Assessing Differential Item Functioning and Differential Test Functioning in an Academic Motivation Scale Using Item Response Theory Methods

A compelling argument can be made about the need for social work measurement researchers to focus on assessing how items and scales used in practice measure constructs equivalently in different populations (Nugent, 2017). Social work researchers who use scales to build statistical models or practitioners who use scales to identity clinically relevant disorders need to ensure these and other measurement tasks use items and scales that are free from possible bias or undesirable differential functioning. Given the diversity of social work populations and the stakes of the data-informed decisions, practitioners must make in assessments, planning, and evaluation at all levels of practice, ensuring measurement equivalence is imperative (Nugent, 2017; Tran et al., 2017; Unick & Stone, 2010).

Differential functioning refers to the condition where factors other than the construct of interest influences responses to an item in a scale (Tay et al., 2015). Measures do not display differential functioning (e.g., are invariant) if it is safely assumed that respondents with the same standing on the construct (latent variable) of interest respond to items in the same way. If any item contains construct-irrelevant variance due to, say, group membership, then a statistical test using this item is confounded by the group membership differential functioning, and conclusions based on the test would be inaccurate. Further, item differential functioning can accumulate to the scale level, the consequence of which is the scale is confounded by differential functioning and statistical tests (e.g., a test comparing group means) will contain artifacts related to group membership that compromise conclusions made about true group differences (Li & Zumbo, 2009).

On a more fundamental level, item and scale differential functioning compromise measurement validity (American Educational Research Association et al., 2014; Gómez-Benito et al., 2018). The most recent *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) stresses that validity refers to the degree to which evidence supports the interpretation of scores for proposed uses of tests and scales. Differential item functioning is a key element in the discussion relating to validity evidence based on the internal structure of a test or scale (p. 16) and in the discussion about test and scale fairness as a lack of measurement bias (pp. 51-52). In sum, the *Standards* clearly stress the importance of assessing differential item and test or scale functioning in developing evidence for validity.

Article Goals

The primary goal of this article is to describe a differential item functioning (DIF) and differential test functioning (DTF) analysis of a scale—The Academic Motivation Scale (AMS) (Anderson-Butcher et al., 2013)—included in a compendium of scales designed for use by school social workers. The Community

and Youth Collaborative Institute School Experience Surveys (CAYCI-SES) (http://cayci.osu.edu/) resource makes available various scales designed for elementary, middle, and high school students, teachers and staff, and parents and caregivers (Anderson-Butcher et al., 2020). The scales are marketed as valid and reliable measures of constructs that are important for developing needs assessments, for program planning, and for program evaluations in school settings. Through a series of field-tested analyses, the AMS developers found the scale to be a psychometrically sound measure of academic motivation (see link above). They did not, however, examine possible differential functioning of items and the scale in their studies.

In this study, we examined possible AMS differential functioning for race, gender, and family composition using Item Response Theory (IRT) methods. Since there is evidence that perceptions of academic motivation may vary by race (Graham & Hudley, 2005), gender (Bugler et al., 2013; Isik et al., 2018; Urdan & Bruchmann, 2018), and family composition (Usher & Kober, 2012), any academic motivation measurement strategy should seek to understand how items and scales are or are not equivalent across these groups.

A Brief Note on Item Response Theory

Since the language of IRT might be new to some readers, the following is brief overview of concepts and terms (for a more detailed description in the social work context, see Nugent, 2017). This terminology is helpful to understand the DIF/DTF discussion. IRT is a set of latent variable techniques designed to examine the process by which individuals respond to items in a measurement instrument. IRT modeling mathematically links each item to an underlying scale typically called theta (θ). This underlying scale is a foundational component of IRT; It is the scale used to represent the latent trait of interest (e.g., in our study academic motivation is the latent trait). In most instances, θ is expressed in a standard normal form. A basic key assumption of IRT is a respondent has a unique location on θ, which influences how he or she responds to each item.

The product of the item-θ linking process is a set parameter estimate that characterizes the relationship between an item and θ. In general, an item will have an *a*-parameter (sometimes referred to as a slope or discrimination parameter), which is an indicator of an item's ability to discriminate between different levels of θ. The *a*-parameter also is a measure of the strength of the relationship between an item and θ where higher values suggest stronger relationships (much like a factor loading in factor analysis). Further, an item will have one or more *b*-parameters (sometimes referred to as a location or difficulty parameters). The general rule is that *m*-1 *b*-parameters are estimated for ordinal scales (where *m* refers to the number of response categories) so for our five-category Likert response scale, four *b*-parameters were estimated. The *b*-parameters represent the point on θ where a respondent has a .5 probability of choosing that response

category or higher. Once an acceptable model has been fit, model parameters (*a*-parameters and *b*-parameters) can then be used to compute other useful IRT components such as item and scale information functions, conditional standard errors, model-based estimates of respondent θ scores, and model-based expected scores. As discussed below, item parameters, model-based estimates for respondent θ scores, and expected true scores play a prominent role in DIF/DTF analyses.

One last point specific to a DIF/DTF analysis is there is a general convention to identify one group as a reference group and one group as a focal group. We followed a recommended approach of using group sample size to designate reference and focal groups, with the larger groups designated as reference groups (see Table 1). Note that Tay et al. (2015) suggest in most studies the designation of the reference and focal groups is arbitrary and does not affect the computation of DIF (p. 23).

Method

Sample

The data used in this study came from 3,221 7th grade students in 17 school districts in a large mid-western U.S. urban county. The characteristics of the sample of students were as follows: 69.8% were in suburban schools, 30.2% were in the inner-city school district; 48.7% were male, 51.3% were females; 68.9% were White, 31.1% were other races; 57.1% lived in households with both parents, 42.9% lived in various other living arrangements (living with one parent, splitting time between parents, foster care). The data were collected as a part of a coordinated, county-wide effort to identify academic, social, and emotional needs of seventh grade students and data collection processes followed consent procedures prescribed in each district.

Instrument

The Academic Motivation Scale (AMS) is composed of six questions presented to respondents as follows:

1. I have a positive attitude towards school
2. I feel I have made the most of my school experiences so far
3. I like the challenges of learning new things in school
4. I am confident in my ability to manage my schoolwork
5. I feel my school experience is preparing me well for adulthood
6. I have enjoyed my school experiences so far.

The ordinal response scale for each item is *Strongly disagree* (0), *Disagree* (1), *Neither agree or disagree* (2), *Agree* (3), and *Strongly agree* (4). The summary

scale score is a total of the six items resulting in a 0-to-24 scale score range with higher scores corresponding to higher levels of perceived academic motivation.

Data Analysis

All analyses were conducted in the R statistical computing environment (R Development Core Team, 2021) using RStudio (RStudio Team, 2021). A basic assumption of unidimensional IRT models is the items composing a scale measure a single construct. We assessed dimensionality using ordinal exploratory factor analyses (EFA) methods implemented in the R package psych: Procedures for Personality and Psychological Research (Revelle, 2021). Specifically, we examined eigenvalues and scree plots following interpretation recommendations made by Tay et al. (2015, p.18).

For the DIF/DTF analysis, we used the mirt: A Multidimensional Item Response Theory Package for the R Environment (Chalmers, 2012) to fit a set of graded response models (GRM) using a full-information marginal maximum likelihood fitting function with an expectation-maximization algorithm. A GRM model is the recommended model for ordered polytomous response data (Hambleton et al., 2010). We assessed model fit using an index, C_2, specifically designed to assess the fit of IRT models for ordinal data (Cai & Monroe, 2014). We used the C_2-based root mean square error of approximation (RMSEA) as the primary fit index. In addition, we used a comparative fit index (CFI) and a standardized root mean square residual (SRMR) to assess adequacy of model fit based on suggestions made by Maydeu-Olivares (2015).

Following the fitting and assessment of the group GRM models, we proceeded to examine race, gender, and family DIF/DTF. We followed steps recommended by Meade (for a full elaboration of terminology and recommendations please refer to Meade, 2010; Meade & Wright, 2012; Tay et al., 2015). Meade recommends a two-stage approach for conducting IRT invariance analyses using a series of likelihood ratio tests and the computation of mean difference and standardized mean difference effect sizes. All the procedures in the Meade framework are implemented as functions in the mirt package.

Results

Dimensionality

EFA results supported the unidimensionality of the AMS in all groups in the study. As noted, we followed recommendations listed by Tay et al. (2015, p18) suggesting that variance accounted for by the first factor should be at least 20 percent and that the first eigenvalue should four to five times larger than the second eigenvalue. In addition, visual examination of the scree plot of eigenvalues should show a clear drop from the first to second eigenvalue.

Race eigenvalues indicated there was a dominate first factor in each group that accounted for substantial variance (48% for White students and 39% for other race students). The first eigenvalue was just over five times higher than the second eigenvalue for White students (3.38 / .66 = 5.12) and was four times higher for other race students (2.96 / .74 = 4.00). The same general pattern of results was obtained for both gender and family groups. For example, scree plots for gender groups indicated there was a dominate first factor in each group that accounted for substantial variance (46% for female students and 44% for male students). The first eigenvalue was just under five times higher than the second eigenvalue for female students (3.29 / .6 = 4.91) and was just four times higher for male students (2.96 / .74 = 4.00). Finally, scree plots indicated there was a dominate first factor in each family composition group that accounted for substantial variance (46% for two-parent students and 42% for other family students). The first eigenvalue was just over five times higher than the second eigenvalue for two-parent students (3.31 / .66 = 5.02) and was just over four times higher for other family students (3.10 / .66 = 4.49). Scree plots for all groups substantiated that a single factor was dominant.

GRM Model Fit

Once unidimensionality is established, the next step is to fit and assess individual group models. Table 1 presents results for various GRM model fit indexes for all the groups. Specifically for race, the White group CFI = .984 and other race CFI = .998, White group RMSEA = .069 and other race RMSEA = .022, and the White group SRMR = .036 and other race group SRMR = .026; For gender, the female group CFI = .992 and other race CFI = .986, female group RMSEA = .047 and male group RMSEA = .059, and the female group SRMR = .029 and male group SRMR = .034; For family, the two-parent group CFI = .987 and other family group CFI = .993, two-parent group RMSEA = .060 and other family group RMSEA = .041, and the two-parent group SRMR = .033 and other family group SRMR = .030. These results indicated a GRM model was plausible for each group using recommended threshold values of RMSEA $\leq .06$, SRMR $\leq .08$ and CFI $\geq .95$ (Maydeu-Olivares, 2015).

DIF/DTF Results

In this section, we discuss results from the Meade two-stage DIF/DTF process. In the first stage, items are assessed for DIF using a series of likelihood ratio tests. A likelihood ratio test involves comparing the fit two models: a baseline model and a comparison model. For this analysis, each item (the comparison model) is compared to a model for all other items (the baseline model) where parameters for baseline items are constrained to be equal. The difference between the comparison and baseline models is assessed by a G^2 value (distributed as χ^2) where a significant p-value indicates possible DIF (see details in Meade, 2012, p. 1017).

Results for the first stage of the DIF analysis for each of the group likelihood ratio tests are presented in Table 2. Two models are analyzed in this stage of the analysis. First, items are screened as possible DIF items using a procedure called *all-others-as-anchors*. For the *all-others-as-anchors* model, the *p*-value threshold for deciding about whether an item displays DIF is typically set at $p \leq .05$. For example, using this threshold in step one for the race comparison, item 4 did not display DIF ($p = .183$) and each of the remaining items (items 1, 2, 3, 5, and 6) were identified as possible DIF items. For the next model, non-DIF items from the first model are used as anchor items and a second series of likelihood ration tests are computed. For the race analysis, item 4 was selected as the anchor item for the *anchor-item* model from the *all-others-as-anchors* model. Results for *anchor-item* model also are shown in Table 2. These results substantiate items 1, 2, 3, 5, and 6 are DIF items as indicated by the values of the adjusted Benjamini-Hochberg $BH\text{-}p \leq .05$.

The results for gender DIF and family DIF shown in Table 2 can be similarly interpreted. For example, for gender item 3 and item 5 demonstrated DIF in the *all-others-as-anchors* model and this DIF was further substantiated in the *anchor-item* model. For the family analysis, only item 3 was detected as a DIF item in both the *all-others-as-anchors* model and the *anchor-item* model. It is interesting to note that item 5 (“School experience is preparing me for adulthood”) displayed DIF in the race and gender groups, and item 3 (“I like the challenges of learning new things in school”) displayed DIF across all groups. In the second stage, effect sizes based on estimated scores are computed and interpreted.

A few comments about these measures are in order since they represent an important feature of the Meade framework. The prior likelihood ratio tests use a traditional *p*-value interpretation to flag DIF for an item. The logic behind this interpretation is similar to the use of a *p*-value in a traditional null hypothesis statistical test situation where a relationship is deemed to be statistically significant if a *p*-value falls below some specified threshold. There are well-known criticisms of this logic, not least of which is that it characterizes a statistical relationship as a simple dichotomy—as statistically significant or not statistically significant. Effect sizes evolved to address this dichotomous line of reasoning by characterizing statistical relationships as lying on a continuum, that is, relationships that could be characterized by measures of size or magnitude (Cohen, 1992). Meade made the case that both DIF and DTF should be thought as lying on a continuum and developed a set of functions and procedures to express those relationships (Meade, 2010).

In the Meade framework, two types of continuous scale effect sizes are computed—mean differences and standardized mean differences. Mean differences are just that, differences between means expressed in the actual metric of the measures involved. Standardized mean differences are mean differences divided by a standard deviation and are, therefore, expressed in a *z*-score like

metric. In the Meade framework, standardized mean differences are similar to the popular Cohen's *d* effect size (Meade, 2010, p.730).

Developing DIF/DTF effect sizes in the IRT modeling process proceeds as follows. Once an acceptable GRM model has been fit, item parameters (*a*-parameters, *b*-parameters) are then be used to estimate two sets of scores for respondents that are optimal from the standpoint they are weighted by the model parameter values. The first set of scores are *estimated θ scores* which are expressed in a standard normal metric. A second set of scores, called *expected scores*, are transformations of the estimated θ scores. The importance of these scores is they are expressed in the original item (0-to-4 scale) and scale (0-to-24 scale) metrics. Group mean differences in expected scores are the basis for effect sizes used in this step of the Meade framework.

Focal group θ estimates play a pivotal role in developing expected scores for group comparisons. These scores are the common metric used to compute both focal group and reference group expected scores. Essentially, focal group item parameters and reference group item parameters are used to estimate expected scores using focal group estimated θ scores. The logic of the process is straightforward; If items do not display DIF, expected scores for each group will be close in value. Mean difference and standardized mean differences effect sizes provide statistical examination of differential functioning and various item and scale plots provide a visual indication of the extent to which differential functioning is present (Meade, 2010, pp. 729-730).

Various effect sizes for race are shown in Table 3. The signed item difference in the sample (SIDS) can be interpreted as the average estimated score difference between the focal group and reference group. A negative sign indicates that the focal group has a lower mean on an item than the reference group. For example, the SIDS = -.195 for item 3 indicates that other race students, on the average, scored .195 points lower than White students with equal estimated θ scores (keep in mind the scale here is the item level 0-to4 scale). The signed designation in SIDS refers to the fact that difference across θ may not consistently favor one group and, therefore, some differences might be both negative and positive values. The unsigned item difference in the sample (UIDS) can be interpreted as the average absolute estimated score difference in the sample between other race students and White across other race group respondents. The UIDS = .195 has the similar interpretation to the SIDS; other race students, on the average, scored .195 lower than White students on the item with equal estimated θ scores. The SIDS and UIDS effect sizes can be similarly interpreted for the gender and family group comparisons.

This difference between race SIDS and UIDS is illustrated in Figure 1 where group expected scores for each item are plotted. When the group curves cross (e.g., item 2), the differences between estimated scores for each group across θ change so that one group may be favored at low levels of θ and then becomes less

favored at higher levels of θ. This pattern is referred to as non-uniform DIF. Uniform DIF, on the other hand, is shown by non-crossing functions (e.g., item 3) where one group is consistently favored over θ. Uniform and non-uniform DIF is indicated by the SIDS and UIDS effect sizes. With uniform DIF, these values are equal (e.g., item 3) since the effects of the form of the DIF is consistent. For non-uniform DIF, SIDS and UIDS will tend to be different, sometimes changing signs depending on the magnitude of the different trace line forms (e.g., item 6).

The ESSD effect size in Table 3 is a standardized mean difference index (Cohen, 1992; Meade, 2010). It is expressed in standard deviation units and can, therefore, have a negative or positive sign. For example, the ESSD = -.339 for item 3 can be interpreted as other race students are .339 standard deviation units below White students with equal estimated θ scores. The ESSD effect size can be similarly interpreted for the gender and family group comparisons. Cohen provided a general framework for interpreting the magnitude of standardized mean differences, e.g., "small" (~.2), "medium" (~.5), and "large" (~.8). Using this framework, all the items had small effect sizes.

Finally, we computed effect sizes at the scale (DTF) level (shown in Table 4). Two measures of DTF are useful. The signed test difference in the sample (STDS) is defined as the difference in the summed scale score expected, on the average, across all focal group sample respondents due to DTF. The expected test score standardized difference (ETSSD) is defined as a Cohen's *d* effect size (Meade, 2010). The race values for each of these measures were: STDS = -.245 and ETSSD = -.073. The STDS value indicates that, on the average, other race students were .245 points lower on the scale score. Keep in mind the scale score range is from 0 to 24 so this difference is a relatively small difference. The ETSSD value indicates that, on the average, the other race group scores .073 standard deviation units below White group scores. The values are illustrated in Figure 2 where the race trace lines are close in form and slightly cross at higher ends of θ.

Further, for gender, the STDS = -.231and ETSSD = -.067. The STDS value indicates that, on the average, male students were .231 points lower on the scale score. The ETSSD value indicated that, on the average, the male group scores were .067 standard deviation units below female group scores. Finally, for family, the STDS = -.105 and ETSSD = -.031. The STDS value indicates that, on the average, other family students were .105 points lower on the scale score. The ETSSD value indicated that, on the average, the other family group scores were .031 standard deviation units below female group scores. Interpreting this effect size in the Cohen's *d* framework, all the ETSSD values would be considered a well below small effect sizes.

Discussion

The primary goal of this article was to describe an IRT-based DIF/DTF analysis of the Academic Motivation Scale in the Community and Youth Collaborative Institute School Experience Surveys (Anderson-Butcher, et al., 2020). Our overall conclusion is that the AMS appears to operate similarly across race, gender, and family groups on both the item level and the scale level. There was evidence that more items showed race differential functioning, but the mean difference and the standardized mean difference effect sizes were small. In addition, mean differences and standardized mean differences for gender and family at the item and scale also were small. This finding is important because it can be the case that small item differential functioning can accumulate across items and have a pronounced impact on the differential functioning of the scale composed of those items. We did not see evidence of accumulated DIF in our DTF assessment.

An important consequence of determining the AMS appears to operate similarly for our study groups is we can have increased confidence results from statistical procedures are not confounded by DIF/DTF. For example, say we use the scale in a school-wide needs assessment and decide to determine if academic motivation differs for male and female students by using a t-test of the difference between means. Say we find a statistically significant difference between the groups indicating females tend to have higher academic motivation scores, on the average, than males. Further, we compute a standardized mean difference effect size and find the difference between groups is both statistically and practically significant. Because of minimal gender DIF/DIF found in this study, we can conclude the academic motivation difference between males and females is a substantive difference, not an artifact of DIF/DIF.

Limitations

As with all studies, our conclusions must be tempered by a few cautions and limitations. First, the study used a convenience sample of 7th grade students. Non-random samples typically place limits on how universally applicable findings from a study can be. That notwithstanding, one of the significant advantages of using IRT methods is unbiased item and scale properties can be obtained from unrepresentative samples (Embretson & Reise, 2000, pp. 23–25). Two IRT properties support this claim. First, the group invariance property holds that the estimated item parameters (slopes and thresholds) are population invariant which means, theoretically, item parameters will be the same (or nearly the same) in different populations. This property is based on the assertion that the values of the item's parameters are a property of the item, not the group who responds to the item (Baker & Kim, 2017, p. 41). The second property—person invariance—asserts a person's standing on a latent trait is independent of the items used to measure it (Baker & Kim, 2017, p. 74). The key concept here is respondents have a location on the construct of interest that influences their responses to items. For

example, we assumed a student's response to each item was a manifestation of that student's underlying perception of his or her academic motivation. Thus, students with low perceptions of academic motivation were more likely to endorse "*Strongly disagree*" or "*Disagree*" response categories than students with higher perceptions of school connectedness.

Second, most IRT-based DIF/DTF analyses compare two groups which is usually accomplished by collapsing categories for multi-category variables. In our study, for example, race category responses were small, thus requiring us to collapse the categories into the generic category of 'other race' students. We did the same collapsing process for family composition. In future studies, there could be an interest to see if academic motivation differs within specific categories of race by ensuring a sufficient number of respondents in each group in the study design and employing methods to examine DIF/DTF with multiple focal groups (Tay et al., 2015, p. 23).

Future Research

Although we suggest our study findings add to the cumulative evidence of the validity of the AMS based on the DIF/DTF findings for race, gender, and family composition, we think a compelling argument for scale validity must be further informed by additional studies. There are other factors and characteristics that should be examined for DIF/DTF. For example, studies have detected academic motivation variability by parent involvement and family background (Usher & Kober, 2012), gender identity (Bugler et al., 2013), sexual preference (Aerts et al., 2015), and cultural attributes and ethnic factors (Isik et al., 2018) all of which should be explored specific to the AMS. In addition, there is evidence that simple word-for-word translations (the AMS is available in Spanish) may not result in DIF/DTF free versions (Chen, 2008; Tran et al., 2019).

In a recent article, Thompson and Frey (2020) stressed the need for school social workers to have access to free, feasible, and *valid* measurement tools (our emphasis added). They argued such tools are a keystone to the proper implementation of evidence-based school social work practice (p.4). We agree with the free, feasible, and valid assertions but would add that a complete validity argument must be informed by studies of differential item and scale functioning. For example, all the scales in The Community and Youth Collaborative Institute School Experience Surveys collection—which are marketed as valid and reliable measures of constructs that are important for school social work and other practitioners—would benefit by further DIF/DTF research.

Data Availability

For readers interested in replicating the above analyses, the data files and the R code are available at the author's GitHub repository: https://github.com/JerryBean46/Academic-Motivation-DIF-DTF.

References

Aerts, S., Van Houtte, M., Dewale, A., Cox, N., & Vincke, J. (2015). School motivation in secondary schools: A survey of LGB and heterosexual students in Flanders. *Youth & Society, 47(3),* 412-437. doi:10.1177/0044118X12467657

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014*). Standards for educational and psychological testing.* Washington, DC: AERA.

Anderson-Butcher, D., Amorose, A., Bates, S.M., Iachini, A.L., Ball, A., & Henderson, T. (2020). Driving school improvement planning with community and youth collaborative institute school experience surveys. *Children & Schools, 42(1),* 7-17. doi:10.1093/cs/cdz028

Anderson-Butcher, D., Amorose, A., Iachini, A.L., & Ball, A. (2013). *Community and Youth Collaborative Institute School Experience Surveys: Middle & High School Student Survey*. Columbus, OH: College of Social Work, The Ohio State University.

Baker, F.B. & Kim, S. (2017). *The basics of item response theory using R.* Cham, Switzerland: Springer International Publishing AG.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57(1),* 289-300. doi:10.2307/2346101

Bugler, M., McGeown, S.P., & St Clair-Thompson, H. (2013). Gender differences in adolescents' academic motivation and classroom behaviour. *Educational Psychology, 35(5),* 541-556. doi:10.1080/01443410.2013.849325

Cai, L. & Monroe, S. (2014). *A New Statistic for Evaluating Item Response Theory Models for Ordinal Data* (CRESST Report 839). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chalmers, P.R. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1-29. doi:10.18637/jss.v048.i06

Chen, F.F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95(5),* 1005-1018. doi:10.1037/a0013193

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112(1),* 155-159. doi:10.1037/0033-2909.112.1.155

Edelen, M.O., Thissen. D., Teresi, T., Kleinman, M., & 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, 44(11), Supplement 3*, S134-S142. doi:10.1097/01.mlr.0000245251.83359.8c.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. New York, NY: Psychology Press.

Gómez-Benito, J., Sireci, S., Padilla, J-L., Hidalgo, M.D., & Benitez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30(1),* 104-109. doi:10.7334/psicothema2017.183

Graham, S., & Hudley, C. (2005). *Race and Ethnicity in the Study of Motivation and Competence.* In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (p. 392–413). Guilford Publications.

Isik, U., El Tahir, O., Meeter, M., Heymans, M.W., Jansma, E.P., Croiset, G., & Kusurkar, R.A., (2018). Factors influencing academic motivation of ethnic minority students: A review. *SAGE Open (April-June),* 1-23. doi:10.1177/21582244018785412

Li, Z., & Zumbo, B.D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica, 30*, 343-370.

Lopez Rivas, G.E., Stark, S., & Chernyshenko, O.S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement, 33(4),* 251-265. doi:10.1177/0146621608321760

Maydeu-Olivares, A. (2015). Evaluating fit in IRT models. In S.P. Reise & D.A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 111-127). New York, NY: Routledge.

Meade, A.W. (2010). A taxonomy of effect size measures for differential function of items and scale. *Journal of Applied Psychology, 95(4),* 728-743. doi:10.1037/a0018966

Meade, A.W., & Wright, N.A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97(5),* 1016-1031. doi:10.1037/a0027934

Nugent, W.R. (2017). Understanding DIF and DTF: Description, methods, and implications for social work research. *Journal of the Society for Social Work & Research, 8(2)*, 305-334. doi:10.1086/691525

R Development Core Team. (2021). R: A language and environment for statistical computing, reference index version 4.1.1. [Computer software]. R Foundation for Statistical Computing.

RStudio Team (2021). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Revelle, W. (2021). psych. Procedures for Personality and Psychological Research. (Version 2.1.6) [Computer software]. Northwestern University, Evanston, IL. https://CRAN.R-Project.org/package=psych

Tay, L., Meade, A.W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18(1)*, 3-46. doi:10.1177/1094428114553062

Thompson, A.M., & Frey, A.J. (2020). Free, feasible, and valid measurement tools for school social workers. *Children & Schools, 42*, 3-6. doi:10.1093/cs/cdz030

Tran, V.T., Nguyen, T.H, & Chan, K.T. (2017). *Developing Cross-Cultural Measurement in Social Work Research and Evaluation (2nd Ed)*. New York: Oxford University Press.

Unick, G. J. & Stone, S. (2010). State of modern measurement approaches in social work research literature. *Social Work Research, 34(2)*, 94-101. doi:10.1093/swr/34.2.94

Urdan, T., & Bruchmann, K. (2018). Examining the academic motivation of a diverse student population: A consideration of methodology. *Educational Psychologist, 53(2)*, 114-130. doi:10.1081/00461520.2018.1440234

Usher, A., & Kober, N. (2012). *What roles do parent involvement, family background, and culture play in student motivation?* [Paper No. 4, Student Motivation—An Overlooked Piece of School Reform Series]. Center on Education Policy, Graduate School of Education and Human Development, George Washington University. https://files.eric.ed.gov/fulltext/ED532667.pdf

Table 1

GRM Model Fit Indexes

Group	*N*	*RMSEA*	*RMSEA 95% CI*	*SRMR*	*CFI*
Race					
White	2220	.069	[.058, .081]	.036	.984
Other Race	1001	.022	[.000, .045]	.026	.998
Gender					
Females	1651	.047	[.033, .061]	.029	.992
Males	1570	.059	[.045, .074]	.034	.986
Family					
Two-parent	1839	.060	[.047, .074]	.033	.987
Other Family	1382	.041	[.025, .057]	.030	.993

Table 2

Results from the Two-stage Likelihood Ratio Tests for DIF for All Groups

	All-others-as anchors model		Anchor-item model	
Item	G^2	*p*	G^2	*BH-p*
Race				
1. Positive attitude towards school	16.82	.005	12.61	.034
2. Made the most of school experiences so far	27.23	<.001	19.36	.003
3. Like the challenges of learning new things in school	44.14	<.001	29.45	<.001
4. Confident in ability to manage school work	7.55	.183	—	—
5. School experience is preparing well for adulthood	11.41	.044	11.46	.043
6. Enjoyed school experience so far	18.82	.003	18.98	.003
Gender				
1. Positive attitude towards school	6.02	.305	—	—
2. Made the most of school experiences so far	3.33	.650	—	—
3. Like the challenges of learning new things in school	18.93	.002	3.66	<.001
4. Confident in ability to manage school work	9.18	.102	—	—
5. School experience is preparing well for adulthood	11.42	.044	16.16	.006
6. Enjoyed school experience so far	9.93	.077	—	—
Family				
1. Positive attitude towards school	4.17	.525	—	—
2. Made the most of school experiences so far	4.37	.497	—	—
3. Like the challenges of learning new things in school	13.06	.023	11.22	.047
4. Confident in ability to manage school work	7.81	.167	—	—
5. School experience is preparing well for adulthood	8.27	.042	—	—
6. Enjoyed school experience so far	4.83	.437	—	—

Table 3

Item-level Effect Sizes for All Groups

Item	*SIDS*	*UIDS*	*ESSD*
Race			
1. Positive attitude towards school	.001	.005	.002
2. Made the most of school experiences so far	.049	.042	.095
3. Like the challenges of learning new things in school	-.195	.195	-.339
4. Confident in ability to manage school work	—	—	—
5. School experience is preparing well for adulthood	-.051	.054	-.089
6. Enjoyed school experience so far	-.050	.058	-.079
Gender			
1. Positive attitude towards school	—	—	—
2. Made the most of school experiences so far	—	—	—
3. Like the challenges of learning new things in school	-.123	.124	-.212
4. Confident in ability to manage school work	—	—	—
5. School experience is preparing well for adulthood	-.108	.108	-.182
6. Enjoyed school experience so far	—	—	—
Family			
1. Positive attitude towards school	—	—	—
2. Made the most of school experiences so far	—	—	—
3. Like the challenges of learning new things in school	-.105	-.105	-.181
4. Confident in ability to manage school work	—	—	—
5. School experience is preparing well for adulthood	—	—	—
6. Enjoyed school experience so far	—	—	—

Note: SIDS = signed item difference in the sample; UIDS = unsigned item difference in the sample; ESSD = expected score standardized difference

Table 4

Scale-level Effect Sizes for All Groups

Scale-level	*STDS*	*ETSSD*
Race	-.246	-.073
Gender	-.231	-.067
Family	-.105	-.031

Note: SIDS = signed test difference in the sample;

ETSSD = expected test (scale) score standardized difference

Figure 1

Item-level expected scores for race groups

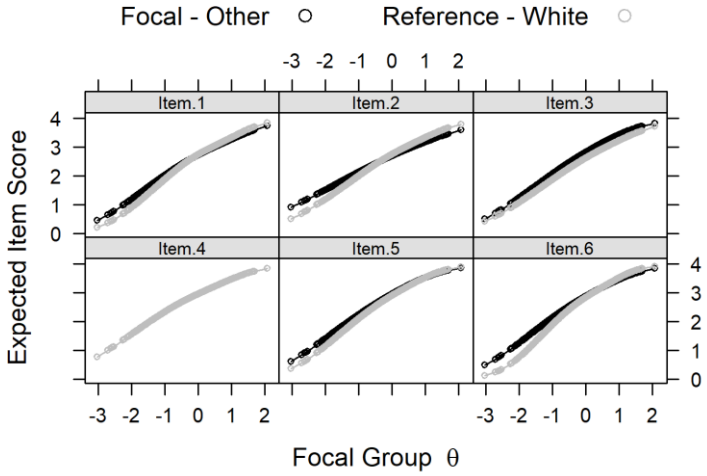

Figure 2

Scale-level expected scores for race groups