Northwestern

Examining Differential Item Functioning (DIF) by Educational Attainment on Measures of Depression

Robert S. Chapman, BA, Bayley J. Taple, MS, James W. Griffith, PhD, Rylee Brower, MA, and Benjamin D. Schalet, PhD

Northwestern University Feinberg School of Medicine

Speaker: Bayley J. Taple, MS

Northwestern Medicine®
Feinberg School of Medicine

Background

- Depression commonly occurs and is a public health concern
  - 29.9% lifetime risk for US adults (Kessler et al., 2012)
- Individuals with lower levels of education face, on average, more social and economic challenges
- Individuals with lower educational attainment and depression education are vulnerable on multiple levels
- Effective treatment for this population begins with accurate and appropriate assessment tools

Differential Item Functioning (DIF)

DIF exists when there is a difference in the strength of the relationship between a questionnaire item and a concept across groups

Previous Research on DIF by Education

- Most DIF by education studies have been done on the Mini-Mental State Exam and have found significant DIF on 4-5 items
  - Mini-Mental State Exam shows DIF by education
  - (Murden et al., 1991; Ramirez et al., 2006; Jones & Gallo, 2002; Crane et al., 2006)

- Little to no DIF by education found on Mattis Dementia Rating Scale
  - 4 items problematic, only 1 showed any large effect (digit span backwards)
  - (Teresi et al., 2000)

- PROMIS Depression scale showed DIF on 3 items
  - DIF by gender, age, and education
  - Small effect overall, large effect for some individuals
  - (Teresi et al., 2009)

Study Method

- Secondary data analysis of data from a study linking PROMIS to legacy measures of depression
- 3 internet panel samples (for details see Choi et al., 2014)
  - **PROMIS 1 Wave 1**: recruited by Polimetrix online
  - **NIH Toolbox calibration**: recruited by Greenfield Online
  - **PROsetta Stone**: recruited by Op4G

Participants

- NIH Toolbox (*N* = 748)
  - Age: 18 – 92 years, *mean* = 47
  - 56% female
  - 78% White, 9% Black
- PROMIS 1 Wave 1 (*N* = 744)
  - Age: 18 – 88 years, *mean* = 51
  - 52% female
  - 80% White, 10% Black
- PROsetta Stone (*N* = 1104)
  - Age: 18 – 88 years, *mean* = 46
  - 52% female
  - 72% White, 11% Black

	NIH Toolbox N = 748	PROMIS 1 Wave 1 N = 744	PROsetta Stone N = 1104
Age, Mean (SD)	47.2 (15.2)	51.0 (18.8)	46.3 (17.5)
Gender, N (%)			
Male	328 (44)	357 (48)	528 (48)
Female	420 (56)	386 (52)	576 (52)
Missing	0 (0)	1 (<1)	0 (0)
Education, N (%)			
High school or less	205 (27)	171 (23)	465 (42)
Some college or technical school	326 (44)	334 (45)	305 (28)
College graduate or above	217 (29)	239 (32)	334 (30)
Missing	0 (0)	0 (0)	0 (0)
Ethnicity, N (%)			
Non-Hispanic or Non-Latinx	634 (85)	670 (90)	929 (84)
Hispanic or Latinx	114 (15)	70 (9)	175 (16)
Missing	0 (0)	4 (1)	0 (0)
Race, N (%)			
White or Caucasian	585 (78)	592 (80)	793 (72)
Black or African American	67 (9)	72 (10)	124 (11)
Asian	16 (2)	3 (<1)	58 (5)
American Indian or Alaska Native	13 (2)	6 (<1)	7 (1)
Native Hawaiian or Pacific Islander	5 (1)	0 (0)	7 (1)
Biracial or Multiracial	16 (2)	71 (10)	33 (3)
Other	46 (6)	N/A	82 (7)
Missing	0 (0)	0 (0)	0 (0)

Measures

	PROMIS Depression (Cella et al., 2010)	BDI-II (Beck et al., 1996)	PHQ-9 (Kroenke et al., 2001)	CES-D (Radloff, 1977)
NIH Toolbox Calibration	✓ (20 items)		✓	✓
PROMIS 1 Wave 1	✓ (28 items)			✓
PROsetta Stone	✓ (15 items)	✓		

Hypothesis

DIF will be observed across levels of educational attainment for complex items on legacy measures of depression (BDI-II, CES-D, PHQ-9)

Analytical Strategy

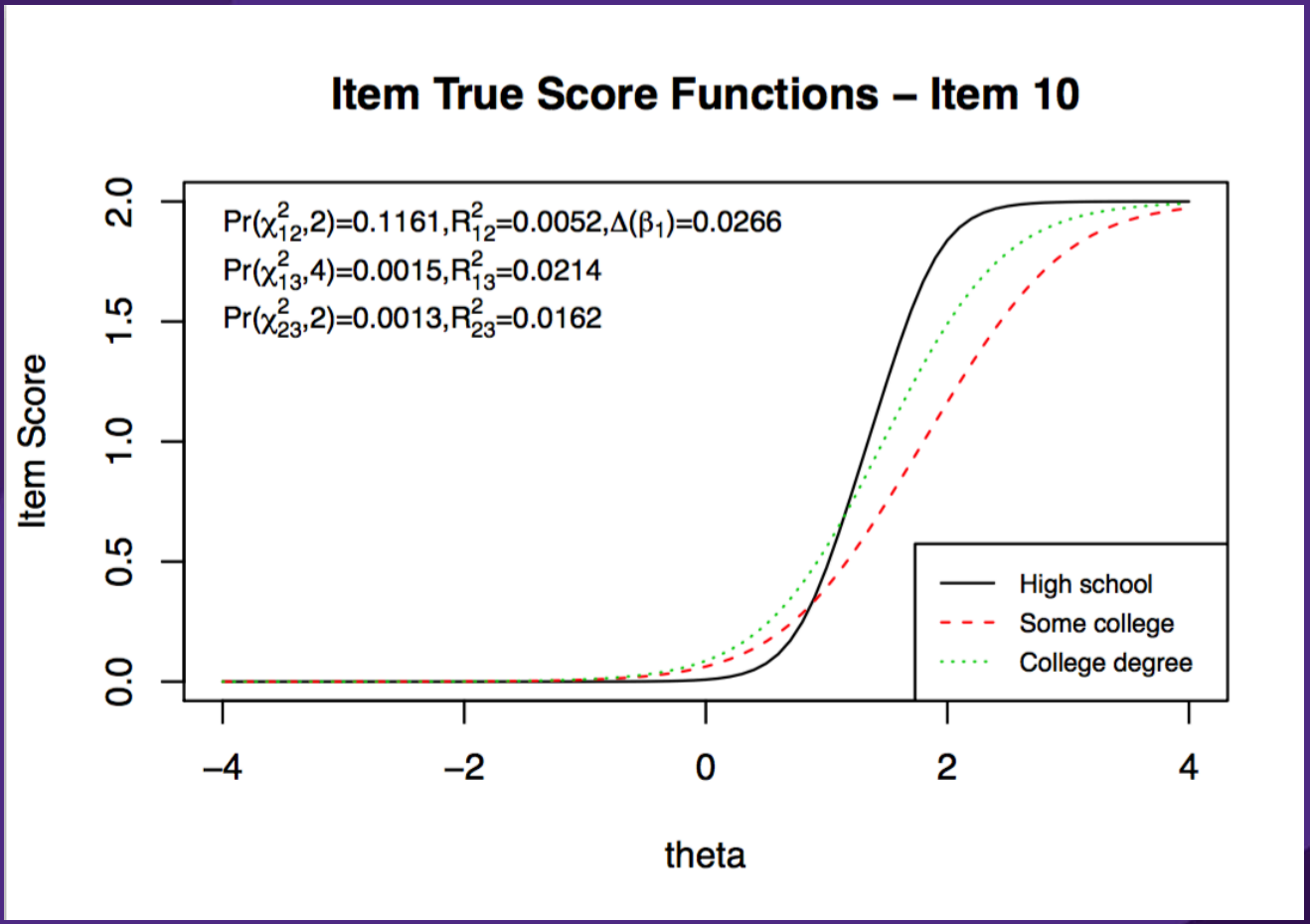
- Educational attainment 3 groups provided most accurate information
  - High school or less, some college or technical school, college graduate or above
- McFadden's pseudo R^2 as cutoff for flagging items for DIF
  - Different cutoffs chosen for each sample and measure
  - Effect size of DIF

DIF Cutoffs

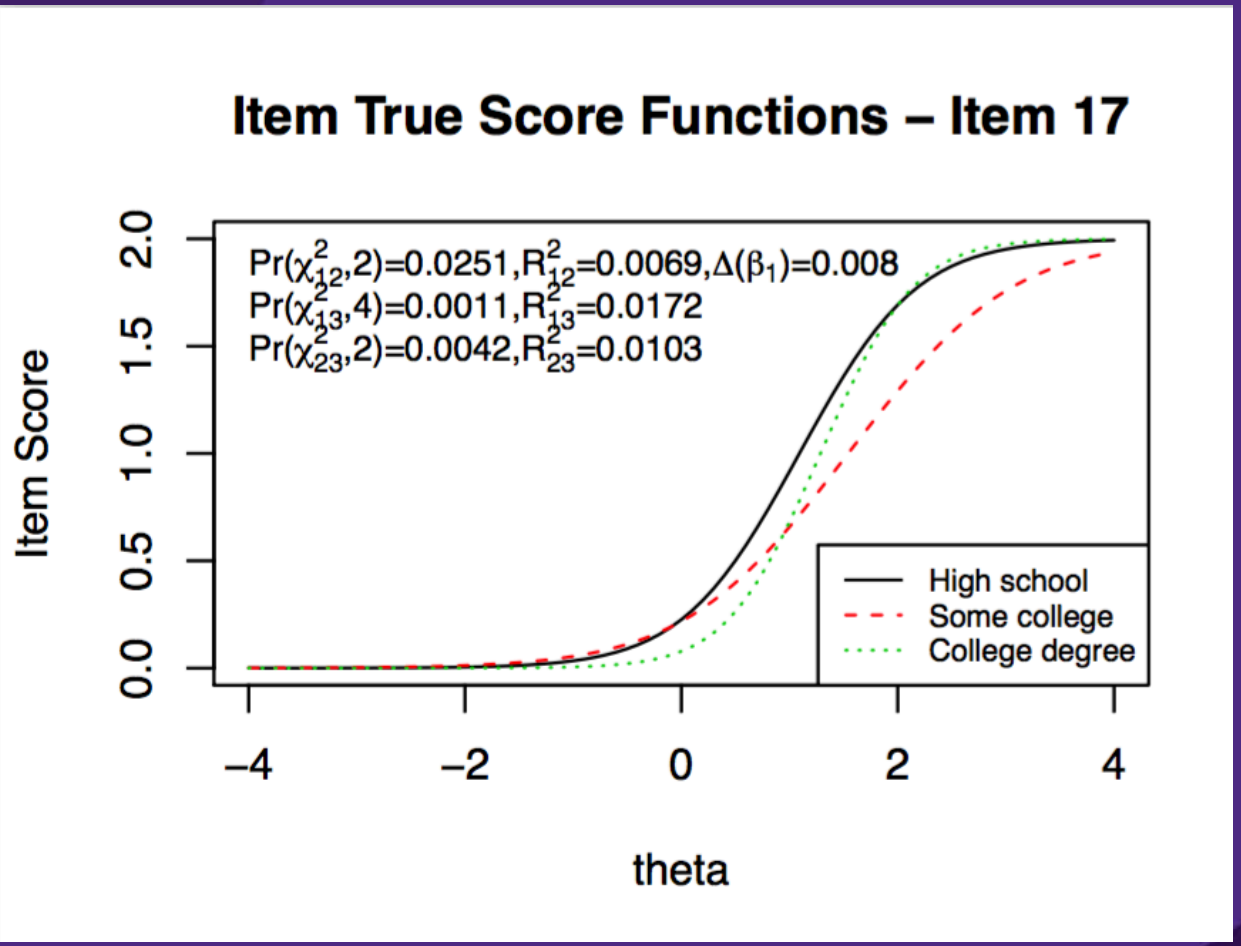
- NIH Toolbox
  - CES-D: R^2 = .01
  - PHQ-9: R^2 = .004
  - PROMIS Depression (20 items): R^2 = .009
- PROMIS 1 Wave 1
  - CES-D: R^2 = .008
  - PROMIS Depression (28 items): R^2 = .007
- PROsetta Stone
  - BDI: R^2 = .01
  - PROMIS Depression (15 items): R^2 = .008

DIF Results

	PROMIS Depression	BDI-II	PHQ-9	CES-D
NIH Toolbox Calibration	**1** .01		**2** .005, .007	**1** **.017**
PROMIS 1 Wave 1	**4** .009, .008, .009, .01			**5** .011, .013, **.022**, .011, .013
PROsetta Stone	**4** **.02**, .011, .013, .008	**1** .012		



CESD Item 10:
I felt fearful



CESD Item 17: I had crying spells

Readability Statistics

- Flesch Kincaid
  = *(0.39 * Average Words/Sentence) + (11.8 * Average Syllables/Word)*
- Gunning Fog Index
  = *0.4 (Average Words/Sentence + Percentage of 'Hard' Words)*
- Coleman Liau
  = *0.0588* Letters/100words – sentences/100 words*
- SMOG
  = *3 + sqrt(# of polysyllabic words)*
- Automated Readability Index
  = *4.71(Characters/Word) + 0.5 (Words/Sentences) – 21.43*
- Average Grade Level
  = *average of the above scores*

Readability Findings

Measure	Flesch Kincaid	Gunning Fog Index	Coleman Liau	SMOG	Automated Readability Index	Average Grade Level
BDI-II	3.8	7.0	3.2	8.2	0	4.4
CES-D	2.0	4.7	2.5	6.4	0	3.1
PHQ-9	5.6	7.4	8.8	8.4	5	7.0
PROMIS Depression 15	1.5	4.5	3.1	6.0	0	3.0
PROMIS Depression 20	2.0	4.9	3.9	6.4	1	3.6
PROMIS Depression 28	1.9	4.8	3.2	6.1	0	3.2

Conclusions

- All measures displayed DIF by education for at least one item
  - Hypothesis partially correct
  - PROMIS items also flagged for DIF
  - Overall, found *higher* item slopes for people with high school education or below
- Level of education needs to be considered during development and administration of instruments measuring depression
- DIF is a useful tool to indicate which items may be more difficult for individuals with lower educational attainment

Clinical Implications

- Measures studied are widely used in clinical settings
- Inaccurate assessment of depressive symptoms in patients with lower educational attainment

→ clinical interview → diagnosis → treatment

Limitations and Future Directions

- Limitations:
  - Internet based samples
  - Number of PROMIS items not consistent across samples
- Going forward closely examine IRT parameters for each item among items flagged
- Explore other methods for analyzing DIF

Northwestern

Acknowledgements

Grant support: National Institute on Minority Health and Health Disparities: Reducing Assessment Barriers for Patients with Low Literacy (1R01MD010440-01A1)

Contact Information:
Bayley J. Taple, MS
btaple@u.northwestern.edu

References

- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.
- Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., . . . Hays, R. (2010). The Patient-Reported Outcomes Measurement Infor- mation System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63,* 1179–1194. doi:10.1016/j.jclinepi.2010.04.011
- Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. *Psychological Assessment*, *26*(2), 513-527. doi: 10.1037/a0035768
- Crane, P. K., Gibbons, L. E., Jolley, L., Van Belle, G., Selleri, R., Dalmonte, E., & De Ronchi, D. (2006). Differential item functioning related to education and age in the Italian version of the Mini-mental State Examination. International psychogeriatrics, 18(3), 505-515. https://doi.org/10.1017/S1041610205002978
- Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., … Teresi, J. A. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research*, *16*(1), 69-84. doi: 10.1007/s11136-007-9185-5

- Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283-284. http://dx.doi.org.turing.library.northwestern.edu/10.1037/h0076540
- Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221-233. http://dx.doi.org.turing.library.northwestern.edu/10.1037/h0057532
- Fry, E. (1968). A readability formula that saves time. *Journal of reading*, *11*(7), 513-578. Retrieved from https://www.jstor.org/stable/40013635
- Gunning, R. (1952). *The technique of clear writing.* New York NY: McGraw-Hill
- Jones, R. N., & Gallo, J. J. (2002). Education and sex differences in the Mini-Mental State Examination: Effects of differential item functioning. *The Journals of Gerontology: Series B*, *57*(6), P548–P558. https://doi.org/10.1093/geronb/57.6.P548
- Kessler, R. C., Petukhova, M., Sampson, N. A., Zaslavsky, A. M., & Wittchen, H.-U. (2012). Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *International Journal of Methods in Psychiatric Research*, *21*(3), 169–184. https://doi.org/10.1002/mpr.1359
- Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9 validity of a brief depression severity measure. *Journal of General Internal Medicine, 16,* 606–613. doi:10.1046/j.1525-1497.2001.016009606.x
- McLaughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, *12*(8), 639-646. Retrieved from https://www.jstor.org/stable/40011226

- Murden, R. A., McRae, T. D., Kaner, S., & Bucknam, M. E. (1991). Mini-Mental State Exam scores vary with education in blacks and whites. *Journal of the American Geriatrics Society*, *39*(2), 149-155. https://doi-org.turing.library.northwestern.edu/10.1111/j.1532-5415.1991.tb01617.x
- Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401. doi:10.1177/014662167700100306
- Ramirez, M., Teresi, J. A., Holmes, D., Gurland, B., & Lantigua, R. (2006). Differential item functioning (DIF) and the Mini-Mental State Examination (MMSE): Overview, sample, and issues of translation. *Medical care*, S95-S106. Retrieved from http://www.jstor.org.turing.library.northwestern.edu/stable/41219509
- Smith, E. A., & Senter, R. J. (1967). Automated readability index. AMRL-TR. Aerospace Medical Research Laboratories (US), 1-14.
- Teresi, J. A., Kleinman, M., Ocepek-Welikson, K., Ramirez, M., Gurland, B., Lantigua, R., & Holmes, D. (2000). Applications of item response theory to the examination of the psychometric properties and differential item functioning of the comprehensive assessment and referral evaluation dementia diagnostic scale among samples of Latino, African American, and white non-Latino elderly. Research on Aging, 22(6), 738-773. https://doi-org.turing.library.northwestern.edu/10.1177/0164027500226007
- Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., … Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science Quarterly*, *51*(2), 148–180. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2844669/
- Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532-547.