Analysis of Differential Item Functioning in PROMIS® Pediatric and Adult Measures between Adolescents and Young Adults with Special Health Care Needs

Dan V. Blalock[1,2], Li Lin[3], Mian Wang[4], David Thissen[5], Darren A. DeWalt[6], I-Chan Huang[7] & Bryce B. Reeve[3,8]

[1] Center of Innovation to Accelerate Discovery and Practice Transformation, Durham Veterans Affairs Medical Center
[2] Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine
[3] Department of Population Health Sciences, Duke University
[4] Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill
[5] Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill
[6] Department of Medicine, UNC School of Medicine
[7] Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital
[8] Department of Pediatrics, Duke University School of Medicine

Abstract:
Purpose: Many research studies seek to assess health outcomes among patients across the adolescent-adult age groups. This age group distinction often leads to independent scale creation and validation in self-report measures, such as the Patient-Reported Outcomes Measurement Information System® (PROMIS®) health-related quality of life (HRQOL) measures. Research studies would benefit from the ability to use a single measure across these age groups.

Method: This study is a secondary data analysis of adolescents (age 14-17) and young adults (age 18-20) with special healthcare needs (n = 874). Participants completed short forms of both PROMIS pediatric and adult measures of physical functioning, pain, fatigue, depression, social health, anxiety, and anger. Differential item functioning (DIF) across age groups was examined using Wald tests for graded response model (GRM) item parameters.

Results: No DIF across age group was observed for any item in any of the pediatric or adult short form measures.

Conclusion: These results support the flexible use of pediatric and adult PROMIS HRQOL scales for adolescents and young adults age 14-20.

Keywords:
PROMIS, Item Response Theory, Differential Item Functioning, Health-Related Quality of Life, Pediatric

1 Introduction

The National Institutes of Health launched the Patient-Reported Outcomes Measurement Information System® (PROMIS®) initiative in 2004 with the goal to provide researchers access to a set of patient-reported measures of health-related quality of life (HRQOL) with strong evidence for its validity and reliability across a broad range of diseases and health conditions (Cella, Yount, Rothrock, Gershon, Cook et al., 2007). Through the collaborative work of a large multi-disciplinary team of investigators, both pediatric and adult PROMIS measures are available for commonly experienced HRQOL domains including, but not limited to, physical functioning, pain, fatigue, depression, social health, anxiety, and anger (see: https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis). The PROMIS measures are now being used globally in both clinical research and healthcare delivery settings (Alonso, Bartlett, Rose, Aaronson, Chaplin et al., 2013).

The PROMIS pediatric measures were designed and evaluated in children and adolescents between 8 and 17 years of age and the adult PROMIS measures were designed and evaluated in adults 18 years of age and older. Often, there are circumstances when a given research study will include participants who are below and above the 18-year threshold. This may be a study that collects cross-sectional or longitudinal data across a broad age range that spans the 18-year threshold or a prospective study that begins data collection in adolescence and follows the participant into young adulthood. The goal would be to have a consistent measure of HRQOL that would allow investigators to compare or to combine scores across age groups, or to have consistent metrics to plot HRQOL trajectories over time for the same individuals.

In previous research, PROMIS investigators collected responses to both pediatric and adult PROMIS measures of common HRQOL domains from 874 adolescents and young adults, between 14 and 20 years of age, with special healthcare needs and who require health services (Reeve, Thissen, DeWalt, Huang, Liu et al., 2016). Using a novel linking methodology called calibrated projection, algorithms were created to allow investigators to convert scores from the PROMIS pediatric measures to the PROMIS adult measures, and vice-versa, for similar HRQOL domains (Reeve et al., 2016; Thissen, Liu, Magnus, & Quinn, 2015; Tulsky, Kisala, Boulton, Jette, Thissen et al., 2019). For the hypothetical research studies described in the previous paragraph, these linking algorithms would allow researchers to administer the PROMIS pediatric measures for adolescents less than 18 and the PROMIS adult measures for participants 18 years or older.

However, a better option is available for researchers that includes having young adults (e.g., between 18-20 years) complete the PROMIS pediatric measures to have the same questions as the adolescents in the study, or that includes having the adolescents (e.g., between 14-17 years) complete the PROMIS adult measures to have the same questions as the young adults in the study. Or the hypothetical prospective study would have the same participant complete just the PROMIS pediatric or just the PROMIS adult measure as they are followed from adolescence into young adulthood. For these options, we need to make sure that the items within the PROMIS pediatric and adult measures are interpreted similarly by adolescents and young adults.

The goal of this secondary analysis of data from individuals with special healthcare needs is to perform tests of differential item functioning (DIF) of the PROMIS measures. DIF occurs when individuals from two different groups (in this study, adolescents and young adults) have different probabilities for selecting the response options for an item (e.g., "I got tired easily") even after controlling for the differences between the two groups on the construct being measured by the scale (e.g., PROMIS measure of fatigue) (Reeve, Pinheiro, Jensen, Teresi, Potosky et al., 2016; Thissen, Steinberg, & Wainer, 1993). In other words, an adolescent and a young adult experiencing the same level of fatigue could interpret the item "I got tired easily" differently, and thus respond differently. Given the rigorous qualitative and quantitative methodology used to design and evaluate the items that went into the PROMIS measures (DeWalt, Rothrock, Yount, & Stone, 2007; Reeve, Hays, Bjorner, Cook, Crane et al., 2007; Irwin, Varni, Yeatts, & DeWalt, 2009), we hypothesize there would be no DIF in the PROMIS items due to age group. If this hypothesis is supported by the data, then this would allow investigators to use the same PROMIS version in their study instead of using both the pediatric and adult measures for individuals between 14 and 20 years of age.

2 Methods

2.1 Participants and Procedure

The current investigation is a secondary data analysis of 874 adolescents and young adults with "Special Health Care Needs" as previously described by Reeve and colleagues (Reeve et al., 2016). Adolescents and young adults were 14 to 20 years old, able to read, write, speak English, and have access to a computer with internet. This sample was collected to represent a diverse set of illnesses affecting HRQOL. Participants aged 14 to 17 were classified as the "adolescent sample." Participants aged 18 to 20 were classified as the "young adult sample."

All participants completed a demographic questionnaire capturing the respondent's age, sex, race, ethnicity, and education level, as well as items related to the presence of any self-reported health conditions. Then, all participants completed PROMIS pediatric short forms and corresponding PROMIS adult short forms for the following concepts: physical functioning, pain, fatigue, social health, depression, anxiety, and anger. In pediatric measures, physical functioning is captured with two subscales – upper extremity and mobility. Social health is captured by the PROMIS pediatric measure of peer relationships. In adult measures, social health is captured with a scale on emotional support. Order of administration of the PROMIS pediatric and adult measures was randomized.

All PROMIS items analyzed in the current study have five ordered response categories. Response categories differed by scale: pediatric physical functioning-upper extremity and mobility (1=with no trouble, 2=with a little trouble, 3=with some trouble, 4=with a lot of trouble, 5=not able to do), adult physical functioning (1=not at all, 2=very little, 3=somewhat, 4=quite a lot, 5=cannot do), pediatric fatigue, pediatric pain, pediatric depression, pediatric anxiety, pediatric anger, pediatric peer relationships (1 = never, 2 = almost never, 3 = sometimes, 4 = often, 5 = almost always), adult fatigue, adult pain (1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much), adult depression, adult anxiety, adult anger, adult social health (1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always).

This secondary data analysis was ruled exempt from IRB review by Duke University School of Medicine.

2.2 Statistical Analysis

Item Response Theory (IRT) item parameters for the polytomous ordered response categories were estimated using Samejima's Graded Response Model (GRM; Samejima, 1969). DIF was examined on every item within each PROMIS pediatric measure and adult measure between the adolescent and young adult samples. The young adult sample was the focal group for DIF analyses on all PROMIS pediatric measures, and the adolescent sample was the focal group for DIF analyses on all PROMIS adult measures. The Wald test was used to test for DIF in the GRM discrimination and threshold parameters simultaneously in an omnibus test, and then separately for each parameter. Models did not include covariates such as sex or health condition, as prior studies have investigated DIF across these key attributes (Reeve et al., 2016; Jones, Tommet, Ramirez, Jensen, & Teresi, 2016; Coster, Ni, Slavin, Kisala, Nandakumar et al., 2016; Irwin, Stucky, Langer, Thissen, DeWitt et al., 2010; DeWalt, Thissen, Stucky, Langer, Morgan et al., 2013). Similar to other IRT investigations of PROMIS measures (Irwin et al., 2010), we used the Benjamini–Hochberg procedure to control for the multiplicity of comparisons involved in checking each item for DIF using $\alpha = 0.05$ (Benjamini & Hochberg, 1995).

There were no a priori hypotheses regarding which items within a PROMIS scale are most appropriate to serve as anchors for the DIF tests. Thus, to examine the sensitivity of the results to anchor item selection, DIF analyses were conducted in two ways (based on guidance from Woods, Cai, & Wang, 2013). First, DIF was examined using a two-step "anchor-all-test-all" method, where focal group factor means and variances were estimated by constraining all item parameters to be equal between groups, followed by free estimation of all item parameters while fixing factor means and variances to their values estimated in the previous step. Second, DIF was examined based on three selected anchor items in each scale (comprising between 30% and 50% of the total items within a scale). The top three anchor items demonstrating nonsignificant Wald χ^2 statistics from the "anchor-all-test-all" method were selected within each scale. Multiple simulation studies have found this method to better control for Type 1 error rate in both binary and polytomous items (Wang & Woods, 2017; Cao, Tay, & Liu, 2017). Examinations of DIF for individual item discrimination or threshold parameters, and magnitude of DIF, were to be conducted only if any omnibus DIF across age group was detected at the item level. R (version 4.0.0) was used for all analyses, with IRT parameters and DIF tested using the 'mirt' package (R Core Team, 2020).

3 Results

3.1 Demographic and Clinical Information

The sample included 415 adolescents between 14 and 17 years of age (Mean = 15.63, SD = 1.20) and 459 young adults between 18-20 years of age (Mean = 18.93, SD = 0.75). Adolescents were 48.2% female, and young adults were 57.7% female. Adolescents were 52.0% White, 22.4% Black, 12.3% "Other," 8.2% Asian, 3.4% Multiple Race, and 1.7% did not report race. Young adults were 49.7% White, 19.4% Black, 13.7% "Other," 8.7% Asian, 4.1% Multiple Races, and 4.4% did not report race. Adolescents were 33.3% Hispanic, and young adults were 42.7% Hispanic. The top five health conditions for adolescents were ADHD (33.3%), Mental Health (i.e., a wide range of conditions that affect mood, thinking, or behavior; 25.1%), Allergies (23.6%), Asthma (22.4%), and Chronic Pain (21.2%). The top five health conditions for young adults were Hypertension (25.7%), ADHD (23.5%), Asthma (22.7%), Chronic Pain (22.4%), and Mental Health (20.7%).

3.2 Differential Item Functioning

See Table 1 for full DIF analysis results. No DIF between age groups was identified for any item across the eight PROMIS pediatric scales using omnibus Wald tests. Additionally, no DIF between age groups was identified for any item across the seven PROMIS adult scales using omnibus Wald tests. This was true both when using an "anchor-all-test-all" procedure, as well as three anchor items for each scale. Wald χ^2 p-values were all above their respective cutoffs based on the Benjamini-Hochberg procedure for multiple comparisons (Benjamini & Hochberg, 1995). Because no omnibus Wald tests were significant, we did not conduct DIF analysis for individual IRT item parameters.

Table 1 DIF Estimates and Significance for PROMIS HRQOL Pediatric and Adult Short Forms

PROMIS Item Content	"Anchor-all-test-all" Wald χ^2	"Anchor-all-test-all" Wald *p*-value	3 Anchor Items Wald χ^2	3 Anchor Items Wald *p*-value
PROMIS Pediatric Physical Functioning – Upper Extremity Measure				
I could button my shirt or pants.	2.21	.82	2.41	.79
I could open a jar by myself.	1.84	.87	2.76	.74
I could open the rings in school binders.	1.56	.91	4.77	.74
I could pour a drink from a full pitcher.	2.29	.81	2.99	.70
I could pull a shirt on over my head by myself.	2.93	.71	Anchor	Anchor
I could pull open heavy doors.	1.12	.95	1.68	.89
I could put my shoes on by myself.	6.63	.25	Anchor	Anchor
I could use a key to unlock a door.	1.89	.86	Anchor	Anchor
PROMIS Pediatric Physical Functioning – Mobility Measure				
I could do sports and exercise that other kids my age could do.	2.56	.77	3.84	.57
I could get up from the floor.	7.04	.22	Anchor	Anchor
I could keep up when I played with other kids.	3.64	.60	1.77	.88
I could move my legs.	1.65	.90	2.14	.83
I could stand up by myself.	3.66	.60	Anchor	Anchor
I could stand up on my tiptoes.	7.71	.17	Anchor	Anchor
I could walk up stairs without holding on to anything.	8.47	.13	7.99	.16
I have been physically able to do the activities I enjoy most.	2.34	.80	4.31	.51
PROMIS Adult Physical Functioning Measure				
Does your health now limit you in doing vigorous activities, such as running, lifting heavy objects, participating in strenuous sports?	1.56	.91	1.05	.96
Does your health now limit you in walking more than a mile?	2.93	.71	1.90	.86
Does your health now limit you in climbing one flight of stairs?	1.99	.85	2.48	.78
Does your health now limit you in lifting or carrying groceries?	2.13	.83	Anchor	Anchor
Does your health now limit you in bending, kneeling, or stooping?	1.56	.91	2.25	.81
Are you able to do chores such as vacuuming or yardwork?	0.31	.99	1.33	.93
Are you able to dress yourself, including tying shoelaces and doing buttons?	1.00	.96	1.33	.93
Are you able to shampoo your hair?	2.28	.81	1.24	.94
Are you able to wash and dry your body?	2.37	.80	Anchor	Anchor
Are you able to get on and off the toilet?	2.63	.76	Anchor	Anchor

PROMIS Pediatric Fatigue Measure				
Being tired made it hard for me to play or go out with my friends as much as I'd like.	5.92	.31	4.82	.44
I felt weak.	5.87	.32	4.97	.42
I got tired easily.	4.90	.43	5.08	.41
Being tired made it hard for me to keep up with my schoolwork.	1.55	.91	3.02	.70
I had trouble finishing things because I was too tired.	1.26	.94	Anchor	Anchor
I had trouble starting things because I was too tired.	5.55	.35	Anchor	Anchor
I was so tired it was hard for me to pay attention.	.51	.99	2.45	.78
I was too tired to do sports or exercise.	2.10	.83	5.06	.41
I was too tired to do things outside.	5.04	.41	Anchor	Anchor
I was too tired to enjoy the things I like to do.	6.24	.28	3.77	.58
PROMIS Adult Fatigue Measure				
I feel fatigued	3.88	.57	2.71	.74
I have trouble starting things because I am tired	1.14	.95	0.91	.97
How run-down did you feel on average?	6.92	.23	Anchor	Anchor
How fatigued were you on average?	6.31	.28	Anchor	Anchor
How much were you bothered by your fatigue on average?	10.62	.06	10.70	.06
To what degree did your fatigue interfere with your physical functioning?	7.70	.17	6.22	.29
How often did you have to push yourself to get things done because of your fatigue?	5.44	.26	3.48	.63
How often did you have trouble finishing things because of your fatigue?	6.56	.26	Anchor	Anchor
PROMIS Pediatric Pain Measure				
I had trouble sleeping when I had pain.	5.62	.34	8.87	.12
I felt angry when I had pain.	1.58	.90	3.51	.62
I had trouble doing schoolwork when I had pain.	3.07	.69	3.30	.65
It was hard for me to pay attention when I had pain.	1.63	.90	2.30	.81
It was hard for me to run when I had pain.	1.64	.90	Anchor	Anchor
It was hard for me to walk one block when I had pain.	2.97	.70	Anchor	Anchor
It was hard for me to have fun when I had pain.	7.02	.22	9.64	.09
It was hard for me to stay standing when I had pain.	3.15	.68	Anchor	Anchor
PROMIS Adult Pain Measure				
How much did pain interfere with your day to day activities?	6.67	.25	5.77	.33
How much did pain interfere with work around the home?	2.97	.70	5.33	.38
How much did pain interfere with your ability to participate in social activities?	1.44	.92	1.66	.89
How much did pain interfere with your enjoyment of life?	4.11	.53	Anchor	Anchor
How much did pain interfere with things you usually do for fun?	2.87	.72	Anchor	Anchor
How much did pain interfere with your enjoyment of social activities?	3.60	.61	Anchor	Anchor
How much did pain interfere with your household chores?	2.06	.84	0.92	.97
How much did pain interfere with your family life?	2.51	.78	1.88	.87

PROMIS Pediatric Depressive Symptoms Measure				
I could not stop feeling sad.	1.38	.93	1.66	.89
I felt alone.	3.22	.67	4.07	.54
I felt everything in my life went wrong.	3.37	.64	Anchor	Anchor
I felt like I couldn't do anything right.	0.80	.98	1.00	.96
I felt lonely.	1.41	.92	1.22	.94
I felt sad.	1.52	.91	Anchor	Anchor
I felt unhappy.	1.92	.86	Anchor	Anchor
I thought that my life was bad.	3.40	.64	4.72	.45
PROMIS Adult Depression Measure				
I felt worthless.	3.36	.65	Anchor	Anchor
I felt like I had nothing to look forward to.	4.18	.52	5.54	.35
I felt helpless.	4.10	.54	3.15	.68
I felt sad.	6.30	.28	Anchor	Anchor
I felt like a failure.	13.16	.02	13.00	.02
I felt depressed.	1.34	.93	1.29	.94
I felt unhappy.	4.75	.45	10.74	.06
I felt hopeless.	5.00	.41	Anchor	Anchor
PROMIS Pediatric Anxiety Measure				
I felt nervous.	5.34	.38	3.66	.60
I felt scared.	7.70	.17	11.01	.05
I felt worried.	9.87	.08	Anchor	Anchor
I felt like something awful might happen.	2.80	.73	Anchor	Anchor
I thought about scary things.	4.51	.49	5.44	.37
I was afraid that I would make mistakes.	10.96	.05	12.39	.03
I worried about what could happen to me.	4.07	.54	2.57	.77
I worried when I went to bed at night.	7.18	.21	Anchor	Anchor
PROMIS Adult Anxiety Measure				
I felt fearful.	2.37	.80	2.78	.73
I found it hard to focus on anything other than my anxiety.	2.01	.85	4.39	.50
My worries overwhelmed me.	3.30	.65	2.81	.73
I felt uneasy.	0.79	.98	Anchor	Anchor
I felt nervous.	2.29	.81	Anchor	Anchor
I felt like I need help for my anxiety.	1.38	.93	0.86	.97
I felt anxious.	0.64	.99	Anchor	Anchor
I felt tense.	4.72	.45	5.10	.40
PROMIS Pediatric Anger Measure				
I felt mad.	4.73	.45	2.91	.72
I felt upset.	5.11	.40	Anchor	Anchor
I felt fed up.	6.19	.29	4.16	.53
I was so angry I felt like throwing something.	6.49	.26	Anchor	Anchor
I was so angry I felt like yelling at somebody.	10.59	.26	Anchor	Anchor
When I got mad, I stayed mad.	3.78	.58	3.12	.68

PROMIS Adult Anger Measure				
I was irritated more than people knew.	0.92	.97	3.37	.64
I made myself angry about something just by thinking about it.	6.93	.23	10.51	.06
I felt angry.	0.83	.98	Anchor	Anchor
I felt like I was ready to explode.	1.46	.92	Anchor	Anchor
I stayed angry for hours.	1.08	.96	Anchor	Anchor
I felt angrier than I thought I should.	11.53	.04	11.97	.04
I was grouchy.	1.32	.93	1.81	.87
I felt annoyed.	1.57	.91	1.38	.93
PROMIS Pediatric Peer Relationships Measure				
I felt accepted by other kids my age.	0.96	.97	2.98	.70
I was able to count on my friends.	3.51	.62	4.32	.51
I was able to talk about everything with my friends.	4.45	.49	5.77	.33
I was good at making friends.	7.86	.16	8.18	.15
My friends and I helped each other out.	3.21	.67	Anchor	Anchor
Other kids wanted to be my friend.	6.26	.28	7.42	.19
Other kids wanted to be with me.	1.42	.92	Anchor	Anchor
Other kids wanted to talk to me.	5.09	.40	Anchor	Anchor
PROMIS Adult Social Health: Emotional Support Measure				
I have someone to confide in or talk to about myself or my problems.	4.78	.44	6.34	.27
I have someone who understands my problems.	5.27	.38	4.90	.43
I have someone who will listen to me when I need to talk.	1.69	.89	1.68	.89
I have someone to talk to when I have a bad day.	5.03	.41	Anchor	Anchor
I have someone I trust to talk with about my problems.	3.53	.62	Anchor	Anchor
I have someone I trust to talk with about my feelings.	3.96	.56	Anchor	Anchor
I have someone with whom I share my most private worries and fears.	1.71	.89	3.56	.62
I have someone who makes me feel appreciated.	3.01	.70	1.96	.86

Note Designated anchors selected based on Wald-2 and discrimination parameters.

p-values reported are uncorrected for multiple comparisons.

4 Discussion

This study described IRT-based DIF analyses conducted across adolescent and young adult age groups on eight PROMIS pediatric and seven PROMIS adult short-form measures, representing domains of physical function, pain, fatigue, social health, depression, anxiety, and anger. The study found no statistically significant DIF across all items of all scales for both PROMIS pediatric and adult measures. Furthermore, DIF was not found when examining varying methods for anchor item selection. These findings suggest that item discrimination and threshold parameters function similarly in each of these scales for adolescents and young adults with a broad range of health conditions and healthcare needs.

These findings yield at least two practical advantages for research using these PROMIS measures. First, longitudinal studies following participants' HRQOL between the ages of 14 and 20 may be able to use the same version of these PROMIS measures (pediatric or adult) throughout these ages in repeated measurements over time. While other sources of variance in these scores over time may still exist, it is unlikely any differences in responses would arise due to participants aging. Second, cross-sectional studies examining participants' HRQOL may also be able to use the same version of these PROMIS measures irrespective of participants' ages between 14 and 20. Selection of which version to use should be driven by an understanding of the patient population, their cognitive functioning, and the nature of effects the research is intended to uncover. Even among those with the same names, the PROMIS pediatric and adult scales measure more or less different constructs, with correlations ranging from 0.93 in this sample for Depressive Symptoms down to a correlation in the 0.6 range between the different Social scales (the latter of which was not linked by Reeve et al., 2016). For any particular study, either the pediatric or adult scale may be judged to include questions that may be more responsive to treatment or the variables at hand.

4.1 Limitations & Future Directions

The results are limited to the items included on the short forms used in this study. We cannot generalize to other items within the banks for the PROMIS domains included in this study, nor can we generalize to items within other PROMIS domains. However, we would expect similar results as all items in the PROMIS banks went through extensive qualitative and quantitative testing to select well performing and easily comprehensible items. Likewise, we cannot generalize results to age groups below or above our age range. Because we combined ages within each group, we can't guarantee the PROMIS adult scale would work as well for just the 14 year olds or the PROMIS pediatric scale would work just as well for just the 20 year olds. We do not have enough sample size to test for each year of age. Thus, some caution would be needed if using an adult form for the youngest age group and the pediatric form for the oldest age group. Lastly, these results do not necessarily generalize to the language translations of the PROMIS measures or in other populations distinct from those with special healthcare needs, including different race/ethnicity demographics.

4.2 Conclusion

This study addresses a key gap in the large literature on PROMIS measures and provides some confidence the either the pediatric or adult forms may be used for studies that involve adolescents and young adults. The findings of no DIF provide another option for investigators who sensibly prefer not to use the linked metrics.

References

Alonso, J., Bartlett, S. J., Rose, M., Aaronson, N. K., Chaplin, J. E., Efficace, F., ... & Ravens-Sieberer, U. (2013). The case for an international patient-reported outcomes measurement information system (PROMIS®) initiative. *Health and quality of life outcomes*, *11*(1), 1-5. https://doi.org/10.1186/1477-7525-11-210.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Cao, M., Tay, L., & Liu, Y. (2017). A Monte Carlo Study of an Iterative Wald Test Procedure for DIF Analysis. *Educational and Psychological Measurement*, *77*(1), 104–118. https://doi.org/10.1177/0013164416637104

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B.B., ... & Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Medical Care*, *45*(5 Suppl 1), S3.

Coster, W. J., Ni, P., Slavin, M. D., Kisala, P. A., Nandakumar, R., Mulcahey, M. J., ... & Jette, A. M. (2016). Differential item functioning in the P atient R eported O utcomes M easurement I nformation S ystem P ediatric S hort F orms in a sample of children and adolescents with cerebral palsy. *Developmental Medicine & Child Neurology*, *58*(11), 1132-1138.

DeWalt, D.A., Thissen, D., Stucky, B.D., Langer, M.M., Morgan DeWitt, E., Irwin, D.E., Lai, J.S., Yeatts, K.B., Gross, H.E., Taylor, O. and Varni, J.W. (2013). PROMIS pediatric peer relationships scale: development of a peer relationships item bank as part of social health measurement. *Health Psychology*, *32*(10), p.1093-1103.

DeWalt, D.A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*, *45*(5 Suppl 1), S12–S21. doi:10.1097/01.mlr.0000254567.79743.e200005650-200705001-00003.

Irwin, D. E., Stucky, B., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J. S., ... & DeWalt, D.A. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research*, *19*(4), 595-607. https://doi.org/10.1007/s11136-010-9619-3

Irwin, D. E., Varni, J. W., Yeatts, K., & DeWalt, D.A. (2009). Cognitive interviewing methodology in the development of a pediatric item bank: a patient reported outcomes measurement information system (PROMIS) study. *Health and Quality of Life Outcomes*, *7*(1), 3. https://doi.org/10.1186/1477-7525-7-3

Jones, R. N., Tommet, D., Ramirez, M., Jensen, R., & Teresi, J. A. (2016). Differential item functioning in Patient Reported Outcomes Measurement Information System®(PROMIS®) physical functioning short forms: Analyses across ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(2), 371-402.

R Core Team. R (2020): A language environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reeve, B.B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... & Liu, H. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, S22-S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04.

Reeve, B.B., Thissen, D., DeWalt, D.A., Huang, I.C., Liu, Y., Magnus, B., ... & Haley, S. (2016). Linkage between the PROMIS® pediatric and adult emotional distress measures. *Quality of Life Research*, *25*(4), 823-833. https://doi.org/10.1007/s11136-015-1143-z

Reeve, B.B., Pinheiro, L. C., Jensen, R. E., Teresi, J. A., Potosky, A. L., McFatrich, M. K., ... & Chen, W. H. (2016). Psychometric evaluation of the PROMIS® fatigue measure in an ethnically and racially diverse population-based sample of cancer patients. *Psychological Test and Assessment Modeling*, *58*(1), 119-139.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph Supplement, 17 (4, Pt. 2). https://doi.org/10.1007/BF03372160

Thissen, D., Liu, Y., Magnus, B., & Quinn, H. (2015). Extending the use of multidimensional IRT calibration as projection: Many-to-one linking and linear computation of projected scores. In *Quantitative Psychology Research* (pp. 1-16). Springer, Cham. https://doi.org/10.1007/978-3-319-19977-1_1

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (p. 67–113). Lawrence Erlbaum Associates, Inc.

Tulsky, D. S., Kisala, P. A., Boulton, A. J., Jette, A. M., Thissen, D., Ni, P., ... & Slavin, M. (2019). Determining a transitional scoring link between PROMIS® pediatric and adult physical health measures. *Quality of Life Research*, *28*(5), 1217-1229. https://doi.org/10.1007/s11136-018-2073-3

Wang, M., & Woods, C. M. (2017). Anchor selection using the Wald test anchor-all-test-all procedure. *Applied Psychological Measurement*, *41*(1), 17-29. https://doi.org/10.1177/0146621616668014

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. https://doi.org/10.1177/0013164412464875

Corresponding author:
Dan V. Blalock, PhD
Durham Center of Innovation to Accelerate Discovery and Practice Transformation (ADAPT)
Center for Health Services Research in Primary Care
Durham Veterans Affairs Health Care System
411 West Chapel Hill Street, Suite 600
Durham, NC 27705
daniel.blalock@duke.edu