COLLEGE OF EDUCATION AND HUMAN ECOLOGY: Quantitative Research, Evaluation & Measurement Program

Item Quality for Cognitive Assessments in Low-to-Middle Income Countries: Evidence from the Ethiopia Young Lives Data.

Winifred Graham Wilberforce & Ann A O'Connell, Ed.D.

INTRODUCTION

Across the globe, testing agencies have a common goal of creating items that maintain measurement invariance and are thus free of group characteristic biases (Khorramdel et al., 2020). Even though test agencies' efforts to consult with local talents to create cultural and gender relevant items in LMIC's, there is evidence that some of these items are biased and may be unfair to some students based on their economic status, gender, and location. We examine the cultural and gender relevant items included on the cognitive assessments for the Young Lives school effectiveness survey data from Ethiopia (2016/2017) (Boyden et al., 2016; Rossiter et al, 2017).

AIM

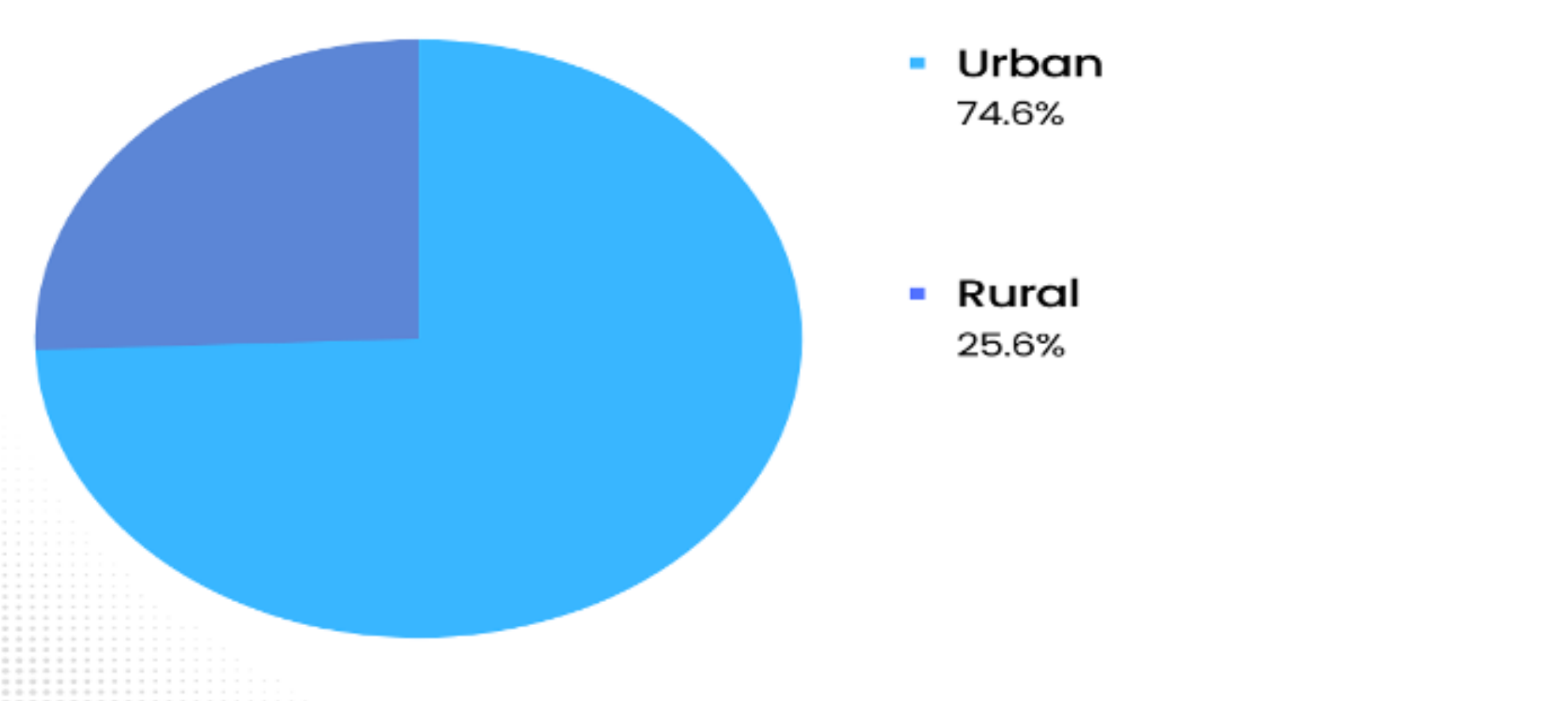
The purpose of this paper is to examine how gender and location (rural vs urban) may impact item quality for assessments administered in low-to-middle income countries (LMIC's). This study examines :

1. The item fit statistics and item targeting of the YL cognitive assessments for Ethiopia under the Rasch model.
2. Differential item functioning analysis (DIF) by gender and location for the cognitive assessment data.

Descriptive Statistics

- We use data from the first wave which was collected in 2016.
- The survey was completed by a total of N = 12,182 children in 7th and 8th grade (M age = 14.36) from 63 schools.
- About half the sample identified as boys (48.5 %) and the other half (48.9%) as girls, with 2.6% missing on this variable.

A PIE CHART SHOWING % URBAN VS RURAL



METHODS

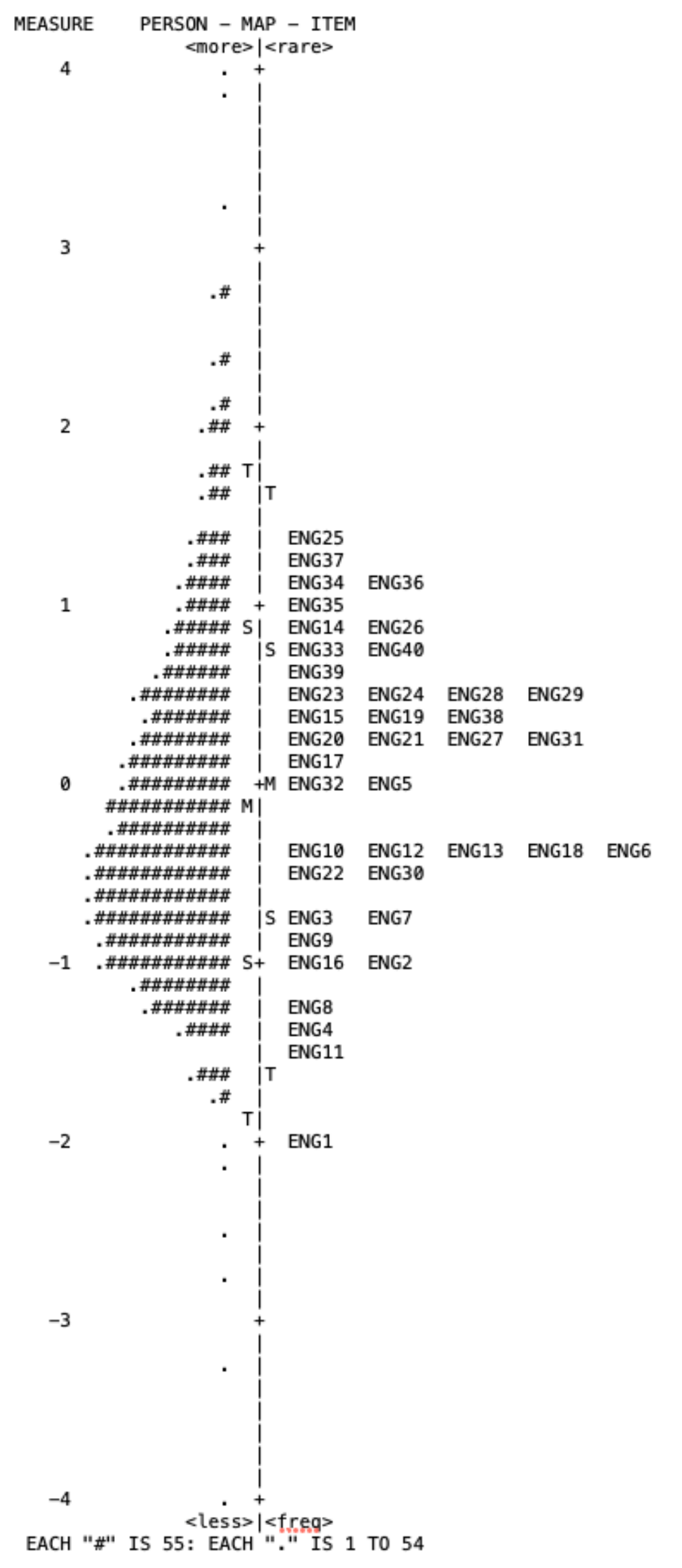
- ❖ Item responses to the cognitive assessments were analyzed using the **Rasch model**. This is consistent with the hypothesized unidimensional nature of this data (Rossiter et al., 2017; Linacre, 2022).
- ❖ Item fit statistics were examined to see how well the data fit the Rasch model, and Wright maps were utilized to clarify item targeting.
- ❖ We conducted differential item functioning **(DIF)** analysis to explore the possibility of item-level performance differences (Walker, 2011) based on gender and location.

RESULTS

Test-Item Targeting

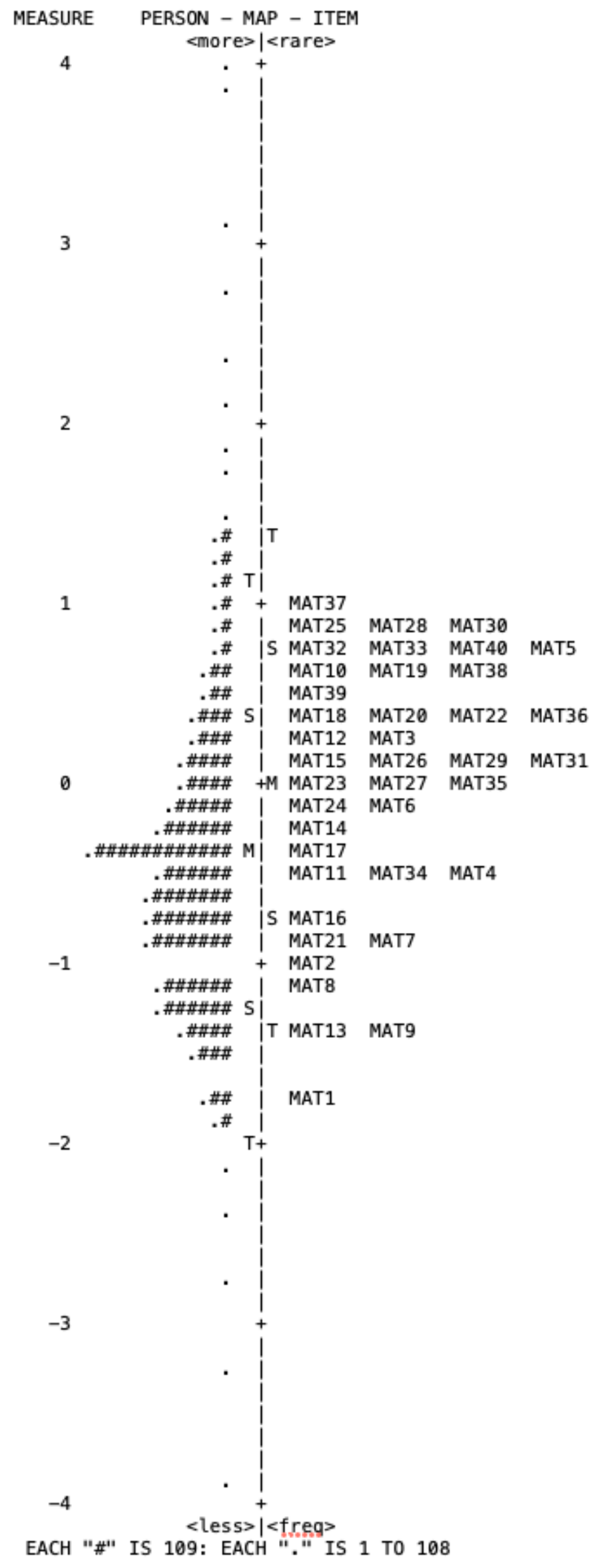
The person-distributions for both the Math & English assessments showed the ability of students on the same measurement scale as the item difficulty; this pattern was acceptable but suggests that more difficult items should be added to the assessment for highly proficient students at the upper part of the scale.

Wright maps showing the person distribution with item labels



```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4             .  +
                  .  |
                     |
                     |
                     |
                     |
                  .  |
                     |
    3                +
                     |
                 .#  |
                     |
                     |
                 .#  |
                     |
                 .#  |
    2           .##  +
                     |
                .## T|
                .##  |T
                     |
               .###  |  ENG25
               .###  |  ENG37
              .####  |  ENG34  ENG36
    1         .####  +  ENG35
             .##### S|  ENG14  ENG26
             .#####  |S ENG33  ENG40
            .######  |  ENG39
          .########  |  ENG23  ENG24  ENG28  ENG29
           .#######  |  ENG15  ENG19  ENG38
          .########  |  ENG20  ENG21  ENG27  ENG31
         .#########  |  ENG17
    0    .#########  +M ENG32  ENG5
        ########### M|
        .##########  |
      .############  |  ENG10  ENG12  ENG13  ENG18  ENG6
      .############  |  ENG22  ENG30
      .############  |
      .############  |S ENG3   ENG7
       .###########  |  ENG9
   -1  .########### S+  ENG16  ENG2
          .########  |
           .#######  |  ENG8
              .####  |  ENG4
                     |  ENG11
               .###  |T
                 .#  |
                    T|
   -2             .  +  ENG1
                  .  |
                     |
                     |
                  .  |
                     |
                  .  |
                     |
   -3                +
                     |
                  .  |
                     |
                     |
                     |
                     |
                     |
   -4             .  +
               <less>|<freq>
EACH "#" IS 55: EACH "." IS 1 TO 54
```

English



```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4             .  +
                  .  |
                     |
                     |
                     |
                     |
                     |
                  .  |
    3                +
                     |
                  .  |
                     |
                     |
                  .  |
                     |
                  .  |
    2                +
                  .  |
                  .  |
                     |
                  .  |
                 .#  |T
                 .#  |
                 .# T|
    1            .#  +  MAT37
                 .#  |  MAT25  MAT28  MAT30
                 .#  |S MAT32  MAT33  MAT40  MAT5
                .##  |  MAT10  MAT19  MAT38
                .##  |  MAT39
               .### S|  MAT18  MAT20  MAT22  MAT36
               .###  |  MAT12  MAT3
              .####  |  MAT15  MAT26  MAT29  MAT31
    0         .####  +M MAT23  MAT27  MAT35
             .#####  |  MAT24  MAT6
            .######  |  MAT14
      .############ M|  MAT17
            .######  |  MAT11  MAT34  MAT4
           .#######  |
           .#######  |S MAT16
           .#######  |  MAT21  MAT7
   -1                +  MAT2
            .######  |  MAT8
            .###### S|
              .####  |T MAT13  MAT9
               .###  |
                     |
                .##  |  MAT1
                 .#  |
   -2               T+
                  .  |
                     |
                  .  |
                     |
                     |
                  .  |
                     |
   -3                +
                     |
                  .  |
                     |
                     |
                     |
                     |
                  .  |
   -4                +
               <less>|<freq>
EACH "#" IS 109: EACH "." IS 1 TO 108
```

Math

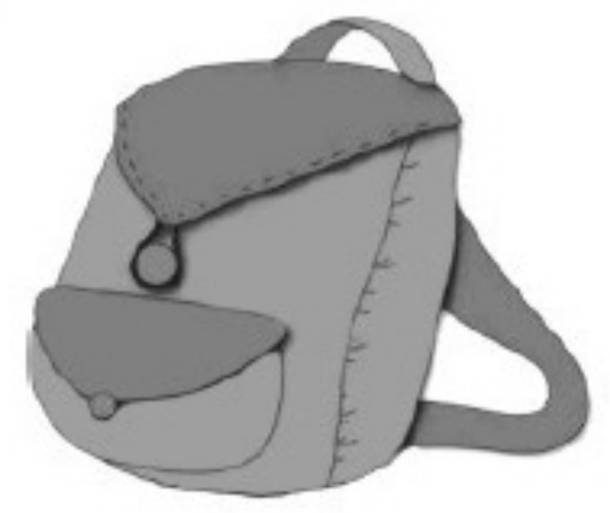
Differential Item Functioning

For both English and Mathematics items, we used the Rasch-Welsh Test and a significance level of 0.05 based on the absolute value of the DIF contrast (Linacre, 2022).

- Overall, there was no support for any individual items that favored boys or girls.
- However, for location (urban versus rural), **15%** of the mathematics items showed moderate to high DIF (|DIF| ≥ 0.64 logits) and slight to moderate DIF (|DIF| ≥ 0.43 logits), (Zwick, Thayer & Lewis, 1999).
- The English test had **17.5%** of items showing moderate to high DIF. For example, the first test item which asked students to identify a school bag, had a DIF contrast of .82. This item was harder for students in rural schools.

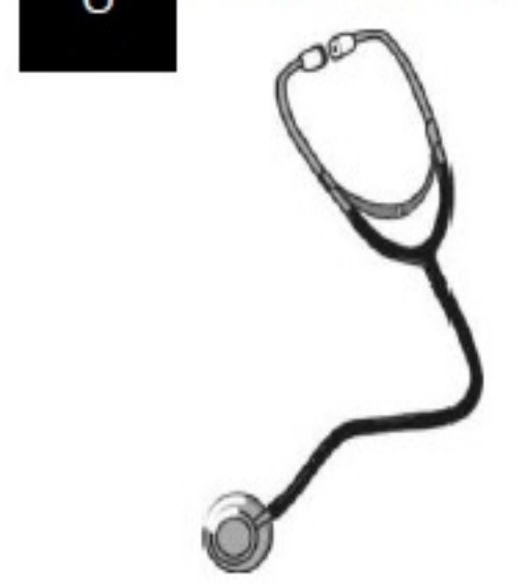
Examples of some English Items that exhibited location DIF

1 Select the correct word for the picture.



A Bat
B Bag
C Pad
D Box

8 Who among the following is most likely to use this object for their work?



A Lawyer
B Doctor
C Farmer
D Carpenter

A picture of a school in rural Ethiopia

Source: Ark Republic

CONCLUSIONS

Based on the literature from LMICs, most large-scale test developers seem to focus on creating unbiased questions to avoid gender DIF than they do for location. Thus, it is no surprise that this same trend is seen within these YL data where the gender-based differences were negligible in terms of DIF but potentially biased items are indicated when location is considered.

E-mail **Wilberforce.2@osu.edu** with suggestions.

BIBLIOGRAPHY

1. Boyden, J., Woldehanna, T., Galab, S., Sanchez, A., Penny, M., Duc, L.T. (2016). Young Lives: An International Study of Childhood Poverty: Round 4, 2013-2014. [data collection]. *UK Data Service.* SN: 7931, http://dx.doi.org/10.5255/UKDA-SN-7931-1.
2. Khorramdel. L., Pokropek, A., & Rijn, P.(2020). Establishing comparability and measurement invariance in large-scale assessments, Part II: Old questions, new challenges, and possible solutions. *Psychological Test and Assessment Modeling. 62. 139-145.*
3. Linacre, J.M. (2022). Winsteps Rasch measurement computer program User's Guide. Version 5.2.0. Portland, Oregon: Winsteps.com
4. Rossiter, J., Azubuike, O., & Rolleston, C. (2017). Young Lives School Survey, 2016-2017: Evidence from Ethiopia. Young Lives Country Report, Oxford. Young Lives.
5. Walker, C. M. (2011). What's the DIF? Why Differential Item Functioning Analyses Are an Important Part of Instrument Development and Validation. *Journal of Psychoeducational Assessment, 29*(4), 364–376. https://doi.org/10.1177/0734282911406666

ACKNOWLEDGEMENT

The data used in this poster come from Young Lives, a 20-year study of childhood poverty and transitions to adulthood in Ethiopia, India, Peru and Vietnam (www.younglives.org.uk). Young Lives is funded by UK aid from the Foreign, Commonwealth & Development Office and a number of further funders. The views expressed here are those of the author(s). They are not necessarily those of Young Lives, the University of Oxford, FCDO or other funders.'

THE OHIO STATE UNIVERSITY