# Measurement invariance

Ben Stenhaug & Ben Domingue

Stanford University

**Abstract**

ABSTRACT TO DO

## Contents

# 1 TO DO

- be sure all my directions match. i think i switched everything to be item easiness at some point but not sure if i did it throughout

# 2 Introduction

Inspired by Camilli (1992), we view differential item functioning (DIF) as bias against a group of students that manifests when an item response model with too few ability dimensions is imposed. From this perspective, the term "differential item functioning" is, perhaps, a misnomer as DIF is a property of the student, not the item. For example, Ackerman (1992) describes a scenario in which a test intends to measure a student's math ability, but their performance also depends on their verbal ability. In this case, math ability is the "target ability," and verbal ability is the "nuisance ability." Fitting a unidimensional item response model to this test results in students with low verbal ability receiving a score systematically lower than their true math ability; therein lies DIF.

Contrarily, the usual setup of DIF simulation studies frames DIF as a property of the item. For example, Kopf, Zeileis, and Strobl (2015) simulate students as belonging to either a reference or focal group. They fix the item difficulties for the reference group, ${b_j}^{\text{ref}}$, to values from a previous study. They set item difficulties for the focal group to ${b_j}^{\text{foc}} = {b_j}^{\text{ref}}$ for items without DIF and ${b_j}^{\text{foc}} = {b_j}^{\text{ref}} + 0.6$ for items with DIF, where 0.6 is the magnitude of DIF in logits. They simulate student ability ${\theta_i}^{\text{ref}} \sim N(0, 1)$ for students in the reference group and ${\theta_i}^{\text{foc}} \sim N(-1, 1)$ for students in the focal group and generate item responses according to the Rasch model. The Rasch model specifies that the probability of student $i$ responding correctly to item $j$ is

$$P(y_{ij} = 1|\theta_i, b_j) = \sigma(\theta_i - b_j) \tag{1}$$

where $\sigma(x) = \frac{e^x}{e^x+1}$ is the standard logistic function.

We find it more informative to describe simulation conditions in terms of single-dimensional items and multidimensional abilities. For example, to translate the Kopf, Zeileis, and

Strobl (2015) simulation from the DIF-as-item-property view to the DIF-as-student-property view, item difficulty is set to what was previously ${b_j}^{\text{ref}}$ for all students. Student ability is expanded to two dimensions, the target ability dimension and nuisance ability dimension. The target ability is what was previously just ability where ${\theta_i}^{\text{ref}} \sim N(0,1)$ for students in the reference group and ${\theta_i}^{\text{foc}} \sim N(-1,1)$ for students in the focal group. The nuisance ability is set to ${\eta_i}^{\text{ref}} = 0$ for the reference group and ${\eta_i}^{\text{ref}} = -1$ for the focal group. We now need to use a 2PL model where the slope on target ability $a_{1j} = 1$ for all items (consistent with the Rasch model) and the slope on nuisance ability $a_{2j} = 0.6$ for all items with DIF and $a_{2j} = 0$ otherwise. Again, 0.6 is the magnitude of DIF in logits. According to the two-dimensional 2PL model, the probability student $i$ responding correctly to item $j$ is

$$P(y_{ij} = 1|\theta_i, \eta_i, a_{1j}, a_{2j}, b_j) = \sigma(a_{1j}\theta_i + a_{2j}\eta_i - b_j). \tag{2}$$

This translation between views makes explicit that nearly all DIF simulation studies have, perhaps suboptimally, examined the unrealistic scenario in which the variance of the nuisance ability is set to 0 for all students of the same group.

If we insist on describing simulation conditions from the DIF-as-student-property view, one might wonder the following: Why not fit a multidimensional item response model which describes the data fully instead of looking for bias in a lower-than-necessary dimensional model? Camilli (1992, 144) tested this idea with the goal of a "more satisfying description of the secondary abilities." He found that the rotational indeterminacy of item response models is challenging to overcome and concluded that "a priori knowledge of the true factor structure" is necessary. It's hard to imagine how one would have such knowledge. Therefore, the best approach, which the DIF literature has nearly unanimously taken, is to fit unidimensional item response models and then look for bias manifesting in the item parameters.

# 3 DIF methods

Psychometricians have long been in search of the perfect DIF detection method. In the following three sections, we summarize DIF methods and propose new DIF methods. We use a simple simulation to demonstrate DIF methods: 10,000 reference group students and 10,000 focal group students take an eight-item test. Target ability is simulated ${\theta_i}^{\text{ref}} \sim N(0,1)$ for students in the reference group and ${\theta_i}^{\text{foc}} \sim N(-1,1)$ for students in the focal group. Nuisance ability is set to ${\eta_i}^{\text{ref}} = 0$ for the reference group and ${\eta_i}^{\text{ref}} = -1$ for the focal group. The slope on target ability is set to $a_{1j} = 1$ for all items, and the slope on nuisance ability is set to $a_{2j} = 0.5$ for the last three items (the items with DIF) and $a_{2j} = 0$ otherwise.

## 3.1 All others as anchors

Meade and Wright (2012) compared the most commonly used methods and unequivocally recommended the all-others-as-anchors (AOAA) method in which each item is tested one at a time for DIF using all other items as anchors. More specifically, testing is done using a nested model comparison between the baseline model, where all item parameters are fixed across groups, and the flexible model, where the parameters of the tested item are freed across groups. Anchor items have fixed parameters across groups and consequently are used to estimate the ability difference between groups. For example, Edelen et al. (2006) used AOAA to look for DIF between the English and Spanish versions of the 21-item Mini-Mental State Examination and found that 10 of the 21 items exhibited DIF. How can we be sure its those 10 items and not the other 11 items with DIF? Implicit in the use of AOAA is the assumption that all items not being tested do not exhibit DIF, which is, of course, impossible. Undesirably, most applications of DIF methods and many simulation studies do not make explicit the assumptions of the DIF method Strobl et al. (2018). In this way, psychometricians might benefit from adopting economists' habit of explicitly stating assumptions and debating their plausibility.

Researchers have noticed the circular logic of AOAA, but have mostly described it indirectly by pointing out inflated Type I errors in simulation studies (Stark, Chernyshenko, and Drasgow 2006). A simple thought experiment illustrates how AOAA fails: Imagine a test with a

sufficiently large number of students and three items where the first item has DIF, and the other two do not. Using AOAA, all items test positive for DIF. The last two items incorrectly test positive because including the first item in the anchor set causes the group ability difference to be misestimated. This phenomenon of items with real DIF inducing the appearance of DIF in other items was only indirectly discussed in the literature until Andrich and Hagquist (2012) coined the term "artificial DIF."

One way to attempt to counter artificial DIF is with purification with iterative anchor selection. For example, (Drasgow 1987) started with AOAA, removed items displaying DIF from the anchor set, then continued iteratively to test the remaining items for DIF until no more items tested positively. (Kopf, Zeileis, and Strobl 2015) named this technique Iterative-backward-AOAA with "backward" (as in reverse, not incorrect) referring to beginning with the assumption that all items are DIF free. We find this name confusing and instead refer to this method as all-others-as-anchors-all-significiant (AOAA-AS) where appending all-significant indicates that anchor selection is done iteratively with all items that show statistical significance for DIF are removed from the anchor set. AOAA-AS might seem like an improvement, but what does one do when all items display DIF in the first AOAA stage? With a sufficient sample size, this will necessarily be the case. Kopf, Zeileis, and Strobl (2015) encountered precisely this problem in their simulation study and chose to select a single anchor item randomly. Woods (2009) suggested a more straightforward, one-step method which uses AOAA and selects the, say, four items that exhibit the least amount of DIF. It's unclear if one should proceed if those four items display DIF themselves.

We propose an extension of these methods, all-others-as-anchors-one-at-a-time (AOAA-OAAT), which, to our knowledge (and surprise), has not previously been proposed. AOAA-OAT starts with AOAA, and the single item exhibiting the most DIF is removed from the anchor set. Each item in the anchor set is then tested for DIF again (with items outside of the anchor set allowed to have free parameters across groups in both the baseline and flexible model). This process continues until no new items display DIF. AOAA-OAAT and AOAA-AS are similar in that they are both iterative; the difference is that AOAA-OAAT takes the more conservative approach of removing only the item exhibiting the most DIF. Applying AOAA-OAT to our thought experiment

demonstrates its logic. The initial AOAA removes the real DIF item from the anchor set because it exhibits the most DIF. In the next step, both of the other items test negative for DIF, and we arrive at the correct conclusion. AOAA-OAT has two assumptions: First, that at least one item does not have DIF, and second, that the majority of items do not have DIF. Table X summarizes these methods. In wrapping one's mind around these methods it's useful to remember that AOAA is not an iterative procedure, while the methods with a dash, AOAA-OAT and AAOA-AS, are iterative procedures with AOAA-OAT being the most conservative.

ALSO MAKE A TABLE OF SIMULATION RESULTS

| Name | Abbreviation | Description | Assumptions | Literature |
|---|---|---|---|---|
| All-others-as-anchors | AOAA | 87837 | 787 | 1 |
| All-others-as-anchors-all-significant | AOAA-AS | 78 | 5415 | 1 |
| All-others-as-anchors-one-at-a-time | AOAA-OAAT | 778 | 7507 | 1 |
| 4 | 545 | 18744 | 7560 | 1 |
| 5 | 88 | 788 | 6344 | 1 |

### 3.2 Presenting information to the analyst

DIF methods, in general, seem to be designed to automatically detect items with DIF without any intervention from a human. We refer to DIF methods that do not use an analyst's judgement as "algorithmic DIF methods." On the other hand, it might make sense to present information to the analyst in a way that empowers them to make decisions in some cases. We propose a technique, the "equal means logit graph" (EMLG), where both group means are to 0 so that differences in performance manifest exclusively in the item parameters. Relatedly, Pohl, Stets, and Carstensen (2017) fit a model with both groups means set to 0 in a pedagogical example, and Talbot III (2013) fixed both pre-test and post-test means to 0 in order to estimate item-specific learning gains.

We fit a unidimensional Rasch, with both group means fixed to 0, to the simulated item response data. We estimate the item parameter covariance matrix using Oakes' identity (Chalmers 2018). Multiple imputations (MI) (Yang, Hansen, and Cai 2012) are used to estimate the distribu-

tion of the difference in difficulty, $b_j^{\text{foc}} - b_j^{\text{ref}}$, for each item. Figure 1 shows these distributions in the EMLG.

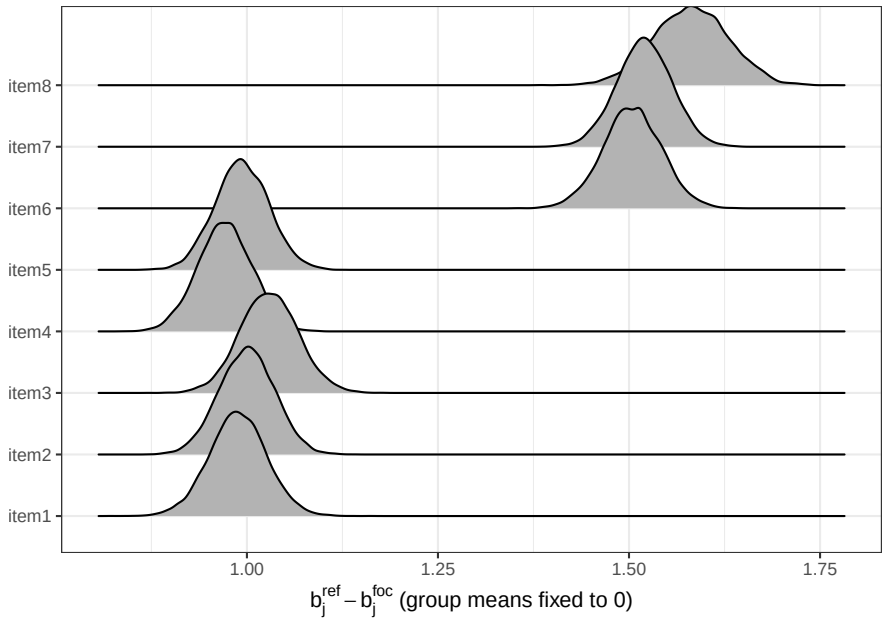


Figure 1: Equal means logit graph (EMLG)

It cannot be stated strongly enough that the EMLG contains all possible information about the difference in group performance. The challenge, then, is to identify the difference in group ability. The analyst might assume that — because there are five items where the reference group outperforms the focal group by 1 logit and only three items where the difference is 1.5 logits — items 1-5 are unbiased and should be used as anchor items. At this point, the analyst re-fits the model with the anchor items having fixed parameters across groups so that they are used to estimate the difference in group ability. The parameters of the other items are free to vary across groups, and, indeed, the new model finds that the item easiness difference, $b_j^{\text{ref}} - b_j^{\text{foc}}$, is about 0.5 logits for each of these items.

One way to show this:

```
## Picking joint bandwidth of 0.00548
```

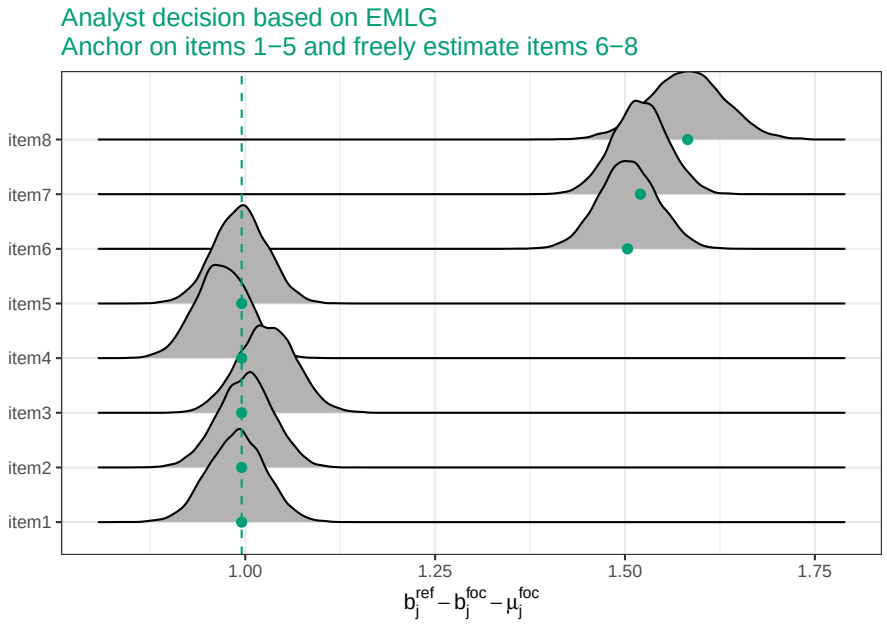


Figure 2: fix this

Another way to show this:

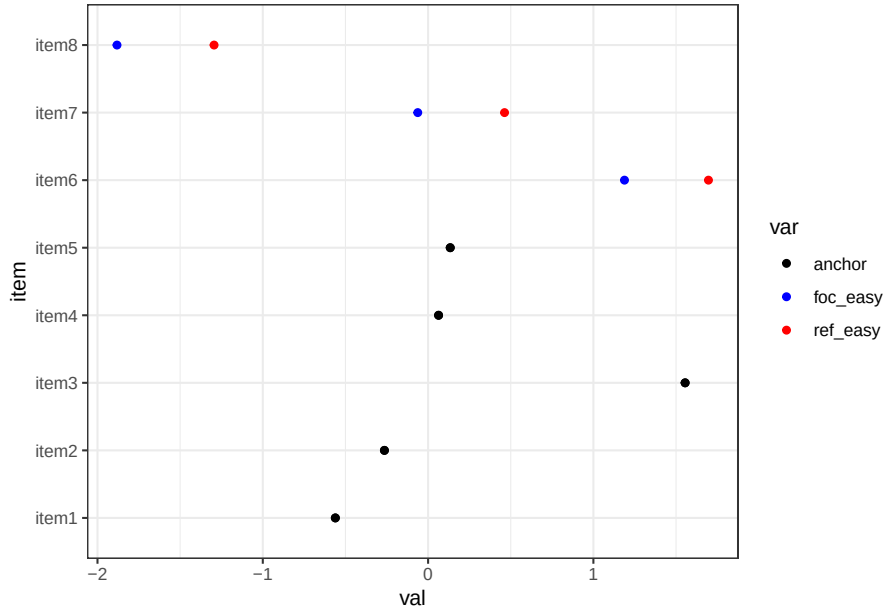


Figure 3: fix this

One of our key concerns with algorithmic DIF methods is that they can lull the analyst

into a false sense of security. Too often, the analyst chooses a method, implements it, and then proceeds as if the method certainly identified the correct anchor items. The EMLG combats this concern by presenting all information clearly to the analyst. In the previous example, the analyst is may be weary of their results having seen how arbitrary it was to conclude that the first five and not the last three items are unbiased. Even in cases where algorithmic DIF methods are used, the analyst should use the EMLG to give them a sense of their item response data.

### 3.3 Clustering

Bechger and Maris (2015) proposed choosing anchor items by identifying clusters of items that function similarly and then choosing one of those clusters to be the "anchor cluster." They pointed out that one way around the unidentifiability issue of item response models is to arbitrarily set an item easiness parameter to 0. For each group, the relative difficulties for each pair of items are stored in the matrix $\mathbf{R}^{\text{ref}}$ with entries ${R_{xy}}^{\text{ref}} = {b_x}^{\text{ref}} - {b_y}^{\text{ref}}$. The ultimate matrix of interest is $\Delta\mathbf{R} \equiv \mathbf{R}^{\text{ref}} - \mathbf{R}^{\text{foc}}$ which is the "differences between groups in the pairwise differences in difficulty between items" [p. 323].

The general idea of identifying clusters of items is interesting. However, their approach is needlessly complicated and they did not give specific suggestions on how to move from $\Delta\mathbf{R}$ to a set of anchor items. Pohl, Stets, and Carstensen (2017) extended their work by proposing an approach for identifying anchor items. $\Delta\mathbf{R}$ is skew-symmetric and of rank 1 so that all information is contained in a single row or column. Accordingly, they use k-means clustering on just the first column of $\Delta\mathbf{R}$ where the number of clusters, k, is chosen by minimizing BIC. They suggest using a combination of cluster size, cluster homogeneity, and cluster parameter precision to choose which of the clusters is the anchor cluster. However, in their simulation study, they find that BIC identifies only a single cluster, so they use all items as anchors.

We propose a new cluster-based approach which we refer to as equal means clustering (EMC). Instead of working with an arbitrary column from $\Delta\mathbf{R}$, we work with the vector of differences in item easiness, ${b_j}^{\text{ref}} - {b_j}^{\text{foc}}$, when both groups have their means fixed to 0 as is the case in the EMLG. Instead of choosing k with BIC, we use the gap statistic method reccommended

by Hastie, Tibshirani, and Walther (2001). The largest cluster is chosen as the anchor cluster. If there is a tie for the largest cluster, the cluster with the lowest standard deviation of item easiness differences is selected.

Using this method, we find that _____.

```
## # A tibble: 8 x 6
##   item  a_ref_easy b_foc_easy difference_in_easy cluster anchor
##   <chr>      <dbl>      <dbl>              <dbl>   <int> <lgl>
## 1 item1    -0.564      -1.55               0.988       1 TRUE
## 2 item2    -0.262      -1.26               1.00        1 TRUE
## 3 item3     1.57        0.546              1.03        1 TRUE
## 4 item4     0.0521     -0.918              0.970       1 TRUE
## 5 item5     0.134      -0.860              0.993       1 TRUE
## 6 item6     1.70        0.193              1.50        2 FALSE
## 7 item7     0.463      -1.06               1.52        2 FALSE
## 8 item8    -1.30       -2.88               1.58        2 FALSE
```

## 3.4 Anchor points

All of the previously discussed DIF methods select a set of anchor items, whether that selection is done by an algorithm or the analyst. The difference in group abilities is then estimated using exclusively the anchor items. An alternative strategy is to select an anchor point, which essentially specifies the difference in group abilities directly. Anchor point methods have the advantage of not assuming that any particular item is DIF-free and therefore, all items can be tested for DIF.

### 3.4.1 Gini index

Strobl et al. (2018) suggest using the Gini index (1912) for selecting anchor points. The Gini index is typically used to measure the inequality of wealth distribution in a country. South Africa typically has the highest Gini index of all measured countries, meaning that it is the country with the most unequal wealth distribution Chitiga, Sekyere, and Tsoanamatsie (2015). In general, the

Gini index "takes high values if, for example, a small minority of persons has a lot of wealth while the vast majority has very little" (Strobl et al. 2018, 7).

The Gini index is maximized to select the anchor point. The intuition is that we want to find the anchor point such that a small minority of items have bias while the vast majority do not. Denote the easiness difference between the two groups when both group means are fixed to zero as $d_j = b_j^{\text{ref}} - b_j^{\text{foc}}$, a function that calculates the Gini index from a vector of non-negative elements as $G(\mathbf{x})$, and the mean of the focal group ability when the reference group ability is fixed to zero (i.e. the anchor point) as $\Theta$. The anchor point is set to

$$\arg\max_{\Theta} G(\mathbf{d} + \Theta)$$

where $\Theta \in (-\infty, \infty)$ and $\Theta$ is added to each element of $\mathbf{d}$.

Figure 4 shows the gini coefficient at a variety of possible values for $\Theta$. We find that the anchor point is -0.99.

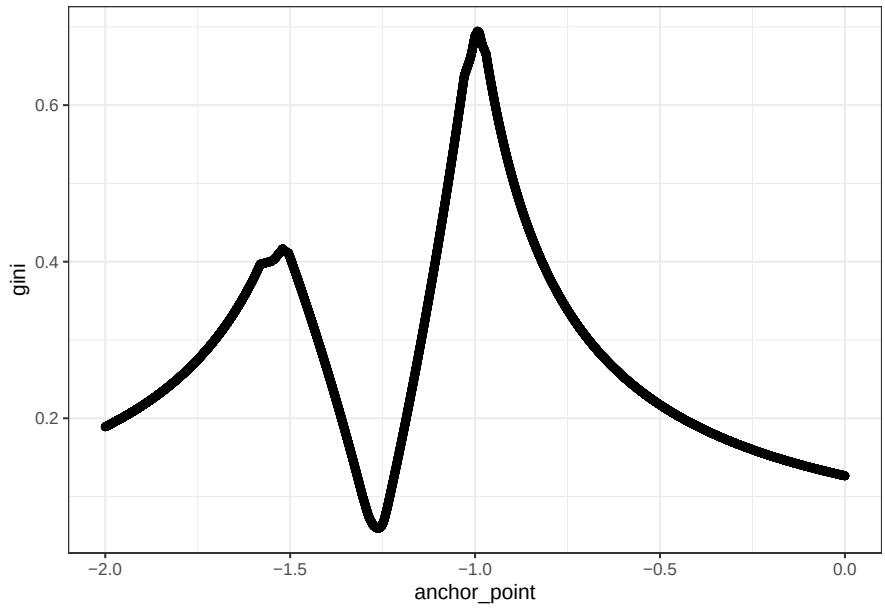


Figure 4: always need a capt

### 3.4.2 Minimizing between curves

Imagine multigroup item response data where $\Theta^{\text{foc}} = \Theta^{\text{ref}}$ and $d_j = 0 \forall j$ so that there is no DIF. The fundamental identification problem of DIF in multipgroup IRT models is that there are an infinite number of models with the same likelihood from which to choose. For example, we could correctly assume that the focal group has the same ability as the reference group and fix $\Theta^{\text{foc}} = 0$. The model would then estimate $d_j \approx 0 \forall j$, and we would correctly conclude the groups have the same ability and there is no DIF. Alternatively, we could assume that the focal group has very high ability and fix $\Theta^{\text{foc}} = 3$. The model would then compensate by finding $d_j \approx -3 \forall j$, and we would incorrectly conclude that the the focal group is high ability but each item contains DIF against the focal group. Both of these models have the same likelihood, so how should we choose which model to believe? This is the fundamental identification problem.

We put forth a method we call minimizing between curves (MBC) which chooses the model with least amount of total DIF on the test. "Curves" refers to the item characteristic curve which maps a student's ability to their probability of correct response. The amount of DIF on each item is calculated as

$$\text{DIF}_j = \int |\Pr(y_j = 1|\theta, b_j^{\text{ref}}) - \Pr(y_j = 1|\theta, b_j^{\text{foc}})| g(\theta) d\theta$$

where $g(\theta)$ is a weighting function such that $\int g(\theta) d\theta = 1$. The total DIF on the test, then, is Total DIF $= \sum_j \text{DIF}_j$. Different choices for $\Theta^{\text{foc}}$ lead to different $\mathbf{b}^{\text{foc}}$, and we select the $\Theta^{\text{foc}}$ that minimizes Total DIF. The idea for considering the difference between item characteristic curves dates back to Stocking and Lord (1983). More recently, Chalmers, Counsell, and Flora (2016) use the difference between test characteristic curves weighted by $g(\theta)$ as a measure of differential test functioning (DTF).

The selection of $g(\theta)$ is important and determines where on the ability spectrum to consider DIF. (2016) does not discuss how to choose $g(\theta)$ and in practice uses $g(\theta)$ uniform for $-6 \leq \theta \leq 6$, which seems like an odd choice. It might seem intuitive to choose $g(\theta) \sim N(0, 1)$ because $\mu^{\text{ref}} = 0$,

but this choice doesn't take into account the ability distribution of the focal group. If $\mu^{\text{foc}} = 3$, wouldn't we also want to prioritize high $\theta$ values? We set $g(\theta)$ to be the average of the reference and focal group ability distributions:

$$g(\theta) = \frac{N(\mu^{\text{ref}}, \sigma^{\text{ref}^2}) + N(\mu^{\text{foc}}, \sigma^{\text{foc}^2})}{2}.$$

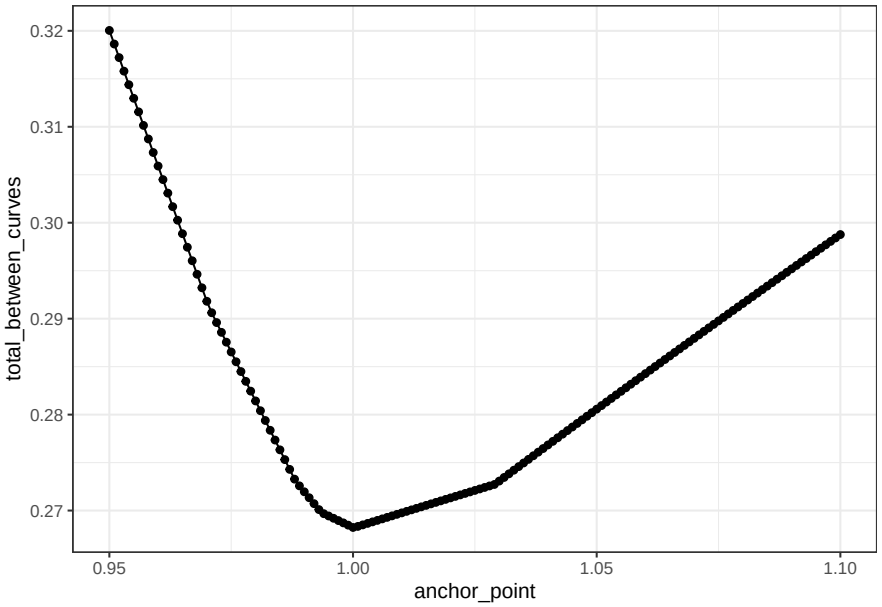


Figure 5: always need a capt

### 3.5 Bounds for DIF

### 3.6 Summary

## 4 Next steps

## 5 Simulate and then walk through as the analyst would with each method

## 6 Mike Frank example

We know that this test contains DIF. Which items depends on ability difference which we can be sure of. We can, however, quantify the amount of DIF on this test. We can take the SD or we can assume between and calculate chalmers statistic for each. We might not know for sure the ability difference but we can quantify the potential for DIF. For each we calculate the SD across item parameters.

Something with DTF?

Do a simulation where the second dimension varies - a more real testing ground

Example with real data perhaps the Mike Frank data

Summer learning loss as applied DIF?

# 7 Notes

- adjust to all be easiness including multidimensional talk

# 8 References


Ackerman, Terry A. 1992. "A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective." *Journal of Educational Measurement* 29 (1): 67–91.

Andrich, David, and Curt Hagquist. 2012. "Real and Artificial Differential Item Functioning." *Journal of Educational and Behavioral Statistics* 37 (3): 387–416.

Bechger, Timo M, and Gunter Maris. 2015. "A Statistical Test for Differential Item Pair Functioning." *Psychometrika* 80 (2): 317–40.

Camilli, Gregory. 1992. "A Conceptual Analysis of Differential Item Functioning in Terms of a Multidimensional Item Response Model." *Applied Psychological Measurement* 16 (2): 129–47.

Chalmers, R Philip. 2018. "Numerical Approximation of the Observed Information Matrix with Oakes' Identity." *British Journal of Mathematical and Statistical Psychology* 71 (3): 415–36.

Chalmers, R Philip, Alyssa Counsell, and David B Flora. 2016. "It Might Not Make a Big Dif: Improved Differential Test Functioning Statistics That Account for Sampling Variability." *Educational and Psychological Measurement* 76 (1): 114–40.

Chitiga, Margaret, E Sekyere, and N Tsoanamatsie. 2015. "Income Inequality and Limitations of the Gini Index: The Case of South Africa." *Human Sciences Research Council (HSRC), Available at: Http://Www. Hsrc. Ac. Za/En/Review/Hsrc-Review-November-2014/Limitations-of-Gini-Index, Site Accessed* 2.

Drasgow, Fritz. 1987. "Study of the Measurement Bias of Two Standardized Psychological

Tests." *Journal of Applied Psychology* 72 (1): 19.

Edelen, Maria Orlando, David Thissen, Jeanne A Teresi, Marjorie Kleinman, and Katja Ocepek-Welikson. 2006. "Identification of Differential Item Functioning Using Item Response Theory and the Likelihood-Based Model Comparison Approach: Application to the Mini-Mental State Examination." *Medical Care*, S134–S142.

Gini, Corrado. 1912. "Variabilità E Mutabilità (Variability and Mutability)." *Tipografia Di Paolo Cuppini, Bologna, Italy*, 156.

Hastie, Trevor, Robert Tibshirani, and Guenther Walther. 2001. "Estimating the Number of Data Clusters via the Gap Statistic." *J Roy Stat Soc B* 63: 411–23.

Kopf, Julia, Achim Zeileis, and Carolin Strobl. 2015. "A Framework for Anchor Methods and an Iterative Forward Approach for Dif Detection." *Applied Psychological Measurement* 39 (2): 83–103.

Meade, Adam W, and Natalie A Wright. 2012. "Solving the Measurement Invariance Anchor Item Problem in Item Response Theory." *Journal of Applied Psychology* 97 (5): 1016.

Pohl, Steffi, Eric Stets, and Claus H Carstensen. 2017. "CluStEr-baSEd anCHor Item Identification and Selection." NEPS Working Paper.

Stark, Stephen, Oleksandr S Chernyshenko, and Fritz Drasgow. 2006. "Detecting Differential Item Functioning with Confirmatory Factor Analysis and Item Response Theory: Toward a Unified Strategy." *Journal of Applied Psychology* 91 (6): 1292.

Stocking, Martha L, and Frederic M Lord. 1983. "Developing a Common Metric in Item Response Theory." *Applied Psychological Measurement* 7 (2): 201–10.

Strobl, Carolin, Julia Kopf, Raphael Hartmann, and Achim Zeileis. 2018. "Anchor Point Selection: An Approach for Anchoring Without Anchor Items." Working Papers in Economics; Statistics.

Talbot III, Robert M. 2013. "Taking an Item-Level Approach to Measuring Change with the Force and Motion Conceptual Evaluation: An Application of Item Response Theory." *School Science and Mathematics* 113 (7): 356–65.

Woods, Carol M. 2009. "Empirical Selection of Anchors for Tests of Differential Item Functioning." *Applied Psychological Measurement* 33 (1): 42–57.

Yang, Ji Seung, Mark Hansen, and Li Cai. 2012. "Characterizing Sources of Uncertainty in Item Response Theory Scale Scores." *Educational and Psychological Measurement* 72 (2): 264–90.