuracy and Efficiency of the Heuristic

Simulated data permitted to evaluate the accuracy and efficiency of the heuristic in estimating the true ILP. The performance of the heuristic with respect to the accuracy and efficiency of the ILP estimates is described in Tables 1 and 2.[7]

Regarding accuracy, the estimate of the true ILP[8], when computed across all conditions (N = 1,593), achieved values of mean absolute error (MAE) and root mean square error (RMSE) lower than .25 logit (MAE = 0.15; RMSE = 0.21). Results reported in Table 1 indicate that when varying the number of items included in the tests to be connected (true ILP ranging between approximately ± 1 logit), the ILP estimate appears to grow more and more precise as the number of items increases. When the test includes only 20 items, the MAE is equal to 0.24 logit (RMSE = 0.31), while at 30 item the MAE is as low as 0.15 logit (RMSE = 0.20). At 50 items, MAE is 0.10 logit (RMSE = 0.13).

TABLE 1
Evaluation of the performance of the heuristic: Accuracy and efficiency of ILP estimation across simulated conditions (all conditions; by number of items; by true ILP size).

	N	MAE	RMSE	E	R^2	g_0
All	1593	0.15	0.21	0.87	0.87	86.38
Number of items						
20	393	0.24	0.31	0.71	0.71	77.86
30	400	0.15	0.20	0.89	0.90	86.50
40	400	0.12	0.16	0.92	0.93	89.00
50	400	0.10	0.13	0.95	0.95	92.00
True ILP size						
Size A	400	0.11	0.16	–0.18	0.40	53.25
Size B	400	0.13	0.17	0.79	0.80	94.25
Size C	396	0.15	0.21	0.89	0.90	98.99
Size D	397	0.21	0.29	0.89	0.91	99.24

Note. True ILP size: A = 0 ≤ | ILP | ≤ 0.25; B = 0.25 ≤ | ILP | ≤ 0.50; C = 0.50 ≤ | ILP | ≤ 0.75; D = 0.75 ≤ | ILP | ≤ 1.00.
MAE = mean absolute error; RMSE = root mean square error; E = Nash-Sutcliffe coefficient; R^2 = coefficient of determination; g_0 = percentage of estimated ILP improving over naïve linking.

Table 1 also reports results concerning the performance of the heuristic when varying the true ILP (size) and results are computed pooling all of the item conditions (20, 30, 40, 50). Results here show that the error tends to increase as the true ILP increases in size. When the true ILP is closer to 0 (Size A condition: –0.25 <= ILP <= 0.25), the MAE is equal to 0.11 logit (RMSE = 0.16), while it increases to 0.21 logit (RMSE = 0.29) when the true ILP is closer to ± 1 (Size D condition).

With respect to efficiency of the estimates, Table 1 also reports relevant statistics, namely R^2, *E*, and g_0 (see Appendix for details). The overall efficiency of the proposed heuristic is quite high: R^2= .87, *E* = .87, and g_0 = 86%. When considering different conditions it is worthy to note that even when tests include 20 items only, efficiency statistics are adequate: *E* (with positive sign) and R^2 are both equal to 0.71, g_0 = 78%. The performance of the heuristic seems to get better as the number of items increases: at 50 items, both *E* and R^2 are as high as 0.95, and g_0 = 92%. With respect to true ILP size, when examining the efficiency of the heuristic at different size conditions, we observe that the heuristic performs better as the size of the true ILP increases (Size B: *E* = 0.79, R^2= 0.80, g_0 = 94%; Size C: *E* = 0.89, R^2= 0.90, g_0 = 99%; Size D: *E* = 0.89, R^2= 0.91, g_0 = 99%). A lower efficiency can be observed when the values of the true ILP are close to zero (Size A condition: *E* = –0.18; R^2= 0.40; g_0 = 53%), which is expected given the low variability of the parameters to be estimated.

Table 2 reports results for each of the conditions resulting from the combination of the two manipulated variables (i.e., number of items, true ILP size). With respect to accuracy, by looking at the table, it is easy to note that the combination between Size D and the 20-items condition results in the worst performance of the heuristic (MAE = 0.34 logit).

TABLE 2
Evaluation of the performance of the heuristic: Accuracy and efficiency of ILP estimation across simulated conditions (Number of items × True ILP size).

Number of items x True ILP size	*N*	MAE	RMSE	*E*	R^2	g_0
20×Size A	100	0.18	0.24	–1.64	0.24	37.00
20×Size B	100	0.22	0.27	0.51	0.55	82.00
20×Size C	96	0.22	0.29	0.79	0.80	96.88
20×Size D	97	0.34	0.43	0.77	0.78	96.91
30×Size A	100	0.11	0.14	0.04	0.45	50.00
30×Size B	100	0.12	0.15	0.85	0.85	97.00
30×Size C	100	0.16	0.21	0.89	0.91	99.00
30×Size D	100	0.20	0.26	0.91	0.93	100.00
40×Size A	100	0.09	0.12	0.32	0.48	58.00
40×Size B	100	0.09	0.12	0.91	0.91	98.00
40×Size C	100	0.13	0.16	0.93	0.95	100.00
40×Size D	100	0.17	0.21	0.94	0.96	100.00
50×Size A	100	0.07	0.09	0.61	0.70	68.00
50×Size B	100	0.09	0.11	0.92	0.92	100.00
50×Size C	100	0.09	0.12	0.96	0.98	100.00
50×Size D	100	0.14	0.18	0.96	0.94	100.00

Note. True ILP size: A = 0 ≤ | ILP | ≤ 0.25; B = 0.25 ≤ | ILP | ≤ 0.50; C = 0.50 ≤ | ILP | ≤ 0.75; D = 0.75 ≤ | ILP | ≤ 1.00.
MAE = mean absolute error; RMSE = root mean square error; *E* = Nash-Sutcliffe coefficient; R^2 = coefficient of determination; g_0 = percentage of estimated ILP improving over naïve linking.

Again, even at this size condition (Size D), the accuracy tends to increase as the number of items grows (30 items: MAE = 0.20 logit; 40 items: MAE = 0.17 logit; 50 items: MAE = 0.14 logit). A similar trend is observed at each of the remaining ILP size conditions. The efficiency statistics show a similar trend: the heuristic shows the lowest efficiency in the condition 20 items × Size A (E = –1.64, R^2 = 0.24, g_0 = 37%). Higher efficiency is achieved, even in the Size A condition, when the number of items increases: positive E values are observed when the number of items is 30 or higher.

A 4 × 4 ANOVA was run to examine the effects of the interaction of the two conditions manipulated in the simulation design (number of items and size) on absolute error (AE) and confirms that both manipulated variables have statistically significant effects on AE. The emerging effects are of opposite sign: the performance of the heuristic significantly improves when the number of test items is higher, F(3, 1577) = 91.85, p = < .001, η^2 = 0.15, and when the true ILP size is smaller, F(3, 1577) = 45.59, p < .001, η^2 = 0.08. Further, a small but significant interaction effect emerged between number of items and size, F(9, 1577) = 2.36, p = .01, η^2 = 0.01. In particular, significant contrasts emerged when comparing the 20-items condition with the other item conditions (i.e., 30, 40, 50) in the Size D group, indicating the 20-items condition as the one with the highest AE value. Figure 2 provides a visualization of this interaction effect.

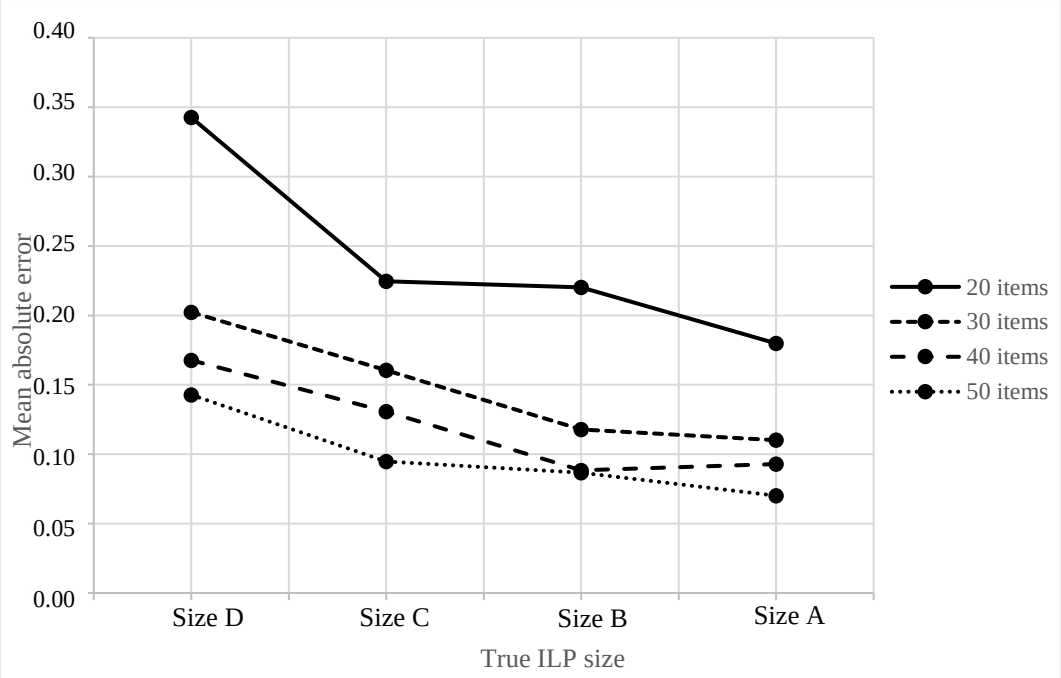

FIGURE 2.
Accuracy of intercept linking parameter (ILP) estimates by number of items and true ILP size.

Accuracy of Estimation of True Difficulty and Ability Parameters

Concerning the results of applying the ILP parameters estimated using the heuristic to recover the "true" difficulty and ability parameters used in the simulation runs, results are in line with what reported above concerning the accuracy of ILP parameters estimation (detailed results are shown in the Appendix, Tables A1 and A2). Overall, across all conditions, the average error in recovering difficulty parameters is quite low (MAE = 0.09 logit; RMSE = 0.10 logit), while the average correlation between generating values and recovered parameters generally approaches +1.00 (range = +.77; +1.00). As noted before for the ILP

parameters, errors in recovering difficulty parameters appear to decrease as test length increases, while errors tend to be positively related with the absolute size of "true" ILP parameters: MAE from 0.13 (at test of 20 items) to 0.06 (at test of 50 items); RMSE from 0.14 to 0.08; MAE from 0.07 (at Size A) to 0.12 (at Size D); RMSE from 0.08 to 0.13.

The accuracy of estimation of ability parameters showed a similar pattern. Overall, the average error in recovering ability estimates was generally low to moderate (range of MAE from 0.20 to 0.75; range of RMSE from 0.25 to 0.86), while the average correlation between generating values and recovered parameters was always quite high (range from .86 to .99). Again, accuracy of ability estimation appears to increase as test length increases and true ILP parameters decrease.

Discussion

In this study we presented, and put to the test using simulated data, a heuristic which can be used to connect two tests when typical linking strategies, such as common items, equivalent groups, or matched samples, are not viable. In the absence of a formal solution for the linking problem, the heuristic offers a tentative solution which requires, as the sole input, the item difficulties for the two tests as estimated separately using the Rasch model. The main idea behind this heuristic is that, based solely on this input, it may be possible to: (a) identify a set of possible solutions to the problem, each consisting of groups of items showing an identical pattern of difficulty across the two tests; (b) for each solution, to use the difficulties of these set of items to compute the ILP for the two tests; and (c) among these solutions, to select at least one as the most appropriate by exploiting features of similarity or dissimilarity between the solution and known data. More specifically, we propose that the most promising solution can been obtained by comparing the range of difficulty of the two tests after they are reconnected using each identified solution with the range of difficulty of the two tests before the reconnection is performed.

The results of analyses on simulated data indicate the performance of the heuristic is satisfactory. The average error is generally quite low (MAE <= 0.2 logit), and in 86% of the cases the solution to the linking problem provided by the heuristic is more precise than that obtainable using a naïve approach, which assumes ILP = 0. The performance of the heuristic is particularly promising when tests include 30 items or more: in 90% of those cases, the linking solution provided by the heuristic is preferable to the naïve linking approach, and the average absolute error in estimating the ILP is as low as 0.12. This is particularly relevant given that test length of 30-40 items has been indicated as a typical lower bound for educational tests, in particular when performing test equating (Kolen & Brennan, 2004; Wells, Subkoviak, & Serlin, 2002).

Of course, there exist also cases in which the heuristic indicates solutions to the linking problems which significantly deviate from the true values. Findings from the exploratory analyses performed on simulate data indicate error is higher for shorter tests including just 20 items. Error also seems to increase as the true ILP increases, but MAE never exceeds 0.2 logits.

In the absence of formal methods for the solution of the linking problem, however, it is perhaps necessary to point out that the relevant question is not to wonder why sometimes the proposed heuristic is wrong, but whether (1) the heuristic is useful in dealing with the linking problem and (2) if the heuristic can be further improved to solve this problem. Concerning the first question, in light of results emerging from the simulation study, the answer appears to be generally positive, and suggests that the present heuristic may represent a useful tool to recover information about the relative distance in difficulty of tests for which traditional information (e.g., common items and persons, collateral information allowing sample matching) required for the linking of the tests is not available. In these situations, the heuristic may represent a more controlled, and less arbitrary, alternative to the use of subjective evaluations of the difficulty of

the tests. Further, it generally provides a significant improvement over naive assumptions of equality of the tests (i.e., arbitrarily assuming the two tests as having an ILP value of zero). However, it is important to note that the simulation design presented in this study was carried out to provide a preliminary test of the functioning of the heuristic and cannot be considered fully exhaustive, as the conditions allowed to vary (i.e., number of items per test, and size of the true ILP) do not cover the full set of possible situations which can be observed in empirical settings. Moreover, fit between response data and the measurement model was generated to be always optimal and comply with the assumptions of the Rasch model (e.g., unidimensionality of measurement, items distributing uniformly across the difficulty continuum), a condition may not always apply to empirical situations. For these reasons, the results presented in this study are best interpreted as a first step in the evaluation the performance of the heuristics.

Concerning the possibility of improving the heuristic, the findings presented in this study should be necessarily seen as a starting point, rather than a point of arrival. Further developments are required, and mostly concern the need to provide a formal description of the approach used in the simulation design to estimate the true ILP, consisting in computing the average of the ILP values obtained in the set of possible solutions (see Step 3, in the Section "A Heuristic Approach to Link Tests when no Linking Data is Available"). The deepening of this aspect, including the examination of possible alternative approaches, could then form a first direction of development of the proposed heuristic.

A detailed assessment of those cases in which the solution provided by the heuristic is far from the true ILP value would also be beneficial for the development of the heuristic. Based on results from the simulation study, the performance of the heuristic seems to become less precise as the number of items gets smaller, and as the true ILP value increases. A more detailed examination in this direction could perhaps help in recognizing a priori conditions which negatively influence the performance of the heuristic, and thus provide information useful to improve its functioning in these situations, as well as defining its applicability.

Finally, it is possible to imagine possible variations of the presented heuristic. A first possible variation relates to the fact that the input for the heuristic consists in the difficulties of the items in the two tests to be connected; but such estimates carry other information that may be useful, such as their associated standard error of the estimate (SEM) and/or the fit statistics (e.g., mean square INFIT and OUTFIT). Such information could be integrated at the beginning of the heuristic process with the aim of removing certain items (e.g., those with a larger SEM and/or with less adequate fit statistics), so that the definition of all possible ILPs (see Step 1, in the Section "A Heuristic Approach to Link Tests when no Linking Data is Available") is carried out starting from a subset of items showing the most promising measurement properties.

In turn, relaxing the assumption of perfect linearity of difficulties in the item couples, that is, allowing the slope linking coefficient to slightly deviate from 1, would allow the heuristic to return a larger set of possible solutions, each still characterized by a low linking error value (as computed by Monseur & Berezner, 2007). This will also help improve the potential applicability of the heuristic to situations in which the discrimination of the tests which need to be equated varies across different groups of examinees, such as is often the case when data is collected on examinees attending different grades (e.g., vertical linking; Humphry, 2010).

Another area in which the introduction of variations might help improve the performance of the heuristic relates to the statistic which is currently used in the selection of the solutions among all those available (see Step 3, Section "A Heuristic Approach to Link Tests when no Linking Data is Available"), that is, the range of item difficulties in the two tests. Of course, it is also plausible that other statistics may prove useful in improving the functioning of the heuristic.

The aforementioned possible improvements of both the simulation design, and the heuristic itself, are relatively easy to implement, and may be the focus of future studies for which the present study aimed to provide a common framework.

NOTES

1. A rounding of item difficulty values in vectors **X** and **Y** is required for this step. Since these vectors of difficulties are the only input information of the heuristic, the chosen rounding precision also applies to the estimate of the ILP provided by the heuristic.
2. A consequence of the requirement of perfect linearity imposed by the heuristic to the difficulties of item couples selected from vectors **X** and **Y** is that for all the linking solutions provided by the heuristic, computed linking error always equals zero (Monseur & Berezner, 2007). For this reason, this formulation of linking error cannot be used to evaluate the performance of the heuristic.
3. Of course, it can happen that no ILP can be obtained using the heuristic, especially if one or both tests are made up of a few items (e.g., less than 20) or if the difficulties of the items in one test or the other are not adequately well distributed on the continuum. In this case, it only remains to note the impossibility to continue; vice versa, you can proceed with the next step.
4. The range statistic used in this study is based on the maximum product of spacing (MPS) method of estimation of unknown end-points of a uniform distribution (Cheng & Amin, 1983). More specifically, the range statistic is obtained as:

$$\widehat{RAN} = \frac{n \cdot [\text{MAX}(\mathbf{V}) - \text{MIN}(\mathbf{V})] + \text{MAX}(\mathbf{V}) - \text{MIN}(\mathbf{V})}{n - 1}$$

   Where:

$$\boldsymbol{V} = \text{vector having the largest range between } X \text{ and } Y;$$
$$n = \text{order of vector } \mathbf{V}.$$

5. Concerning the size variable, because of our decision to round values to the second decimal number, we can obtain 50 different positive and negative true ILP values for each size group, each of them used twice in the analyses, resulting in 100 replications for each size group. True ILP values are not directly generated, but instead estimated by calibrating item response data using the Rasch model. More specifically, a matrix including responses from all generated person on all generated items (Test A + Test B) is created, and calibrated to the Rasch model. True ILP values are obtained as the mean difference in difficulty between the two tests (i.e., the difference in mean difficulty between Test B and Test A). Assignment of items to each test (A, B) which is performed with a maximum of 600 iteration in order to obtain a true ILP parameters as close as possible to the different desired value within each size.
6. In an empirical situation, as the true ILP is unknown, it would be convenient if the heuristic could provide more than one possible solution, the best choice then being determined by human evaluators based on the content of the items. In case of simulated test data, as in the present study, the "true" solution is known, while information on the content of test items is not available by design. For this reason, in the present study, for each simulated case only one solution is selected for the evaluation of the performance of the heuristic.
7. In 7 of the total 1,600 tested simulated cases (0.43%), the heuristic was not able to provide a solution. These 7 cases consisted of simulated cases (all 7 cases were in the 20-item condition; 4 in the Size C, and 3 in Size D conditions) in which solutions were not found in the range established. For this reason, the statistics reported in the text refer to the remaining 1,593 cases.
8. Overall (N = 1,593), the true ILP varies between $-1.08 \leq \text{ILP} \leq +1.08$; $\overline{\text{ILP}} = 0.00$; $SD_{\text{ILP}} = 0.58$. The estimated ILP varies between $-1.50 \leq \widehat{\text{ILP}} \leq +1.16$; $\overline{\widehat{\text{ILP}}} = 0.00$; $SD_{\widehat{\text{ILP}}} = 0.00$. The ILP estimates were obtained using a number of item couples varying between 3 and 12.

REFERENCES

Angoff, W. H. (1971). *Educational measurement*. Washington, DC: American Council on Education.

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement*, *21*(2), 175-189. doi:10.1111/j.1745-3984.1984.tb00228.x

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*(2), 131-155. doi:10.1111/j.1745-3984.2007.00031.x

Cheng, R. C. H., & Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society. Series B (Methodological)*, *45*(3), 394-403. doi:10.1111/j.2517-6161.1983.tb01268.x

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, *41*, 43-68. doi:10.1111/j.1745-3984.2004.tb01158.x

Fisher, W. P. (1997). Physical disability construct convergence across instruments: Towards a universal metric. *Journal of Outcome Measurement*, *1*(2), 87-113.

Humphry, S. M. (2010) Modeling the effects of person group factors on discrimination. *Educational and Psychological Measurement*, *70*(2), 215-231. doi:10.1177/0013164409344553

Hsu, T. C., Wu, K. L., Yu, J. Y. W., & Lee, M. Y. (2002). Exploring the feasibility of collateral information test equating. *International Journal of Testing*, *2*(1), 1-14. doi:10.1207/S15327574IJT0201_1

Kaspersen, E., Pepin, B., & Sikko, S. A. (2017). Measuring student teachers' practices and beliefs about teaching mathematics using the Rasch model. *International Journal of Research & Method in Education*, *40*(4), 421-442. doi:10.1080/1743727X.2016.1152468

Kim, S., Livingston, S. A., & Lewis, C. (2011). Collateral information for equating in small samples: A preliminary investigation. *Applied Measurement in Education*, *24*(4), 302-323. doi:10.1080/08957347.2011.607057

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.

Linacre, J. M. (2009). Winsteps (Version 3.68. 0)[Computer software]. Chicago, IL: Winsteps.com.

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

Marengo, D., Miceli, R., Rosato, R., & Settanni, M. (2018). Placing multiple tests on a common scale using a post-test anchor design: Effects of item position and order on the stability of parameter estimates. *Frontiers in Applied Mathematics and Statistics*, *4*, 50. doi:10.3389/fams.2018.00050

Marengo, D., Miceli, R., & Settanni, M. (2016). Do mixed item formats threaten test unidimensionality? Results from a standardized math achievement test. *TPM – Testing, Psychometrics, Methodology in Applied Psychology*, *23*(1), 25-36. doi:10.4473/TPM23.1.2

Masters, G. N. (1985). Common-person equating with the Rasch model. *Applied Psychological Measurement*, *9*(1), 73-82. doi:10.1177/014662168500900107

Miceli, R., Marengo, D., Molinengo, G., & Settanni, M. (2015). Emerging problems and IRT-based operational solutions in large-scale programs of student assessment: The Italian case. *TPM – Testing, Psychometrics, Methodology in Applied Psychology*, *22*(1), 53-70. doi:10.4473/TPM22.1.5

Miceli, R., & Molinengo, G. (2005). Somministrazione di test computerizzati di tipo adattivo. Un'applicazione del modello di misurazione di Rasch [Administration of computerized and adaptive tests: An application of the Rasch model]. *Testing, Psychometrics, Methodology in Applied Psychology*, *12*(3), 131-149.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1992). How to equate tests with little or no data. *ETS Research Report Series*, *1*. doi:10.1002/j.2333-8504.1992.tb01451.x

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*(3), 323-335.

Nagy, P. (2000). The three roles of assessment: Gatekeeping, accountability, and instructional diagnosis. *Canadian Journal of Education/Revue canadienne de l'education*, *25*(4), 262-279. doi:10.2307/1585850

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I – A discussion of principles. *Journal of Hydrology*, *10*(3), 282-290. doi:10.1016/0022-1694(70)90255-6

Powers, S. J., & Kolen, M. J. (2012). Using matched samples equating methods to improve equating accuracy. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 2, pp. 87-114). Iowa City, IA: CASMA, The University of Iowa.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*(1), 77-87. doi:10.1177/0146621602261005

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (ETS Research Report-93-04). Princeton, NJ: Educational Testing Service.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement*. Chicago, IL: University of Chicago, MESA Press.

APPENDIX

Statistics Used to Evaluate the Performance of the Heuristic

Accuracy of Estimation of True ILP

For each simulated case, the absolute error in estimating the ILP by taking the absolute difference between the estimated ILP ($\widehat{\Theta}$) and the true ILP value (Θ). That is, the following formula was used:

$$\text{Absolute error: } \mathrm{AE} = \left|\Theta - \widehat{\Theta}\right|$$

Next, we computed the average absolute error in overall condition, as well as in each condition of the simulation. That is, for *N* simulated cases, we computed the average absolute difference between the estimated ILP ($\widehat{\Theta}$) and the true ILP value (Θ):

$$\text{Mean absolute error: } \mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\Theta_i - \widehat{\Theta_i}\right|$$

Further, in order to obtain an indicator of accuracy which is sensitive to large errors, we computed the root mean square error:

$$\text{Root mean square error: } \mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\Theta_i - \widehat{\Theta_i}\right)^2}$$

Efficiency in Estimating the True ILP

The efficiency of the heuristic in estimating the true ILP was investigated using the coefficient of determination (R^2) and the *E* coefficient (Nash-Sutcliffe coefficient; Nash & Sutcliffe, 1970).

In more detail, for *N* simulated cases, R^2 was computed by estimating a simple regression model in which the vector of true ILPs ($\boldsymbol{\Theta}$) served as dependent variable, while the vector of ILP ($\widehat{\boldsymbol{\Theta}}$) was included in the model as independent variable. The *E* coefficient was computed using the following formula:

$$\text{Nash-Sutcliffe coefficient: } E = 1 - \frac{\sum_{i=1}^{N}\left(\boldsymbol{\Theta}_i - \widehat{\boldsymbol{\Theta}_i}\right)^2}{\sum_{i=1}^{N}\left(\boldsymbol{\Theta}_i - \bar{\boldsymbol{\Theta}}\right)^2};$$

When computed on the same data, the value of the E coefficient is generally close to R^2 for positive values and as values approach 1. However, in contrast to the R^2, the *E* cofficient is expected to be negative when the ILP estimates provided by the heuristic are less accurate than the average true ILP values, that is when the heuristic is not efficient.

Finally, we computed g_0 statistic, which is the percentage of simulated cases in which the use of the heuristic is linked to a gain in accuracy in estimating the true ILP over the use of a naïve linking approach (i.e., assuming true ILP is 0). That is, for *N* simulated cases, we computed the percentage of cases in which the AE statistic is lower than the absolute difference between the true ILP and 0, using the following formula.

$$g_0 = \left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{A}_i\right)\cdot 100; \text{where: } \mathrm{A}_i = 1 \text{ if } \mathrm{AE}_i < |\Theta_i|$$

Accuracy of Estimation of True Difficulty and Ability Values

For *N* simulated cases, the accuracy of heuristic in estimating the true difficulty (*W*) and ability (*Z*) values was computed using the following statistics.

Mean absolute error:

$$\text{MAE}(W) = \frac{1}{L}\sum_{k=1}^{L}|w_k - \widehat{w_k}|;\ \text{MAE}(Z) = \frac{1}{P}\sum_{k=1}^{L}|z_k - \widehat{z_k}|;$$

$$\text{Root mean square error: RMSE}(W) = \sqrt{\frac{1}{L}\sum_{k=1}^{L}(w_k - \widehat{w_k})^2};\ \text{RMSE}(Z) = \sqrt{\frac{1}{P}\sum_{k=1}^{P}(z_k - \widehat{z_k})^2};$$

where L and P respectively indicate the order of the vectors of estimated difficulty ($\widehat{W}$) and ability ($\hat{Z}$) values after the reconnection of vectors **X** and **Y**.

Finally, strength of associations between estimated ILP and true ILP values was investigated using Bravais-Pearson's correlation coefficient r ($r_{W;\widehat{W}}$; $r_{Z;\hat{Z}}$).

Accuracy in Recovering True Item Difficulty and Person Ability Values

TABLE A1
Evaluation of the performance of the heuristic:
Accuracy of recovering true item difficulty values across simulated conditions

	N	Mean absolute error			Root mean square error			*r* (Pearson)		
		(Min)	(MAE)	(Max)	(Min)	RMSE	(Max)	(Min)	(Mean)	(Max)
All	1,593	0.03	0.09	0.73	0.04	0.10	0.73	0.77	1.00	1.00
Number of items										
20	393	0.04	0.13	0.73	0.05	0.14	0.73	0.77	0.99	1.00
30	400	0.03	0.09	0.37	0.04	0.10	0.37	0.96	1.00	1.00
40	400	0.03	0.07	0.37	0.04	0.09	0.37	0.96	1.00	1.00
50	400	0.03	0.06	0.23	0.04	0.08	0.24	0.98	1.00	1.00
True ILP size										
Size A	400	0.03	0.07	0.33	0.04	0.08	0.34	0.95	1.00	1.00
Size B	400	0.03	0.08	0.37	0.04	0.09	0.37	0.92	1.00	1.00
Size C	396	0.03	0.09	0.47	0.04	0.10	0.48	0.93	1.00	1.00
Size D	397	0.04	0.12	0.73	0.04	0.13	0.73	0.77	0.99	1.00
Number of items× True ILP size										
20×Size A	100	0.04	0.10	0.33	0.05	0.12	0.34	0.95	0.99	1.00
20×Size B	100	0.04	0.12	0.37	0.05	0.13	0.37	0.92	0.99	1.00
20×Size C	96	0.04	0.12	0.47	0.06	0.14	0.48	0.93	0.99	1.00
20×Size D	97	0.04	0.18	0.73	0.05	0.19	0.73	0.77	0.98	1.00
30×Size A	100	0.03	0.07	0.18	0.04	0.08	0.19	0.99	1.00	1.00
30×Size B	100	0.04	0.08	0.21	0.04	0.09	0.21	0.99	1.00	1.00
30×Size C	100	0.04	0.09	0.26	0.05	0.10	0.26	0.97	0.99	1.00
30×Size D	100	0.04	0.11	0.37	0.05	0.13	0.37	0.96	1.00	1.00
40×Size A	100	0.03	0.06	0.22	0.04	0.07	0.22	0.98	1.00	1.00
40×Size B	100	0.03	0.06	0.25	0.04	0.07	0.25	0.96	1.00	1.00
40×Size C	100	0.03	0.08	0.21	0.04	0.09	0.22	0.98	1.00	1.00
40×Size D	100	0.04	0.10	0.37	0.04	0.11	0.37	0.96	1.00	1.00
50×Size A	100	0.03	0.05	0.14	0.04	0.06	0.14	0.99	1.00	1.00
50×Size B	100	0.03	0.06	0.15	0.04	0.07	0.16	0.99	1.00	1.00
50×Size C	100	0.04	0.06	0.21	0.04	0.07	0.22	0.98	1.00	1.00
50×Size D	100	0.04	0.08	0.23	0.05	0.10	0.24	0.99	1.00	1.00

Note. True ILP size: A = 0 ≤ | ILP | ≤ 0.25; B = 0.25 ≤ | ILP | ≤ 0.50; C = 0.50 ≤ | ILP | ≤ 0.75; D = 0.75 ≤ | ILP | ≤ 1.00.

TABLE A2
Evaluation of the performance of the heuristic:
Accuracy in recovering true ability values across simulated conditions

	N	Mean absolute error			Root mean square error			*r* (Pearson)		
		(Min)	(MAE)	(Max)	(Min)	RMSE	(Max)	(Min)	(Mean)	(Max)
All	1,593	0.20	0.28	0.75	0.25	0.37	0.86	0.86	0.97	0.99
Number of items										
20	393	0.32	0.36	0.75	0.41	0.46	0.86	0.93	0.94	0.95
30	400	0.26	0.29	0.43	0.33	0.38	0.53	0.93	0.97	0.98
40	400	0.22	0.25	0.42	0.28	0.33	0.51	0.96	0.98	0.98
50	400	0.20	0.23	0.31	0.25	0.29	0.40	0.96	0.98	0.99
True ILP size										
Size A	400	0.20	0.28	0.44	0.26	0.35	0.54	0.91	0.97	0.99
Size B	400	0.20	0.28	0.46	0.25	0.36	0.57	0.93	0.97	0.99
Size C	396	0.20	0.28	0.54	0.25	0.36	0.66	0.92	0.97	0.99
Size D	397	0.20	0.30	0.75	0.26	0.38	0.86	0.86	0.97	0.99
Number of items × True ILP size										
20×Size A	100	0.33	0.35	0.44	0.42	0.45	0.54	0.91	0.95	0.97
20×Size B	100	0.32	0.36	0.46	0.41	0.46	0.57	0.93	0.95	0.97
20×Size C	96	0.33	0.36	0.54	0.42	0.46	0.66	0.92	0.95	0.97
20×Size D	97	0.32	0.39	0.75	0.41	0.49	0.86	0.86	0.94	0.97
30×Size A	100	0.26	0.28	0.42	0.33	0.36	0.40	0.96	0.97	0.98
30×Size B	100	0.26	0.29	0.33	0.34	0.37	0.41	0.95	0.97	0.98
30×Size C	100	0.27	0.30	0.35	0.34	0.38	0.45	0.94	0.98	0.98
30×Size D	100	0.27	0.31	0.43	0.34	0.31	0.43	0.94	0.97	0.98
40×Size A	100	0.23	0.25	0.29	0.29	0.32	0.36	0.96	0.98	0.98
40×Size B	100	0.22	0.25	0.33	0.28	0.32	0.41	0.97	0.98	0.98
40×Size C	100	0.22	0.25	0.31	0.29	0.33	0.39	0.96	0.97	0.98
40×Size D	100	0.23	0.27	0.42	0.29	0.34	0.51	0.96	0.97	0.98
50×Size A	100	0.20	0.22	0.25	0.26	0.29	0.33	0.97	0.98	0.99
50×Size B	100	0.20	0.22	0.25	0.25	0.29	0.33	0.97	0.98	0.99
50×Size C	100	0.20	0.22	0.26	0.25	0.29	0.38	0.97	0.98	0.99
50×Size D	100	0.20	0.24	0.31	0.26	0.30	0.40	0.96	0.98	0.99

Note. True ILP size: A = 0 ≤ |ILP| ≤ 0.25; B = 0.25 ≤ | ILP | ≤ 0.50; C = 0.50 ≤ | ILP | ≤ 0.75; D = 0.75 ≤ | ILP | ≤ 1.00.