2.4 Equate groups

Iris Eekhout[1]
Stef van Buuren[1,2]
[1]Netherlands Organisation for Applied Scientific Research TNO, Leiden, 2316 ZL, The Netherlands
[2]University of Utrecht, Utrecht, 3584 CH, The Netherlands

This section introduces the concepts and tools needed to link assessments made by different instruments administered across multiple cohorts. Our methodology introduces the idea of an equate group. Systematic application of equate groups provides a robust yet flexible methodology to link different instruments. Once the links are in place, we may combine the data to enable meta-analyses and related methods.

- What is an equate group? (2.4.1)
- Concurrent calibration (2.4.2)
- Strategy to form and test equate groups (2.4.3)
- Statistical framework (2.4.4)
- Common latent scale (2.4.5)
- Quantifying equate fit (2.4.6)
- Differential item functioning (2.4.7)

2.4.1 WHAT IS AN EQUATE GROUP?

An *equate group* is a set of two or more milestones that measure the same thing in (perhaps slightly) different ways. Table 2.3.2 contains an example of an equate group, containing items that measure the ability to form two-word sentences. Also, Figure 2.3.2 and Figure 2.3.3 show examples of equate groups.

Equate groups vary in quality. We can use high-quality equate groups to link instruments by restricting the difficulty of all milestones in the equate group to be identical. Equate groups thus provide a method for bridging different tools.

Figure 2.4.1 displays items from three different instruments with overlapping sets of milestones. The shared items make up equate groups, as presented by the arrows between them. In the example, all three instruments share one milestone ("walk alone"). The "sitting" and "clap hand" items appear in two tools. So in total, there are three equate groups.

DOI: 10.1201/9781003216315-15

2.4.2 CONCURRENT CALIBRATION

Patterns as in Figure 2.4.1 occur if we have multiple forms of the same instrument. Although in theory, there might be sequence effects, the usual working assumption is that we may ignore them. Equate groups with truly shared items that work in the same way across samples are of high quality. We may collect the responses on identical items into the same column of the data matrix. As a consequence, usual estimation methods will automatically produce one difficulty estimate for that column (i.e. common item).

The procedure described above is known as *concurrent calibration*. See Kim & Cohen (1998) for more background. The method simultaneously estimates the item parameters for all instruments. Concurrent calibration is an attractive option for various reasons:

- It yields a common latent scale across all instruments;
- It is efficient because it calibrates all items in a single run;
- It produces more stable estimates for common items in small samples.

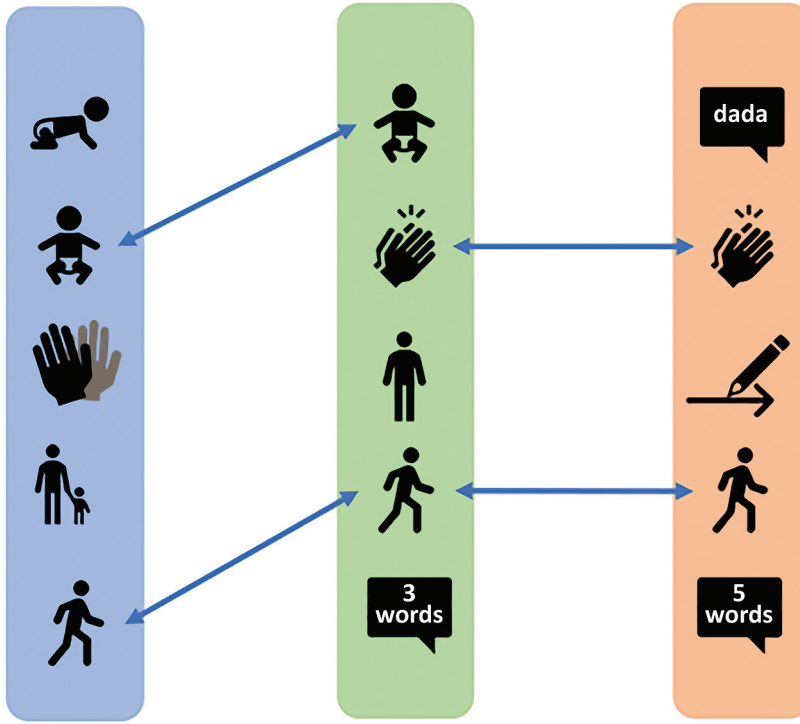

FIGURE 2.4.1 Example of three instruments that are bridged by common items in equate groups.

However, concurrent calibration depends on a strict distinction between items that are indeed the same across instruments and items that differ.

In practice, strict black-white distinctions may not be possible. Items that measure the same skill may have been adapted to suit the format of the instrument (e.g. number of response options, question formulation, and so on). Also, investigators may have altered the item to suit the local language and cultural context. Such changes may or may not affect the measurement properties. The challenge is to find out whether items measure the underlying construct in the same way.

In practice, we may need to perform concurrent calibration to multiple - perhaps slightly dissimilar - milestones. When confronted with similar - but not identical - items, our strategy is first to form provisional equate groups. We then explore, test and rearrange these equate groups, in the hope of finding enough high-quality equate groups that will bridge instruments.

2.4.3 STRATEGY TO FORM AND TEST EQUATE GROUPS

An equate group is a collection of items. Content matter experts may form equate groups by evaluating the contents of items and organizing them into groups with similar meaning. The modelling phase takes this set of equate groups (which may be hundreds) as input. Based on the analytic result, we may activate or modify equate groups. It is useful to distinguish between *active* and *passive* equate groups. What do we mean by these terms?

- *Active equate group*: The analysis treats all items within an active equate group as one super-item. The items obtain the same difficulty estimate and are assumed to yield equivalent measurements. As the items in an active equate group may originate from different instruments, such a group acts as a bridge between instruments.
- *Passive equate group*: Any non-active equate groups are called passive. The model does not restrict the difficulty estimates, i.e., the milestones within a passive equate group will have separate difficulty estimates.

Since active equate groups bridge different instruments, they have an essential role in the analysis. In general, we will set the status of an equate group to active *only* if we believe that the milestones in that group measure the underlying construct in the same way. Note that this does not necessarily imply that all items need to be identical. In Table 2.3.2, for example, small differences exist in item formulation. We may nevertheless believe that these are irrelevant and ignore these in practice. Reversely, there is no guarantee that the same milestone will measure child development in the same way in different samples. For example, a milestone like “climb stairs” (Figure 2.4.2) could be more difficult (and more dangerous) for children who have never seen a staircase.

The data analysis informs decisions to activate equate groups. The following steps implement our strategy for forming and enabling equate groups:

FIGURE 2.4.2 One year old child climbs stairs.
Photo by Iris Eekhout.

- Content matter experts compare milestones from different instruments and sort similar milestones into equate groups. It may be convenient to select one instrument as a starting point, and map items from others to that (see section 2.3.2);
- Visualize age profiles of mapped items (see section 2.3.3). Verify the plausibility of potential matches through similar age profiles. Break up mappings for which age profiles appear implausible. This step requires both statistical and subject matter expertise;
- Fit the model to the data using a subset of equate groups as active. Review the quality of the solution and optimize the quality of the links between tools by editing the equate group structure. The technical details

of this model are explained in section 2.4.4. Refit the model until (1) active equate groups link all cohorts and instruments, (2) active equate groups are distributed over the full-scale range (rather than being centred at one point);
- Assess the quality of equate groups by the infit and outfit (see section 2.4.6).
- Test performance of the equate groups across subgroups or cohorts by methods designed to detect differential item functioning (see section 2.4.7).

The application of equate groups is needed to connect different instruments to a universal scale. The technique is especially helpful in the situation where abilities differ across cohorts.

If the cohort abilities are relatively uniform (for example as a result of experimental design) and if the risk of misspecification of the equate groups is high, a good alternative is to rely on the equality of ability distribution. In our application, this was not an option due to the substantial age variation between cohorts.

2.4.4 PARAMETER ESTIMATION WITH EQUATE GROUPS

The Rasch model is the preferred measurement model for child development data. Section 1.4 provides an introduction of the Rasch model geared towards the D-score.

The Rasch model expresses the probability of passing an item as a logistic function of the difference between the person ability β_n and the item difficulty δ_i. The model (2.4.1) is defined as

$$\pi_{ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}$$

Formula 2.4.1.

One way to interpret the formula is as follows. The logarithm of the odds that a person with ability β_n passes an item of difficulty δ_i is equal to the difference $\beta_i - \delta_i$ (Wright & Masters, 1982). See the logistic model in Section 1. 4.6.1 for more detail.

In model (2.4.1) every milestone i has one parameter δ_i. We extend the Rasch model by restricting the δ_i of all items within the same equate group to the same value. We thereby effectively say that these items are interchangeable measures of child development.

Estimation of the parameter for the equate group is straightforward. Wright & Masters (1982) present a simple method for aligning two test forms with common items. There are three steps:

- Estimate the separate δ_i's per item;
- Combine these estimates into δ_q by calculating their weighted average;
- Overwrite each δ_i by δ_q.

TABLE 2.4.1
Overview of the symbols used in equations (2.4.1) and (2.4.2).

Symbol	Term	Description
β_n	Ability	True (but unknown) developmental score of child n
δ_i	Difficulty	True (but unknown) difficulty of item i
δ_q	Difficulty	The combined difficulty of the items in equate group q
π_{ni}	Probability	Probability that child n passes item i
l		The number of items in the equate group
w_i		The number of respondents with an observed score on item i

Suppose that Q is the collection of items in equate group q, and that w_i is the number of respondents for item i. The parameter estimate δ_q for the equate group is

$$\delta_q = \frac{\sum_{i \in Q} \delta_i w_i}{\sum_{i \in Q} w_i}$$

Equation 2.4.2.

2.4.5 COMMON LATENT SCALE

The end goal for using the equate group method to model development items is to measure development on one common latent scale, the D-score. That way, the measure (i.e. D-score) can be obtained, irrespective of which instrument is used in which population.

Figure 2.4.3 displays the D-score estimates by age in three cohorts from the GCDG study: Netherlands 1 (`GCDG-NLS-SMOCC`), Ethiopia (`GCDG-ETH`) and Colombia 2 (`GCDG-COL-LT42M`) for two different analyses. As described in section 2.2.2, the Netherlands 1 study administered the `ddi`; Ethiopia measured children by the `by3`; and Colombia collected data on `by3`, `den`, `aqi` and `bdi`. Accordingly, there is an overlap in items between Ethiopia and Colombia via the `by3`, but the Netherlands 1 cohort is not linked.

We created the plot on the left-hand side without active equate groups. The large overlap between Ethiopian and Columbian children occurs because the scales for these studies are linked naturally via shared items from `by3`. Since the `ddi` instrument is not connected, the Dutch cohort follows a different track. While we can compare D-scores between Ethiopia and Colombia, it is nonsensical to compare Dutch to either Ethiopia or Colombia. The right-hand side plot is based on an analysis that used active equate groups to link the cohorts. Since the analysis connected the scales for all three cohorts, we can now compare D-scores obtained between all three cohorts.

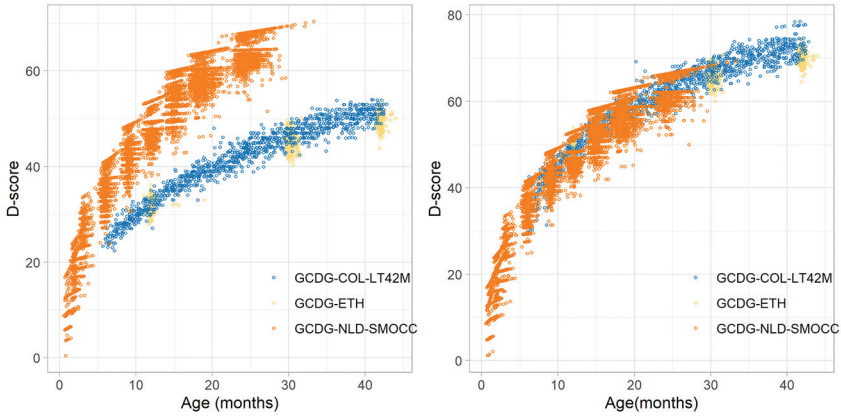

FIGURE 2.4.3 Example of three cohorts with and without equate group linking.

This example demonstrates that active equate groups form the key for converting ability estimates for children from different cohorts using different instruments onto the same scale.

2.4.6 QUANTIFYING EQUATE FIT

It is essential to activate only those equate groups for which the assumption of equivalent measurement holds. We have already seen the *item fit* and *person fit* diagnostics of the Rasch model. This section describes a similar measure for the quality of an active equate group.

2.4.6.1 EQUATE FIT

Section 1.6 defines the observed response of person n on item i as x_{ni}. The accompanying standardized residual z_{ni} is the difference between x_{ni} and the expected response P_{ni}, divided by the expected binomial standard deviation,

$$z_{ni} = \frac{x_{ni} - P_{ni}}{\sqrt{W_{ni}}}$$

with variances $W_{ni} = P_{ni}(1 - P_{ni})$.

Equate infit is an extension of item infit that takes an aggregate over all items i in active equate group q, i.e.,

$$\text{Equateinfit} = \frac{\sum_{i \in q} \sum_{n}^{N} (x_{ni} - P_{ni})^2}{\sum_{i \in q} \sum_{n}^{N} W_{ni}}.$$

Likewise, we calculate *Equate outfit* of group q as

$$\textbf{Equate outfit} = \frac{\sum_{i \in q} \sum_{n}^{N_i} z_{ni}^2}{\sum_{i \in q} N_i},$$

where N_i is the total number of responses observed on item i. The interpretation of these diagnostics is the same as for item infit and item outfit.

Note that these definitions implicitly assume that the expected response P_{ni} is calculated under a model in which all items in equate group q have the same difficulty. This is not true for passive equate groups. Of course, no one can stop us from calculating the above equate fit statistics for passive groups, but such estimates would ignore the between-item variation in difficulties, and hence gives a too optimistic estimate of quality. The bottom line is: *The interpretation of the equate fit statistics should be restricted to active equate groups only.*

2.4.6.2 Examples of well fitting equate groups

The evaluation of *equate fit* involves comparing the observed probabilities of endorsing the items in the equate group to the estimated probability of endorsing the items in the equate group. For an equate group there is an empirical curve for each item in the equate group and one shared estimated curve. The empirical curves should all be close together, and close to the estimated curve for a good equate fit.

Figure 2.4.4 shows a diagnostic plot for equate groups `REC6` (Turns head to sound of bell) and `GM42` (Walks alone). The items within `REC6` have slightly different formats in the Bayley I (`by1`), Dutch Development Instrument (`ddi`), and the Denver (`den`). The empirical curves in the upper figure show good overlap, but note that hardly any negative responses were recorded for four of the five studies, so the shared estimate depends primarily on the Dutch sample. Items from equate group `GM42` appear in six instruments: `bar`, `by1`, `by2`, `by3`, `ddi`, and `gri`. Also, here the empirical data are close together, and even a little steeper than the fitted dashed line, which indicates a good equate fit. The infit and outfit indices, shown in the upper left corners, confirm the good fit (fit < 1).

2.4.6.3 Examples of equate groups with poor equate fit

Poor fitting equate groups are best treated as passive equate groups, so that items in those groups are not restricted to the same difficulty. Empirical item curves with different locations and slopes indicate a poor fit. Additionally, the equate fit indices will indicate a poor fit (fit > 1).

Figure 2.4.5 shows examples for groups `COG24` (Bangs in play / Bangs 2 blocks) and `EXP12` (Babbles). In both cases there is substantial variation in location between the empirical curves. For `COG24` we find that the fitted curve

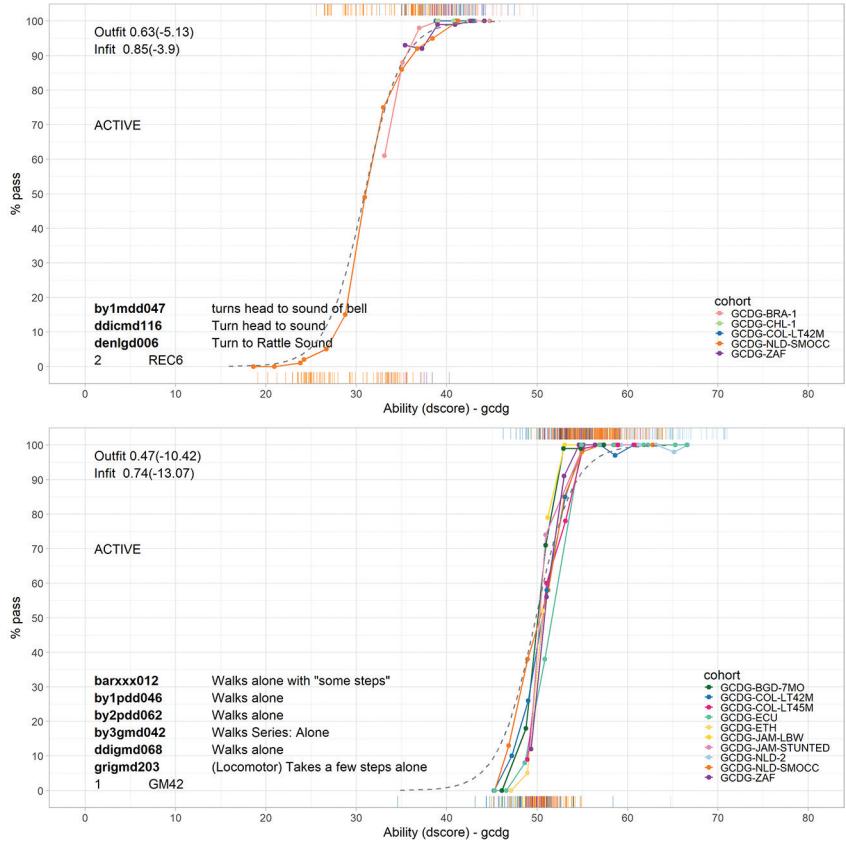

FIGURE 2.4.4 Two equate groups that present a good equate fit.

is closer to the `den` item, which suggests that the equate difficulty is mostly based on the `den` item. Items from equate group `EXP12` have a different format in instruments `by1`, `ddi` and `gri`. The empirical curves, with different colours for each instrument, are not close to each other, nor close to the fitted curve. Note that all infit and outfit statistics are fairly high, indicating poor fit. Both equates are candidates for deactivation in a next modelling step.

2.4.7 DIFFERENTIAL ITEM FUNCTIONING

Items within an active equate group should work in the same way across the different cohorts, i.e., they have no differential item functioning (DIF). The assumption of no DIF is critical for active equate groups. If violated, restricting the difficulty parameters as equal across cohorts may introduce unwanted bias in comparisons between cohorts. This section illustrates the role of DIF in equate groups.

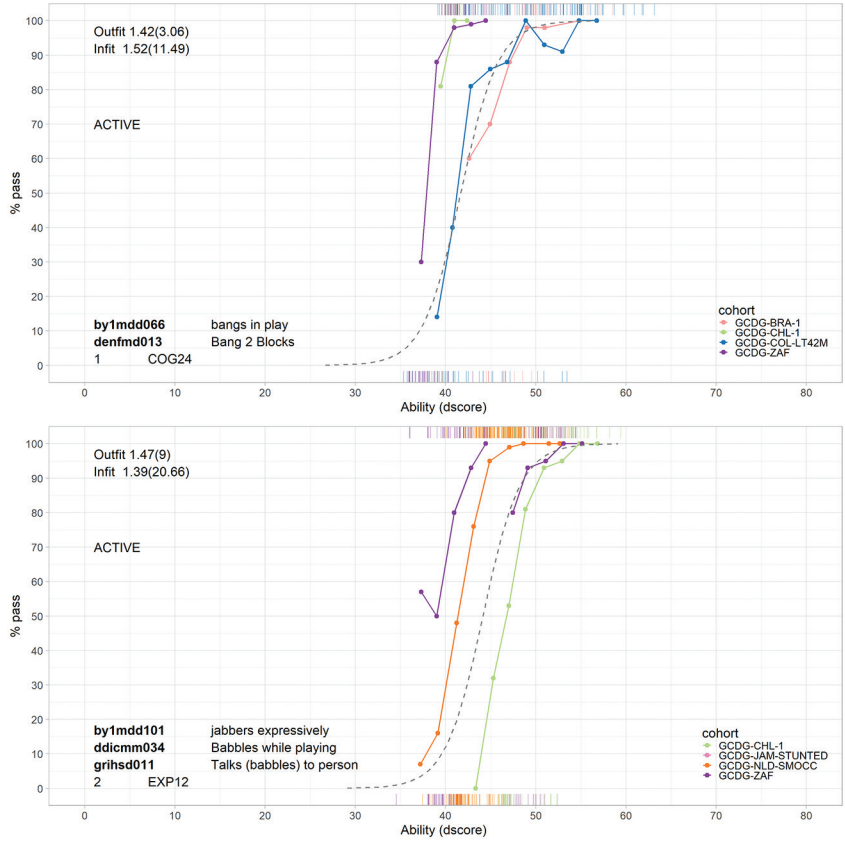

FIGURE 2.4.5 Two equate groups that present a poor equate fit.

2.4.7.1 Good equate groups without DIF

Section 1.6.3 discusses the role of DIF in the evaluation of the fit of items to the Rasch model. This section illustrates similar issues in the context of equate groups.

Figure 2.4.6 shows the empirical curves of two equate groups, `FM31` (two cubes) and `EXP26` (two-word sentence). All curves are close to each other, so there is no differential item functioning here.

2.4.7.2 Poor equate groups with DIF for study

Figure 2.4.7 plots the empirical curves for equate groups `GM44` (throws ball) and `EXP23` (5 or more words). The substantial variation between these curves

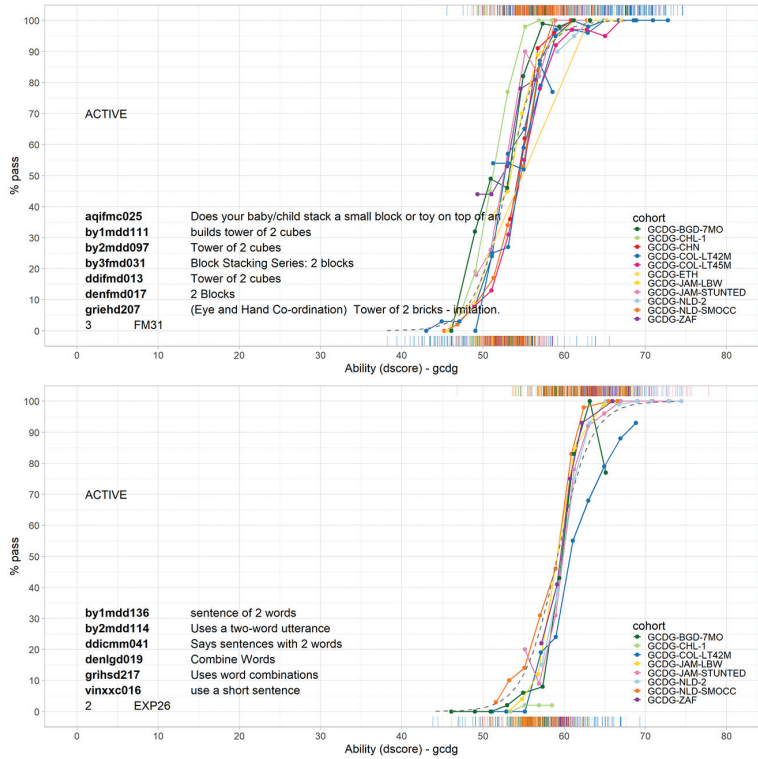

FIGURE 2.4.6 Two equate groups that present no differential item functioning between cohorts.

is a sign of differential item functioning. For example, *Throws ball* is easier for children in the South-Africa cohort (purple curve; `GCDG-ZAF`) and more difficult for children in Colombia (blue curve; `GCDG-COL-LT42M`). In other words, the probability of passing the item given the D-score (i.e. item difficulty) differs between the cohorts. Likewise, there is differential item functioning for *Says more than 5 words*. This milestone is easier for children in Jamaica (yellow and pink curves; `GCDG-JAM-LBW` and `GCDG-JAM-STUNTED`) than for children from Ecuador (green; `GCDG-ECU`).

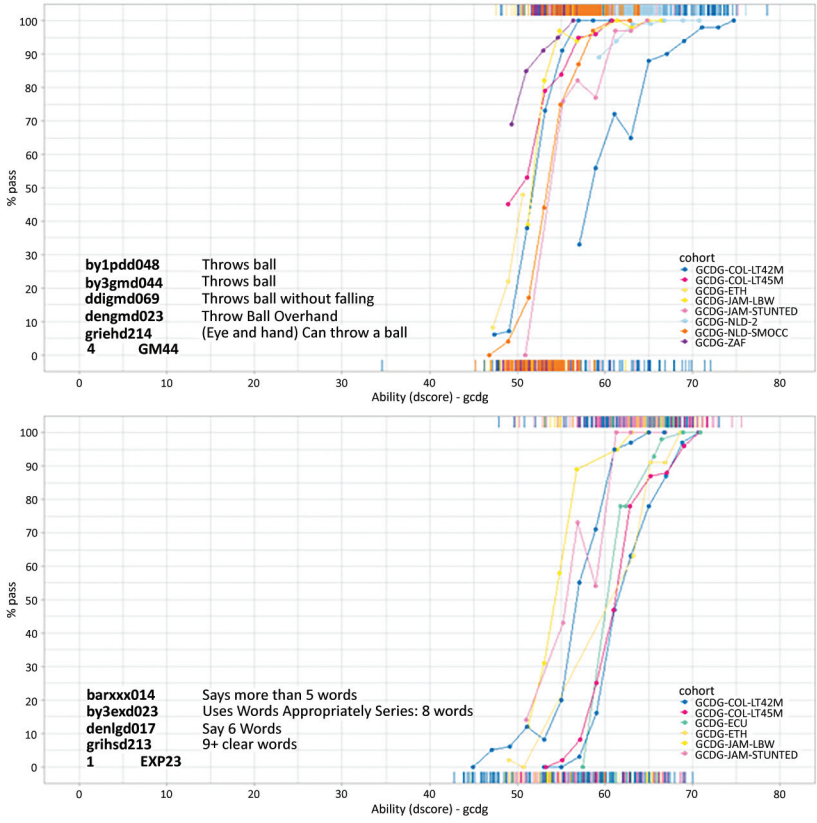

FIGURE 2.4.7 Two equate groups that present differential item functioning between cohorts.