Measurement invariance

Ben Stenhaug & Ben Domingue
Stanford University

Contents

1 Intro

2 Intro

Inspired by Camilli (1992), we view differential item functioning (DIF) as bias towards a group of students that manifests when we impose an item response model with too few ability dimensions. Viewed in this way, 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 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 will result in students with low verbal ability being given a score systematically lower than their true math ability; this is 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 the reference or the focal group. The item difficulties for the reference group, b_j^{ref}, are fixed. The item difficulties for the focal group are $b_j^{\text{foc}} = b_j^{\text{ref}}$ 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. Student ability is simulated $\theta_i^{\text{ref}} \sim N(0, 1)$ for students in the reference group and $\theta_i^{\text{foc}} \sim N(0, 1)$ for students in the focal group. The item responses are generated according to the Rasch model. The Rasch model specifies that the probability of student i responding correctly to item j is

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

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

This simulation study can be translated from the DIF-as-item-property view to the DIF-as-student-property view. Item difficulty is set to what was previously b_j^{ref} for all students. We must then expand student ability 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 $\theta_i^{\text{foc}} \sim N(0, 1)$ 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 multidimensional 2PL model, the probability student i responding correctly to item j is

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

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.

We insist on simulating from the dif-as-asdf view. When looking for DIF, it is still useful to fit models with a single ability dimension to avoid interpreting a rotationally indetermined latent space.

There have beeen a variety of DIF methods proposed tested and used. The literature seems to be in search of a method that can perfectly detect items with DIF while paradoxically admitting that it is sometimes impossible to disentangle group ability differences and DIF. For example, Meade 2012 concluded that AOAA where asdkfhjasdklf was best. AOAA has been used in many studies for example that one paper where they report results with certainty. These studies often don't make explicit the assumptions of DIF methods invoked. It's easy to show that AOAA is imperfect: Imagine a test with 3 items. Each item will test positive for DIF. a phenomenon that went indirectly address in the literature until andrich at all named it artificial DIF.

We advocate for methods that present information to the analyst where the analyst makes decisions as opposed to methods which automate the detection and subsequently the removal of items with DIF as is frequently the case (e.g. in PISA). The most obvious method is to at first punt on disentangling between ability and bias and instead set both group means to zero and estimate the item parameters separately for each group. The difference in item parameters, then, includes both the difference in ability and bias. To our knowledge, this method has not been suggested in the literature with the exception of pohl who uses it in a pedagogical example and tashjkdf who uses it pre to post. For example, consider a didactic scenario in which an eight item exam where the focal group and reference group. Where the difference in ability between groups is one logit. And additionally, the first two items have a one logit bias against the students who. We fit a model then fix the means. We then sample as leah describes (look at how she describes it) then we get this plot. It cannot be said strongly enough that this plot contains all of the information about the combination of ability and bias that is available in the data. It is then up to the analyst and/or method to choos what's known anchor point, which is essentially the difference in abilities with remaining differences being attributed to bias.

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.

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.

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.