Equate groups: An innovative method to link multi-item instruments across studies

Iris Eekhout (iris.eekhout@tno.nl)
Netherlands Organization for Applied Scientific Research TNO
Ann M. Weber
University of Nevada Reno
Stef Buuren
Netherlands Organization for Applied Scientific Research TNO

Research Article

Keywords: Rasch Model, Item Response Theory, Equating, Concurrent Calibration, Meta-analyses

Posted Date: March 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1439153/v1

Abstract

Background: Combining existing data sets for meta-analyses enables analyses across multiple contexts and conditions. However, the use of different measurement instruments in different studies hampers comparability. This paper proposes a novel method to link tools administered across multiple studies.

Methods: The proposed methodology starts by forming equate groups, i.e., sets of items from different instruments with similar meanings. We fit the Rasch model to all data, with the additional restriction that items within the same equate group obtain identical difficulty estimates. The main modelling task is to divide equate groups into active (where the restriction holds) and passive (without the restriction) groups. We studied the method's performance in a simulation study and illustrated its practical application to early childhood development.

Results: If we treat all equate groups as passive, the difficulty estimates of the identical items are unduly affected by differences in abilities between the study samples. On the other hand, when abilities are similar across samples, it is safer to set all equate groups to passive because any mis-specified equate groups can lead the restricted model to underperform. Our method deals with the common case between these two extremes where 1) we suspect a priori ability differences between samples, and 2) at least some of the items are equivalent across instruments.

Conclusions: Our equating method presents a flexible alternative to the classic common-item nonequivalent-group design for existing data. We conclude that equate groups are an economical and exciting concept that enables insightful statistical analysis from seemingly disparate data sources.

Background

Aggregating data from multiple studies into a single data set to conduct meta-analyses of person data in an integrative manner has become a common practice that enables researchers and funding organizations to synthesize findings across multiple contexts and conditions [1, 2]. The use of existing data can yield valuable information without the financial and time investment required to gather new data. However, studies collect patient-reported outcome data by different multi-item instruments validated in other languages, cultures and contexts, even when measuring the same underlying construct. Examples from medical research include health-related quality-of-life instruments and depression scales [3]. The challenge is to integrate the empirical data to express the study outcomes on the same scale. This paper describes a solution to create a common scale for a common construct using data obtained from different samples assessed with different instruments that share common items.

A measurement model specifies the relation between measured items and a latent construct. The Rasch model is a fundamental measurement model that specifies the probability of passing an item as a function of the difficulty of the item and the person's ability [4]. The model is of great practical use for achieving comparability. Suppose that our data consist of item responses collected on different instruments measuring the same construct in *one sample* of persons. In that case, a Rasch analysis can obtain both difficulty and ability estimates for all items and all persons on a common scale. Suppose now that we create two new instruments from the items that fit the model. Instrument A contains only easy items, and instrument B consists of difficult items. We administer A to a sample of low ability (e.g. younger children) and B to a more mature sample. Even though both instruments and samples differ, we are still able to estimate the ability of each person in both samples on the same scale. However, without the set of difficulty estimates from the first sample, the scales of A and B would be unconnected, and we cannot get scores on a common scale. The problem is that there is no overlap between either the instruments or the samples. Instruments or samples need to be linked in some way. The Rasch model is a way to create linkages.

In this paper, we will focus on a situation where different instruments administered in the same and different settings share common items that are not strictly equivalent, requiring innovative extensions to existing methods to link them. When common items across different measurement instruments are identical, one may implement the common-item nonequivalent groups design or the nonequivalent anchor test (NEAT) [5, 6]. Equating this kind of data is also known as vertical equating or vertical scaling [7]. Test equating is a statistical method to convert scores obtained from different instruments to the same scale. If successful, we may compare scores from one instrument to scores obtained by another. Various equating methods exist, for example, transformation methods and fixed parameter calibration [5].

Transformation methods fit a separate Rasch model to each instrument and adjust either the person parameters (i.e. ability) or the item parameters to align the scales. Transformation methods applied to person parameters are well developed and work best for samples with similar abilities. Examples of such methods include simple equipercentile equating and more advanced chain equating, and post-stratification equating [5, 6, 8]. Transformation methods for item parameters are less well developed. In a Rasch model, the parameters from one instrument can be rescaled onto the scale of the second instrument by linear transformations as $\delta_{Bi} = Z\delta_{Ai} + V$, where δ_{Bi} is the difficulty of item i in instrument B and δ_{Ai} is the difficulty of item in instrument A. Z and V are constants [9]. Specification of reasonable values for Z and V is still an open problem.

Fixed parameter calibration is a linking method focused on the item parameters, also known as pre-equating [10]. In fixed parameter calibration, different instruments are linked by fixing the item difficulties to known calibration values for a common set of anchor items. This method assumes that the remaining items fit the same model, and it requires prior information.

Both transformation methods and fixed parameter calibration require separate parameter estimation procedures for each instrument, which are cumbersome to apply when different studies administered different instruments. Concurrent calibration is an alternative that attempts to estimate the item parameters for all instruments simultaneously. Concurrent calibration works from the merged data from all study samples by stacking common items from different instruments into the same column to form a "common item link". Item x from instrument A forms a common item link with the same item x from instrument B. We place the respondents' responses on item x from instrument A and B into the same column for model estimation. The combined data typically have many missing values because some items were administered to respondents from one study only and some only to the other. We may then use the resulting calibrations for comparing study samples without the need for post-processing or transformations [11]. Several authors have warned that concurrent calibration with stacking may not work well with items that cover a wide range of difficulties and when the number of common items is small [5, 12]. Others pointed out that it might be better to have fewer, tightly linked common items that are well-distributed across the range of difficulties [13].

Concurrent calibration with stacking is an attractive option, but its application requires making a strict distinction between items that are indeed the same in different instruments and items that differ. In reality, when instruments have been developed, adapted and translated for different samples and contexts, this dichotomy becomes somewhat blurry. Items measuring the same skill but from different instruments may have been adapted to suit the local context, language and format (e.g. the number of response options). Such changes may or may not affect the measurement properties of the item. Moreover, if we administer two instruments with a common item to a single sample, we have no place to store both responses in the merged observation-item matrix. A solution to these issues is to equate the relevant parameters of the measurement model, for example, as previously proposed in the context of multiple correspondence analysis [14]. For a Rasch model, an equality constraint restricts difficulty estimates to be identical, thus forming common item links between different instruments without putting the common items into the same column. Equating is considerably more flexible than data merging, and the process of finding an adequate set of equate items is a distinct modelling task. Once we have selected a suitable collection, applying equality constraints in a Rasch model enables mapping and linking multiple samples and instruments onto one common scale.

This article explores a method to link different instruments of the same construct administered in sometimes the same, but mostly different study samples. The next section describes the procedure in more detail. We use a small simulation study to investigate the performance of the Rasch model under equating and illustrate the new method in an application to existing measurement data of early child development across multiple studies and contexts.

Methods

The methodology to generate a common scale that links data from different samples across different multi-item instruments starts with collecting item-level data from multiple studies on the same construct. We then need to synchronize variable names and scoring formats across studies. After harmonization, we locate connections between study samples and measurement instruments. Two samples connect when they use the same tool. Samples could also connect indirectly via a third sample. Connections via the same instruments are natural links because identical items occur in both samples. In the absence of natural links, subject-matter experts identify potential bridges across studies by placing items into groups based on similarity. We use the term “equate group” to refer to a group of items that measure the same feature in (perhaps slightly) different ways.

Figure 1 displays items from three different instruments that measure child development. The tools contain several common items that are measured in multiple instruments but also unique items. Common items are part of an equate group, as displayed by the arrows between them. In the example, there is one (common) item that is equivalent in all three instruments (i.e. walks alone). The item "sits without support" occurs in both the first (i.e. blue) and the second (i.e. green) instrument, and the item "claps hands together" appears in the second (i.e. green) and third (i.e. orange instrument). When we administer each instrument in a different study, we can link these by placing common items in the same column and estimate the model with concurrent calibration. However, linking through data reorganization is not always possible. For example, in Fig. 1, the first and second instrument is administered in the same study. We cannot place the two responses on the early two instruments into the same column. In situations like these, the equate group method provides a more flexible way to link items used in the model. Eekhout & van Buuren use this generic strategy to connect 16 studies measuring child development [15].

Both statistical information and subject matter experts' input provide the basis of the assignment of items to eligible equate groups. An equate group can be either passive or active. An active equate group links items across instruments by restricting item difficulty estimates to be identical. A passive equate group does not enforce this restriction. Equate group status is initially unknown and should be determined as part of the modelling effort. This flexibility in testing and modifying the equate groups is an excellent strategy for improving comparability. In general, high quality equate groups contain items that function similarly in different tests. We evaluate similarity based on face validity and statistical measures, like infit and outfit. In this way, we have the flexibility to assess the goodness of an equate group for items that are not quite identical, such as "says three words" and "says five words" in Fig. 1. Since we store those items in separate columns, it is possible to test whether they can be equated or not without re-organizing the data.

The Rasch model models the probability of passing an item as a logistic function of the difference between each person's ability and the difficulty of the item as follows:

$$P_{ni} = \frac{\exp\left(\beta_n - \delta_i\right)}{1 + \exp\left(\beta_n - \delta_i\right)}$$

where P_{ni} is the probability of passing item i , δ_i is the item calibration dependent on the attributes of item , i.e. the difficulty of item i and β_n depends on the attribute of person n, i.e. the ability of the person. [16] The log-odds that a person with ability β_n answers an item with difficulty δ_i correctly is the difference between the person's ability and the item's difficulty $\left(\beta_n - \delta_i\right)$.

We extended the Rasch model to facilitate the use of equate groups. This extension requires that item difficulties of similar items should be identical. Wright and Stone present a simple method to equivalate the difficulty between two test forms with common item links [16]. In their case, this is done by estimating the shift in difficulty as the weighted average of difficulty differences of the linking items and using this weighted average to align the difficulties of the test forms. Wright and Stone align the scales after fitting separate Rasch models to each instrument, similar to the transformation methods described previously. We adopt their method for concurrent estimation, leading to a constrained Rasch model. In particular, suppose that δ_i is the difficulty of each item in the equate group, l is the number of items in the equate group and w_i is the number of respondents with an observed score on item i. We apply the constraint δ_q = $\delta_i$for all *i* during concurrent model estimating, where

$$\delta_q = \frac{\sum_i^l \delta_i w_i}{\sum_i^l w_i}$$

is the combined difficulty of the items in the equate group.

The modelling task consists of selecting the items that best fit the Rasch model and activating equate groups that bridge all instruments. We recommend to exclude items with less than ten observations in the smallest category. As a modelling strategy, we advise an iterative procedure that balances between preserving the best items, using active equate groups that perform well and bridge all instruments and reasonable distributions of the latent scores in the study population. We diagnose the fit of the model by the quality of equate groups through fit measures and visualizations. We also evaluate the plausibility of the latent variable's distribution and the infit of the items. An important assumption underlying equate groups is that the items in the group work in the same way across the different study samples, i.e., there is no differential item functioning. This assumption is critical for active equate groups because when it is not met, restricting the difficulty parameters to be equal across studies may introduce unwanted bias in comparisons between study samples [17, 18].

To estimate the constrained Rasch model, we developed new software in an R package called dmetric [19]. This package contains various tools to work with equate groups (see Appendix A for code). The rasch() function in dmetric package extends the rasch.pairwise.itemcluster() function from the sirt package [20, 21]. The dmetric package also includes functions that calculate fit measures for items and equate groups and that visualise item response curves. At the time of writing, dmetric is not yet available on CRAN. For the time being, please contact the authors for access to the package.

Simulation

Objective

Previous simulation studies have investigated the performance of concurrent calibration methods compared to transformation methods, for example, or the performance of concurrent calibration across different ability distributions [12, 22–24]. Here, we conducted a simulation study to investigate the quality of the constrained solution using equate groups as a function of the measurement range of the instruments, the number of equate groups, the location of equates along the scale and the difference in abilities between samples under various amounts of misspecification.

Simulation design

Data were simulated for two or three instruments, and each instrument consisted of 10 unique items plus additional items in equate groups. Table 1 presents a summary of the simulation design. The item difficulties ($\delta_{i..I}$) varied in three simulation conditions. Difficulties of items could overlap, i.e. be in the same range for both instruments, or could not overlap, where only the items in the equate groups connected the instruments. Where the difficulties did not overlap, the difficulties of the items could be close to one another or not. Accordingly, the range of item difficulties of the instruments (1) did not overlap and were not close: [-5,-3] and [3, 5]; (2) did not overlap but were close: [-3,-0.1] and [0.1,-3]; or (3) overlapped: [-2,1] and [-1,2]. The number of equate groups was 1, 2 or 5. Depending on the number of equate groups, we added items to the second instrument with item difficulties that were present in the first instrument. The sensitivity to the specification of "wrong" equate groups was investigated by gradually increasing the difference between difficulties of items in one of the equate groups, starting at 0 (no deviation) to 2 logits, with steps of 0.1 logits.

We varied the locations of the equate groups. We suspected that the best locations for equates would be relatively far apart and would cover a wide range of the scale. Equate groups were placed in the centre of the instruments, in the range of one instrument but not in the other, spread equally over both instruments, or at the end of one instrument. Furthermore, we varied two conditions for the person abilities. Both samples had the same ability in the first condition: normally distributed with N(0 ~ 1). In the second condition both samples had different abilities: normally distributed with a mean difference of 2, so N(-1 ~ 1) and N(1 ~ 1). In the condition with three instruments, we simulated samples that were sensitive at distinct ranges, at N(-1.5 ~ 1), N(0.5 ~ 1) and N(2.5 ~ 1).

Given the set of items and difficulty parameters, we simulated a data set with 1000 rows per instrument and one column for each item in the instruments. The person ability settings and the item difficulty settings were used in the sim.raschtype function of the sirt package to generate the data (see R code in Appendix A) [20]. A Rasch model was fitted on the full data to obtain the true difficulty parameters for the reference situation where all items were administered to all respondents. Subsequently, the data were split such that 1000 persons had data for the first instrument, another 1000 for the second and, if the condition required, another 1000 for the third instrument. Two additional Rasch models were fitted to these data: one model where the equate group items had the same difficulty, and another model where all item parameters were estimated freely (i.e. no equate groups were indicated). The estimated difficulties from these two Rasch models were compared to the true reference item difficulties from the full data.

Table 1
Summary of the conditions in the simulation design

Parameter	**Variation**	**Number of variations**
Number of instruments[a]	2 or 3	2
Difficulty ranges for the items in the instruments ($\delta_{i..I}$)	- No overlap: [-5,-3] and [3, 5] - Close: [-3,-0.1] and [0.1,-3] - Overlap: [-2,1] and [-1,2]	3
Number equates	1, 2, or 5	3
Location equates	- In the center of the instruments - In range of one instrument (not the other) - Evenly spread over both instruments - At the extreme end of the instruments	4
Equate misspecification	Difficulty deviation of 0 to 2 logits with steps of 0.1	21
Abilities (β_n)[b]	- Equal: N(0 ~ 1) - Different: N(-1 ~ 1) and N(1 ~ 1) (2 instruments) or N(-1.5 ~ 1),N(0,5 ~ 1), and N(2,5 ~ 1) (3 instruments)	2

Note: [a] Each instrument contained 10 items, with additional equate items. [b] Data were generated for 1000 persons per instrument.

Comparing the model performance

The quality of the scaling of the difficulties for the two (or three) instruments was evaluated using two statistical measures: the correlation and the misalignment. The correlation (ρ) was calculated between the true difficulty parameters and the estimated difficulty parameters obtained from the models with and without equate groups. Higher correlation corresponds with modelled estimates closer to the scale of the true difficulties.

The misalignment (γ) was measured by estimating the vertical distance between two lines (Fig. 2). One line presents the regression of modelled difficulty estimates for instrument A with the true difficulty parameters. The second line is the regression of estimates for instrument B. The coefficient for misalignment captures whether the estimated difficulty parameters for the instruments are aligned to the same scale and we estimate it by

$$\delta = c + \beta\hat{\delta} + \gamma z$$

where δ is the true difficulty parameters of the items, c is the constant, β is the coefficient for $\hat{\delta}$ which are the estimated difficulties, and γ is the mis-alignment for the instruments z. A larger coefficient indicates more misalignment of the difficulty estimates between the instruments.

Results

Correct equate group specification

Table 2 displays a summary of the most important findings from the simulation study. Appendix B provides a full tabulation of the results. In the scenario where the equate groups were correctly specified, the model with the equate groups correlated larger than 0.990 with the true difficulties. The model without equate groups had correlations with the true difficulties that were smaller than 0.850 in the conditions where difficulties were close or overlapping, and sample abilities were different. The location of the equate groups and the amount of equate groups did not affect the results. We found similar trends for misalignments. When sample abilities differed, the misalignment was > 1.75 logits for the model without equate groups, compared to < 0.35 logits for the model with equate groups. Figure 3presents an example of such a condition. In the model without equate groups, the difficulty estimates for one instrument are structurally larger than the truth (solid line). The difficulty estimates for the other are structurally lower than the truth. Accordingly, instruments are not aligned (y = 1.94) and correlation is 0.85. With equate groups, the estimated difficulties of both instruments are higher than the truth, have near-perfect alignment (y= -0.03) and high correlation with the true difficulties (ρ = 0.99).

Table 2
Simulation study results to compare the model with equate groups to the model without equate groups.

		No Equate groups		***With Equate groups***	
Difficulties	*Abilities*	*ρ*	*γ*	*ρ*	*γ*
No overlap	Equal	0.997	0.625	0.999	0.161
Close	Equal	0.996	0.394	0.999	-0.010
Overlap	Equal	0.996	0.132	0.998	0.032
No overlap	Different	0.963	2.460	0.997	0.243
Close	Different	0.832	2.090	0.999	0.000
Overlap	Different	0.779	1.810	0.998	-0.008

Note: ρ is the correlation between the estimated and the true difficulties, γ is the mis-alignment. The results are averaged over the other conditions.

Equate group misspecification

An equate group is misspecified if its items have different difficulties. If differences are large, such misspecification can affect the performance of the model. Appendix C shows the misspecification in logits that caused the model without equate groups to outperform the model with equate groups on correlation and mis-alignment, respectively.

A small difference between difficulties in an equate group still results in a better model with equate groups, than not using equate groups at all. However, when the difference between difficulties in an equate group increases, i.e. mis-specification is larger, the model performs better when equates are not used. This is especially true when the sample abilities are the same between groups. Nevertheless, when sample abilities were different, the model with one incorrectly specified equate group (i.e. one out of one, two or five) still performed better than the model without any equate groups up to misspecification of 1.7 logits. Also, when the sample abilities were the same, and difficulties of the instruments were not overlapping or close, the model with equate groups was sometimes better than the model without equate groups, for a mis-specification up to about 0.5 logits. When the difficulties overlapped, the model without equate groups mostly outperformed the model with equate groups with a mis-specification equal or larger than 0.2 logits.

Illustrative example

To illustrate the methodology in practice, we applied the Rasch model to three example studies. These studies were part of a larger project to construct a generic score for child development (D-score) performed by the Global Child Development Group (GCDG) [25][26]. These studies were the "Netherlands 1" study, the "Ethiopia" study and the "Colombia 2" study [27][28][29]. The studies used different tools to measure child development. The "Netherlands 1" study gathered longitudinal data for 55 items from the Dutch Development Instrument (DDI) on 2038 children aged 0–2 years [30]. The "Ethiopia" study collected scores on 177 items from the Bayley Scales of Infant and Toddler Development, third edition (Bayley-III) on 506 children aged 1–4 years [31]. The "Colombia 2" study contained data from multiple instruments: 99 items from the Ages and Stages Questionnaire (ASQ), 84 items from the Denver Developmental Screening test (Denver), and 231 items from the Bayley-III, on 1311 children aged 0.5–3.5 years [32][33]. In general, children received subsets of items suited for their age.

Previous research on the DDI identified that the Rasch model provided a good summary of child development scores [25, 34]. To obtain a measurement model on the combined data, we fitted the Rasch model on the combined data, with and without active groups. We applied a strict item selection based on the fit of the items to the Rasch model to select only the best items for the combined scale. Accordingly, items were removed based on item infit and outfit (z-score > 0) until the model contained only items with excellent fit. The fitted model included 185 remaining items: 31 ASQ items, 84 Bayley-III items, 53 DDI items, and 17 Denver items. Eleven candidate equate groups were carefully selected based on expert judgement as part of the GCDG project. Also, we calculated infit and outfit statistics for active equate groups. The resulting solution connects the studies by eight active equate groups. Table 3 provides an overview of items per study and how these are connected.

Table 3
Overview of the equates used to link the three studies and four instruments.

	ASQ **31 items**	**Bayley-III** **84 items**	**DDI** **46 items**	**Denver** **17 items**

Netherlands 1
N = 2038
Rows = 16650
			EQ1; EQ2; EQ3; EQ5; EQ6; EQ8	

Ethiopia
N = 506
Rows = 1089
		EQ1; EQ3; EQ4; EQ5; EQ6; EQ7; EQ8		

Colombia 2
N = 1311
Rows = 1311
	EQ4; EQ7	EQ1; EQ3; EQ4; EQ5; EQ6; EQ7; EQ8		EQ1; EQ2; EQ1; EQ8

Note: N indicates the number of children in the study and Rows the number of measurements.				

Figure 4 displays the latent ability scores for the model without (top-panel) and with (bottom-panel) equate groups. The model without equate groups places each ability distribution around the global mean, thus severely distorting relation with child age. This solution is not suitable for comparing child development across studies. The model with equate groups resolves these issues and results in one common scale. One might wonder why Ethiopia and Colombia appear on a similar scale, even without equate groups. The reason is that these studies have common Bayley-III items. As we may expect, the estimated item difficulties differ between the models with and without equates. For example, when we place items "Sit no support" from Denver, "Sits without support (30 sec)" from Bayley-III and "Sit in stable position without support" from the DDI into an equate group, these have a common difficulty of -9.23. In the unconstrained model, item difficulties vary widely (-16.37, -16.22, -2.46, respectively), which destroys the common scale. Figure 5 displays the probability to pass the item by ability for each equate group. These plots confirm that differences between the item characteristic curves from the different studies and instruments are small, as desired.

Discussion

Combining existing data yields valuable information efficiently. Such big data psychometrics leads to new challenges in scale comparability. This paper presents a solution for the problem when strictly equivalent items across different measurement instruments are lacking. Our method places similar items into equate groups, determines which equate groups to activate, and estimates difficulty parameters of items simultaneously for all instruments. All items within an active equate group receive the same difficulty estimates, thus providing a bridge between different instruments. The equate group method is more flexible and general than methods based on stacking identical items into the same column of the data.

The equate group method can be helpful in meta-analyses of individual person data from different studies that measure the same construct with different instruments. Assuming high-quality equate groups, the method links data from different sources to the same scale. While our application is in child development, the method extends to settings where the combination of data sets would offer new research insights. Examples include the measurement of quality of life, the severity of disabilities, depression and physical activity. Also, when designing a new study, incorporating strategic overlap with other studies improves future equate group possibilities.

Equate groups enhance existing methodologies for improving comparability. Standard concurrent calibration place common-items into the same column [35–38], thereby effectively constraining item parameters to be identical across studies. For example, the McHorney study developed a common metric for physical functioning using concurrent calibration, which requires that all studies are linked by identical common items [36]. Our method applies the constraints within the estimation algorithm. Since this does not require a specific organization of the data, we can easily equate multiple items within and across studies and within and across instruments. The method offers enough flexibility to deal with situations where common items are not perfectly identical or less abundant.

To determine optimal equate group composition and status (active or passive) is a new type of modelling activity. Defining the optimal combination of equate groups is a part of the modelling process that should not be taken lightly. Equate-group diagnostics like Fig. 5, may reveal that one or more items do not fit within the group. In such a case, we may need to remove a misfitting item from the equate group, split the equate group into two more homogeneous equate groups, or decide to inactivate the equate group. There are no cut-and-dried criteria yet for such actions, but, as our simulations show, these decisions may have substantial effects on the solution. Based on our experience thus far, we make the following recommendations in working with equate groups. First, collaborate with subject-matter experts to identify important similarities and differences in item formulations, and ask for a starting assignment of items into equate groups. Second, assess the quality of active equate groups by studying the correspondence between the item characteristics curves and calculating equate fit statistics. Third, compare the ability distributions between studies, and evaluate whether any systematic differences are plausible. Fourth, try to distribute active equate groups across the full range of the measurement scale. Finally, when the abilities of the samples are relatively uniform, try a model without any equate groups, and see whether that solution may be preferable.

A challenge of the equating methods we describe might be that the items included in the model are restricted to items that do not show country-level (or study-level) differential item functioning. In essence, our methodology corresponds to a one-stage meta-analysis with fixed effects for the equate groups. This is contrary to methods described by von Davier & von Davier [39], where the linking functions used are less strict, or the methods described in Oliveri & von Davier [40, 41], where the model is adjusted for country level information in a mixed model. That model is comparable to adding a random effect to allow for heterogeneity between studies in a meta-analysis [42, 43]. These methods may result in models that fit a large amount of items, yet with country specific parameters, which may be a limitation for global applications. The strength or our equate group method is, that the model is restricted to the best items and results in item parameters that are globally interpretable.

The local independence assumption of the Rasch model states that items should be unrelated to each other, given the latent construct. Violation of local independence may be due to a high correlation within a person and overfit. We expect that the impact of violations will not be large in practice, but further study is needed to evaluate the extent of this problem. Another topic for further study could be the extension of equate group methods to less strict item response models. In a Rasch model, only one parameter (i.e. item difficulty) needs to be restricted across the items in the equate group. However, in item response models with more than one parameter (i.e. two or three-parameter logistic models), additional parameters may have to be restricted across equate group items. To our knowledge, there is yet no method to accomplish this.

Conclusions

In general, it is efficient to use and combine existing data sources. However, different sources are often incomparable. We conclude that equate groups are an economical and exciting concept that enables insightful statistical analysis from seemingly disparate data sources. We hope that the broader use of equate groups may advance the utility of existing data for answering new questions.

Abbreviations

GCDG Global Child Development Group

NEAT Nonequivalent anchor test

D-score Generic score for child development

DDI Dutch Development Instrument

Bayley-III Bayley Scales of Infant and Toddler Development, third edition

ASQ Ages and Stages Questionnaire

Denver Denver Developmental Screening test

Declarations

Ethics approval and consent to participate

Not applicable. The illustrative data analysis was performed using deidentified non-clinical data from completed studies without patient or public involvement. No new participants were recruited and no new data were collected.

Consent for publication

Not applicable.

Availability of data and materials

The raw data for the Netherlands 1 and Colombia 2 studies are publicly available under a CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/)1. Authorship remains with the study coordinator, but users are free to redistribute, alter and combine the data, on the condition of giving appropriate credit with any redistributions of the material. The URL of the public data is https://d-score.org/childdevdata/. The raw data for the Ethiopia data are not public and cannot be shared with this publication. However , the reader can apply for access to the data through the study contact: Charlotte Hanlon (charlotte.hanlon@kcl.ac.uk).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation [Grant #OPP1138517]. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

Authors' contributions

The method was developed and defined by all authors. The simulation study and analyses were designed and interpreted by SvB and IE and performed by IE. IE drafted the manuscript, SvB assisted in drafting and critically reviewed the manuscript. AW critically reviewed and edited the article for English language. All authors read and approved the final manuscript.

Acknowledgements

We thank the working group members and data contributors of the Global Child Development Group (GCDG) for their contributions: Orazio Attanasio, Department of Economics, University College London, UK; M. Caridad Araujo, Inter-American Development Bank, Washington DC, USA; Maureen M. Black, RTI International and Department of Pediatrics, University of Maryland School of Medicine, Baltimore, MD, USA; Susan M. Chang, Caribbean Institute for Health Research, The University of the West Indies, Kingston, Jamaica; Bernice M. Doove, Envida JGZ, Maastricht, The Netherlands; Lia C.H. Fernald, School of Public Health, University of California, Berkeley, CA, USA; Emanuela Galasso, The World Bank, Washington DC, USA; Sally Grantham-McGregor, Emerita Professor, Institute of Child Health, University College London, UK; Jena D. Hamadani, International Centre for Diarrhoeal Disease Research, Bangladesh; Charlotte Hanlon, Centre for Global Mental Health, Health Services and Population Research Department, Institute of Psychiatry, Psychology and Neuroscience, King's College London; Department of Psychiatry, School of Medicine, College of Health Sciences, Addis Ababa University, Ethiopia; Pamela Jervis, Centre for the Evaluation of Development Policies, The Institute for Fiscal Studies, UK; Simone M. Karam, Medical Faculty, Department of Pediatrics, PPG Public Health, Federal University of Rio Grande, Brazil; Betsy Lozoff, Center for Human Growth and Development, University of Michigan, Ann Arbor, MI, USA; Girmay Medhin, Aklilu Lemma Institute of Pathobiology, Addis Ababa University, Addis Ababa, Ethiopia; Ana M. B. Menezes and the 1993 Pelotas cohort team, Postgraduate Program in Epidemiology, Federal University of Pelotas, Brazil; Helen Pitchik, School of Public Health, University of California, Berkeley, CA, USA; Lisy Ratsifandrihamanana, Centre Médico-Educatif "Les Orchidées Blanches", Antananarivo, Madagascar; Sarah Reynolds, School of Public Health, University of California, Berkeley, CA, USA; Marta Rubio-Codina, Inter-American Development Bank, Washington DC, USA; Norbert Schady, Inter-American Development Bank, Washington DC, USA; Susan P. Walker, Caribbean Institute for Health Research, The University of the West Indies, Kingston, Jamaica; Linda Richter, DST-NRF Centre of Excellence in Human Development, University of the Witwatersrand, Johannesburg, South Africa

Data availability statement

The raw data for the Netherlands 1 and Colombia 2 studies are publicly available under a CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/)1. Authorship remains with the study coordinator, but users are free to redistribute, alter and combine the data, on the condition of giving appropriate credit with any redistributions of the material. The URL of the public data is https://d-score.org/childdevdata/. The raw data for the Ethiopia data are not public and cannot be shared with this publication. However , the reader can apply for access to the data through the study contact: Charlotte Hanlon (charlotte.hanlon@kcl.ac.uk).

References

1. Curran PJ, Hussong AM. Integrative Data Analysis: The Simultaneous Analysis of Multiple Data Sets. Psychol Methods [Internet]. 1992;14:81–100. Available from: .
2. Hussong AM, Curran PJ, Bauer DJ. Integrative Data Analysis in Clinical Psychology Research. Annu Rev Clin Psychol [Internet]. Annual Reviews; 2013;9:61–89. Available from: http://www.annualreviews.org/doi/10.1146/annurev-clinpsy-050212-185522
3. Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. Psychol Assess [Internet]. American Psychological Association Inc.; 2014;26:513–27. Available from: /pmc/articles/PMC5515387/
4. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedogogische Instituut; 1960.
5. Kolen MJ, Brennan RL. Test equating, scaling, and linking: methods and practices. 3rd ed. Springer-Verlag New York; 2014.
6. Dorans NJ, Pommerich M, Holland PW, editors. Linking and Aligning Scores and Scales [Internet]. New York, NY: Springer New York; 2007. Available from: http://link.springer.com/10.1007/978-0-387-49771-6
7. Harris DJ. Practical Issues in Vertical Scaling. Link Aligning Scores Scales [Internet]. New York, NY: Springer New York; 2007. p. 233–51. Available from: http://link.springer.com/10.1007/978-0-387-49771-6_13
8. von Davier AA, Holland PW, Thayer DT. The Kernel Method of Test Equating [Internet]. New York, NY: Springer New York; 2004. Available from: http://link.springer.com/10.1007/b97446
9. Kolen MJ, Brennan RL. Item Response Theory Methods. Test Equating, Scaling, Link [Internet]. New York, NY: Springer New York; 2014. p. 171–245. Available from: http://link.springer.com/10.1007/978-1-4939-0317-7_6
10. Marco GL. Item Characteristic Curve Solutions to Three Intractable Testing Problems [Internet]. J. Educ. Meas. National Council on Measurement in Education; 1977. p. 139–60. Available from: https://www-jstor-org.vu-nl.idm.oclc.org/stable/1434012
11. Wingersky MS, Lord FM. An Investigation of Methods for Reducing Sampling Error in Certain IRT Procedures. Appl Psychol Meas [Internet]. Sage PublicationsSage CA: Thousand Oaks, CA; 1984;8:347–64. Available from: http://journals.sagepub.com/doi/10.1177/014662168400800312
12. Kim S-H, Cohen AS. A Comparison of Linking and Concurrent Calibration Under Item Response Theory. Appl Psychol Meas [Internet]. SAGE PUBLICATIONS, INC.2455 Teller Road, Thousand Oaks, CA 91320; 1998;22:131–43. Available from: http://journals.sagepub.com/doi/10.1177/01466216980222003
13. Taherbhai HM, Seo DY. Comparing concurrent versus fixed parameter equating with common items: using the dichotomous and partial credit models in a mixed-item format test. J Appl Meas [Internet]. 2007;8:84–96. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17215567
14. van Buuren S, de Leeuw J. Equality Constraints in Multiple Correspondence Analysis. Multivariate Behav Res [Internet]. Lawrence Erlbaum Associates, Inc.; 1992;27:567–83. Available from:

http://www.tandfonline.com/doi/abs/10.1207/s15327906mbr2704_4

15. Eekhout I, van Buuren S. Tuning instruments to Unity. In: van Buuren S, Eekhout I, editors. Child Dev with D-score. 2021.
16. Wright B, Stone M. Measurement Essentials. 2nd Editio. Wilmington, Delaware: Wide Range; 1999.
17. Kopf J, Zeileis A, Strobl C. Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches. Educ Psychol Meas [Internet]. SAGE Publications; 2015;75:22–56. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29795811
18. Millsap RE, Everson HT. Methodology Review: Statistical Approaches for Assessing Measurement Bias. Appl Psychol Meas [Internet]. Sage PublicationsSage CA: Thousand Oaks, CA; 1993;17:297–334. Available from: http://journals.sagepub.com/doi/10.1177/014662169301700401
19. van Buuren S, Eekhout I. dmetric: Tools to Investigate the D-score Metric, R package version 0.52.0 [Internet]. 2021. Available from: https://github.com/D-score/dmetric
20. Robitzsch A. sirt: Supplementary Item Response Theory Models, R package version 3.9-4 [Internet]. CRAN; 2020. Available from: https://cran.r-project.org/package=sirt
21. Robitzsch A, Steinfeld J. Item response models for human ratings: Overview, estimation methods, and implementation in R [Internet]. Psychol. Test Assess. Model. 2018. Available from: https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2018_20180323/6_PTAM_IRMHR_Main__2018-03-13_1416.pdf
22. Hanson BA, Béguin AA. Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design. Appl Psychol Meas [Internet]. Sage PublicationsSage CA: Thousand Oaks, CA; 2002;26:3–24. Available from: http://journals.sagepub.com/doi/10.1177/0146621602026001001
23. Kim S, Kolen MJ. Effects on Scale Linking of Different Definitions of Criterion Functions for the IRT Characteristic Curve Methods. J Educ Behav Stat [Internet]. SAGE PublicationsSage CA: Thousand Oaks, CA; 2007;32:371–97. Available from: http://journals.sagepub.com/doi/10.3102/1076998607302632
24. LeBeau B. Ability and Prior Distribution Mismatch: An Exploration of Common-Item Linking Methods. Appl Psychol Meas [Internet]. SAGE PublicationsSage CA: Los Angeles, CA; 2017;41:545–60. Available from: http://journals.sagepub.com/doi/10.1177/0146621617707508
25. van Buuren S. Growth charts of human development. Stat Methods Med Res [Internet]. 2014;23:346–68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23487019
26. Weber AM, Rubio-Codina M, Walker SP, van Buuren S, Eekhout I, Grantham-McGregor S, et al. The D-score: a metric for interpreting the early development of infants and toddlers across global settings. BMJ Glob Heal [Internet]. 2019;4:e001724. Available from: https://pubmed.ncbi.nlm.nih.gov/31803508/
27. Herngreen WP, Reerink JD, van Noord-Zaadstra BM, Verloover-Vanhorick SP, Ruys JH. SMOCC: Design of a Representative Cohort-study of Live-born Infants in the Netherlands. Eur J Public Health

[Internet]. Oxford University Press; 1992;2:117–22. Available from: https://academic.oup.com/eurpub/article-lookup/doi/10.1093/eurpub/2.2.117

28. Hanlon C, Medhin G, Alem A, Tesfaye F, Lakew Z, Worku B, et al. Impact of antenatal common mental disorders upon perinatal outcomes in Ethiopia: the P-MaMiE population-based cohort study. Trop Med Int Heal [Internet]. 2009;14:156–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19187514
29. Rubio-Codina M, Araujo MC, Attanasio O, Muñoz P, Grantham-McGregor S. Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies. Carpenter DO, editor. PLoS One [Internet]. Public Library of Science; 2016;11:e0160962. Available from: http://dx.plos.org/10.1371/journal.pone.0160962
30. Van Wiechen W. Ontwikkelingsonderzoek op het consultatiebureau. Nationale Kruisvereniging; 1988.
31. Bayley N. Bayley Scales of Infant and Toddler Development®, Third Edition [Internet]. San Antonio, TX: Harcourt Assessment; 2006. Available from: https://www.pearsonclinical.com/products/100000123/bayley-scales-of-infant-and-toddler-development-third-edition-bayley-iii.html
32. Squires J, Bricker D, Twombly E, Nickel R, Clifford J, Murplhy K. Ages & Stages English Questionnaires, Third Edition (ASQ-3): A Parent-Completed, Child-Monitoring System. Baltimore, MD: Paul H. Brookes Publishing Co; 2009.
33. Frankenburg WK, Dodds J, Archer P, Shapiro H, Bresnick B. The Denver II: a major revision and restandardization of the Denver Developmental Screening Test. Pediatrics [Internet]. 1992;89:91–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1370185
34. Jacobusse G, van Buuren S, Verkerk PH. An interval scale for development of children aged 0–2 years. Stat Med [Internet]. 2006;25:2272–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16143995
35. Hopman-Rock M, Dusseldorp E, Chorus A, Jacobusse G, Ruetten A, Van Buuren S. Response Conversion for Improving Comparability of International Physical Activity Data [Internet]. J. Phys. Act. Heal. 2012. Available from: http://www.public-health.tu-dresden.de/dotnetnuke3/Portals/5/Projects/EUPASS/appendix b.pdf.
36. McHorney C, Cohen A. Equating health status measures with item response theory: illustrations with functional status items. Med Care [Internet]. 2000;38:II43–59. Available from: https://www.jstor.org/stable/3768062
37. van Buuren S, Hopman-Rock M. Revision of the ICIDH Severity of Disabilities Scale by data linking and item response theory. Stat Med [Internet]. 2001;20:1061–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11276036
38. Van Buuren S, Eyres S, Tennant A, Hopman-Rock M. Improving Comparability of Existing Data by Response Conversion. J Off Stat [Internet]. 2005;21:53–72. Available from: https://stefvanbuuren.name/publications/Improving comparabilty - JOS 2005.pdf

39. von Davier M, von Davier AA. A unified approach to IRT scale linking and scale transformations. ETS Res Rep Ser [Internet]. 2004;2004:i–21. Available from: http://doi.wiley.com/10.1002/j.2333-8504.2004.tb01936.x

40. Oliveri ME, Von Davier M. Investigation of model fit and score scale comparability in international assessments [Internet]. Psychol. Test Assess. Model. 2011. Available from: https://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

41. Oliveri ME, von Davier M. Toward Increasing Fairness in Score Scale Calibrations Employed in International Large-Scale Assessments. Int J Test [Internet]. Taylor & Francis Group; 2014;14:1–21. Available from: http://www.tandfonline.com/doi/abs/10.1080/15305058.2013.825265

42. Barili F, Parolari A, Kappetein PA, Freemantle N. Statistical Primer: heterogeneity, random- or fixed-effects model analyses?†. Interact Cardiovasc Thorac Surg [Internet]. 2018;27:317–21. Available from: https://doi.org/10.1093/icvts/ivy163

43. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ [Internet]. BMJ Publishing Group Ltd; 2011;342. Available from: https://www.bmj.com/content/342/bmj.d549

Figures

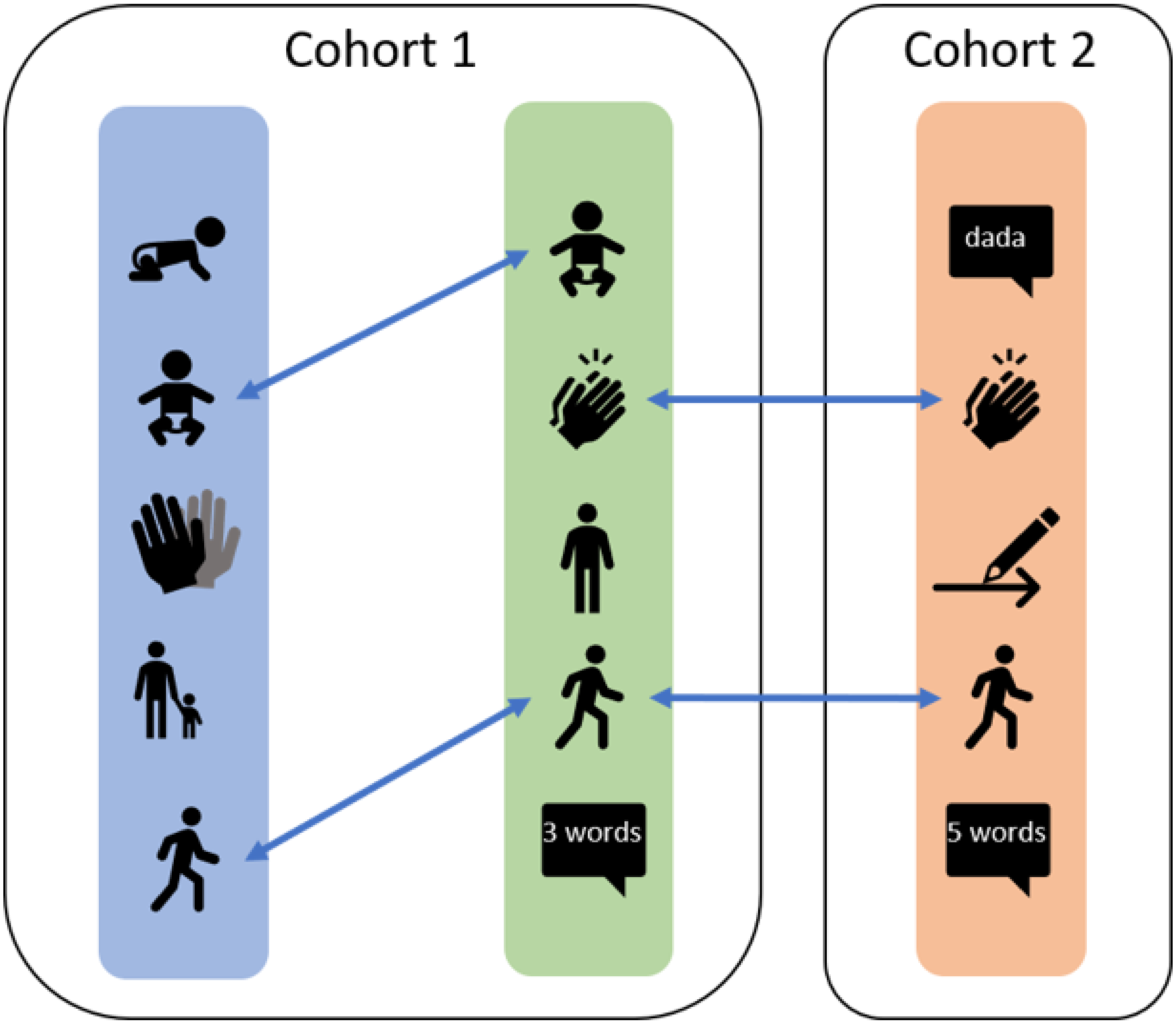

Figure 1

Example of three instruments, that can be linked via common items.

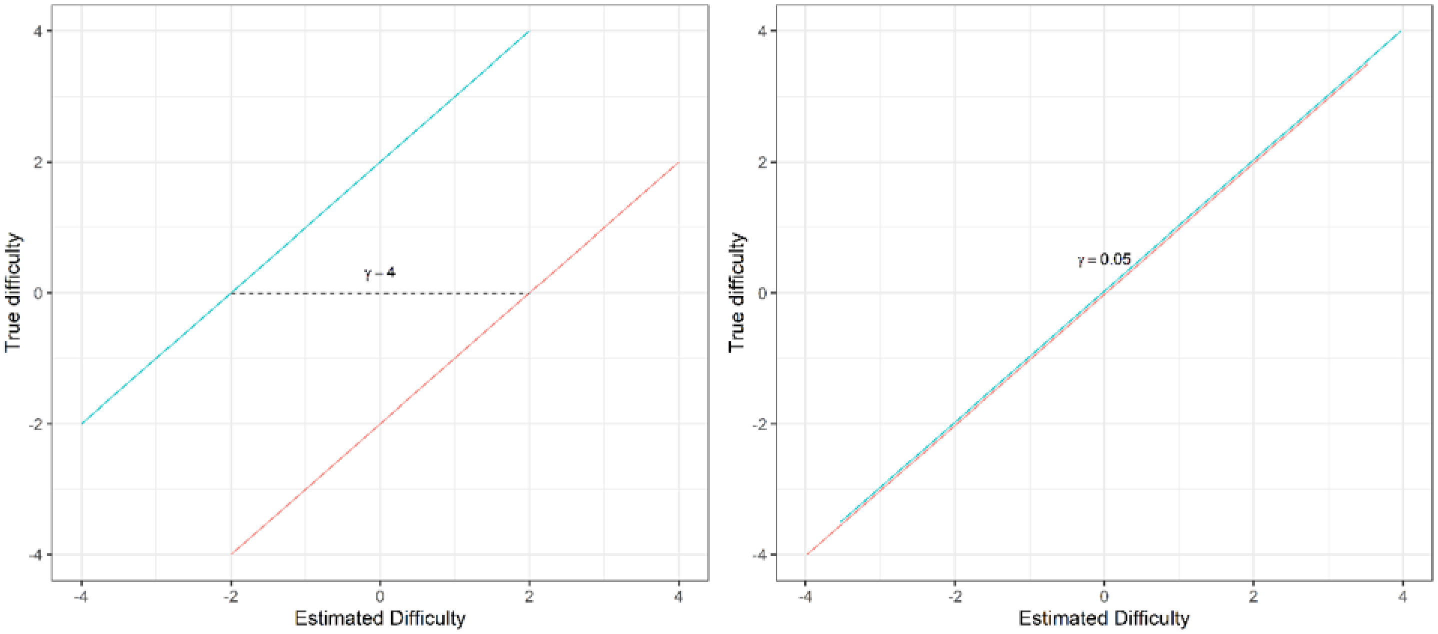

Figure 2

Illustration of the mis-alignment parameter.

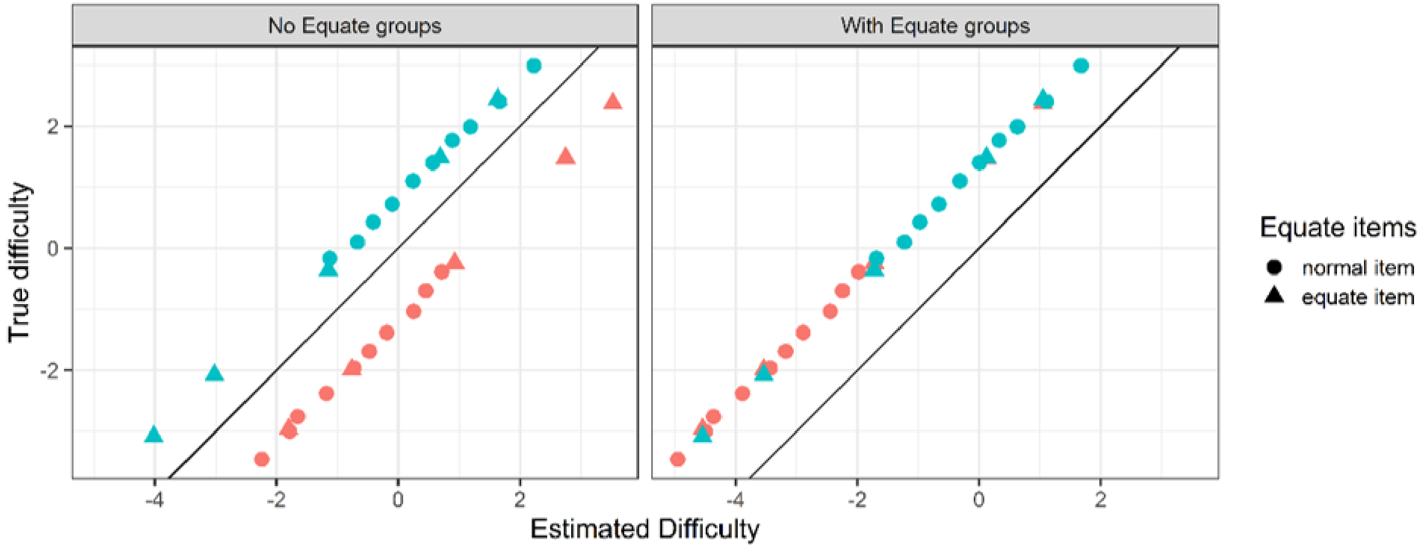

Figure 3

Difficulty estimates from model without equate groups (ρ = 0.85; γ =1.94) (left) and model with equate groups (ρ = 0.99; γ =-0.03) (right) where difficulty ranges are close, cohort abilities differ and five equate groups are spread through both instruments. Difficulty estimates are colored by instrument.

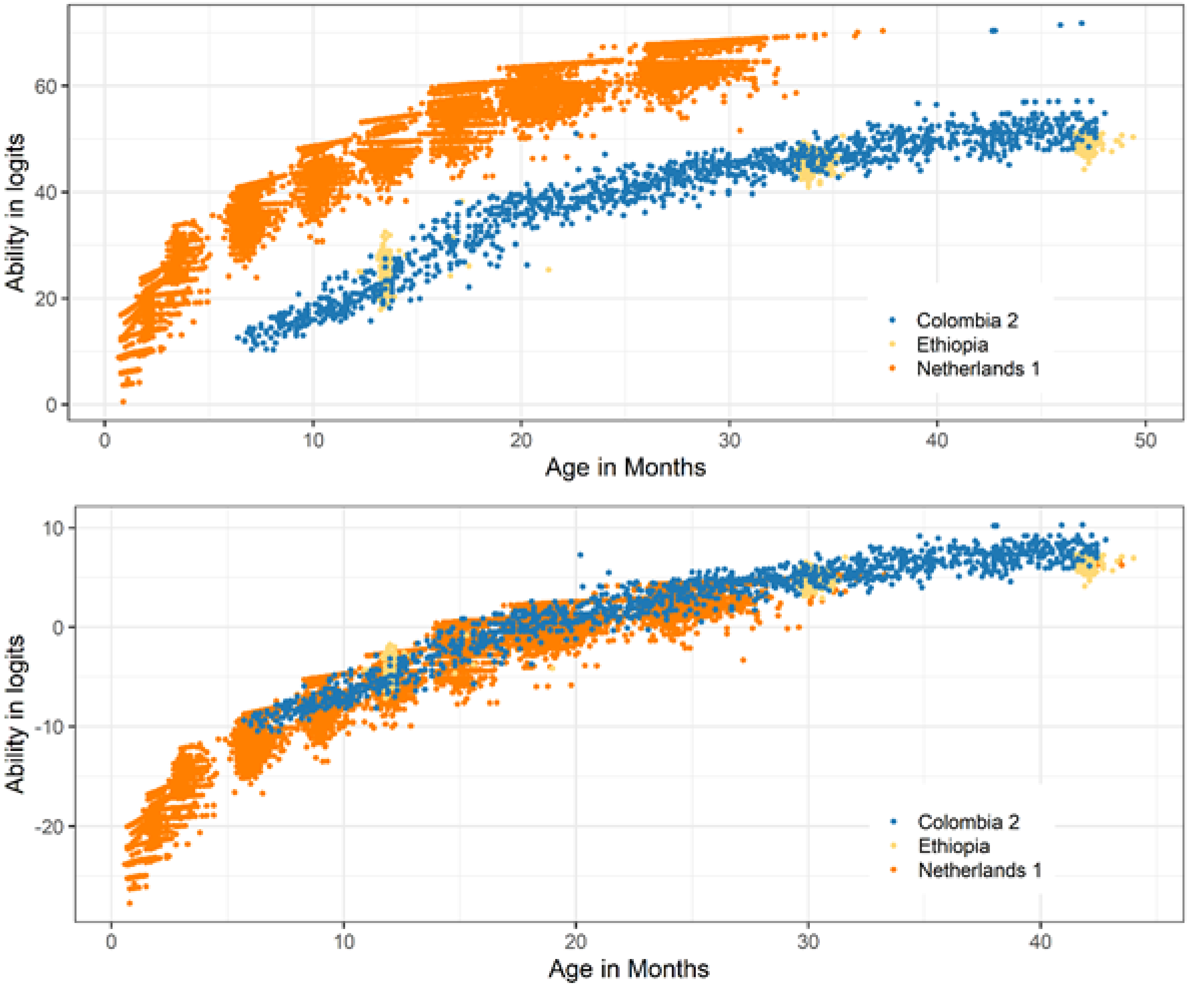

Figure 4

Latent ability in logits for age for the three illustrative studies. The top panel results from the model without equate groups and the bottom panel results from the model with equate groups.

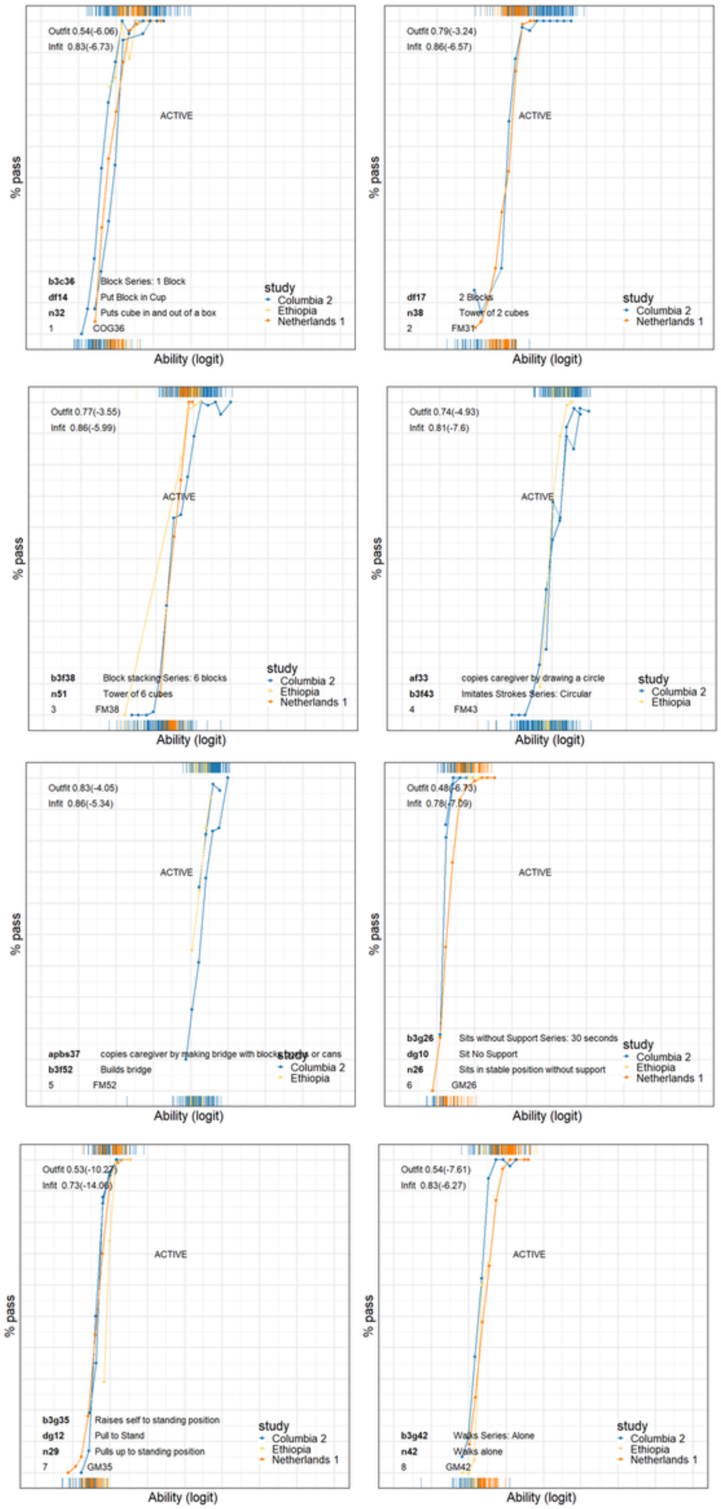

Figure 5

Percentage pass for ability in the illustrative data for the equate groups.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixAcodedatasimulationforpaper.pdf
- AppendixB.pdf
- AppendixC.pdf