Measuring Being Bullied in the Context of Racial and Religious DIF

Michael C. Rodriguez, Kory Vue, José Palma
University of Minnesota
April, 2016

Paper presented at the annual meeting of the
National Council on Measurement in Education, Washington DC.

As we continue to address many persistent challenges in education, particularly those regarding educational equity and, in particular, achievement gaps, educators, community leaders, and youth development researchers have turned their attention to non-cognitive factors in school achievement, also referred to as developmental assets or social-emotional skills. In addition to these developmental skills, some have identified critical developmental supports that must also be in place to secure positive youth development. However, even though we may be able to promote and enhance developmental skills and supports for youth, many continue to face developmental challenges. One of these challenges receiving a great deal of attention is bullying.

The measurement of developmental skills and supports has a relatively recent but rich history, including prominently the work of Search Institute (2005), whose researchers developed tools to measure aspects of their developmental asset profile, with some evidence of common features across diverse communities of youth (Sesma & Roehlkepartain, 2003). The presence of developmental skills, such as positive identity, commitment to learning, and social competence, are developed and reinforced optimally through multiple contexts and sources of support (Scales, Benson, & Mannes, 2006). Others have recognized the difficulty of measuring such skills in diverse populations and across different developmental stages (Griffin, McGaw, & Care, 2012; Kyllonen, 2012).

There are fewer significant attempts to measure the developmental challenges America's youth face on a regular basis. Most measures of risk-taking behaviors and challenging features of family, school, and community contexts are based on single-items in youth surveys. Perhaps with the exception of measures of school climate (see the National School Climate Center, 2015), which have seen more research and development, other measures of contexts like school violence, family violence, or bullying have received less psychometric attention. For decades, youth development researchers have provided evidence of the importance of developmental skills, supports, and challenges for outcomes from cradle to career (Benson, Scales, Hamilton, & Sesma, 2006; Erikson, 1968; Farrington et al., 2012; Lerner et al., 2006). But psychometric work on these tools has lagged significantly.

The Centers for Disease Control and Prevention have had a long-standing interest in violence prevention, as they refer to it, "a major public health problem" (p. 1; CDC, 2015). They released a compendium of assessments measuring bullying, victimization, perpetration, and bystander experiences. This compendium provides a thorough description and research-based discussion of the construct of bullying from three perspectives, victims, perpetrators, and bystanders. The compendium also provides general criteria for evaluating measures, based on criteria recommended by Robinson, Shaver, & Wrightsman (1991). These include ranges of inter-item correlations, coefficient alpha, test-retest reliability, "convergent validity," and "discriminant validity," with four levels of each including minimal, moderate, extensive, and

exemplary. The moderate levels of each criteria include inter-item correlations of .10 to .19, alpha of .60 to .69, test-retest reliability greater than .30 within one to three months, significant correlations with two related measures for convergent evidence and significant differences on one unrelated measure. Unfortunately, these criteria are provided without consideration of the purpose or intended use of the measure – and more notably, there are no criteria regarding evidence supporting the interpretation or use across diverse communities.

Given the importance of these factors and increasing efforts to measure them, the measurement community can play an important role in securing evidence to support their interpretation and use (Kane, 2013). Aside from early attempts to measure attitudes including the work of Thurstone and others in the 1920s (and to some degree perceptions of social contexts), we have not held the measurement of developmental skills, supports, and challenges to the same rigorous standards as measures of achievement. As a growing arena for measurement, rigorous evaluation of measurement quality addressing the validity of score interpretation and use has been limited. "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). This challenge is particularly acute in the use of measures in areas of great need – where graduation rates, achievement levels, and other educational outcomes are disparate among diverse communities. In diverse communities, meaningful and appropriate interpretations must be permissible among diverse groups for scores to be useful to educators and other decision makers. This is critical for score-use fairness, to prevent misuse with marginalized communities and especially with students facing persistent academic challenges.

To support the establishment of a common interpretation framework in diverse communities, some degree of measurement invariance (consistent score quality and meaning) should be confirmed across those communities. In many contexts for the use of developmental skills, supports, and challenges, we are less concerned about testing mean differences, but in addressing the levels of skills, supports, and challenges within each community. We need scores to be appropriate indicators of the levels of the specific trait if we are to make appropriate decisions about the needs of each community – and appropriate decisions about potential interventions. We need scores to reflect levels of the construct being measured rather than irrelevant group differences due to measurement misspecifications (Millsap, 2010; Rupp & Zumbo, 2006; Zumbo, 2007) or construct-irrelevant features (Haladyna & Downing, 2004). The investigation of measurement quality and validity has always been a standard component of test development for most large-scale tests, particularly those with high stakes (AERA, APA, & NCME, 2014). Similar expectations should be held for measures of developmental skills, supports, and challenges.

DIF

Measurement invariance includes a class of methods appropriate for assessing invariance in measurement, addressing the question of whether an instrument is measuring the same trait across subgroups in a population or over measurement conditions. Measurement invariance analysis is often implemented at the level of the total scale via factor analytic methods, for example, through multi-group confirmatory factor analysis. Although a multi-group CFA across relevant subgroups may indicate equivalent factorial structures, item level distortions may still be present (Zumbo & Koh, 2005). Therefore, it is important to conduct item level analyses for

evaluating item-level invariance across subgroups so that we may identify items that could affect score interpretation (Zumbo, 2007).

This study is an examination of a measure of being bullied (Bullied, capitalized when referring explicitly to the measure of being Bullied), an assessment of differential item functioning (DIF). DIF, the extent to which an item functions differently for members of different groups, controlling for differences in group trait levels, was expected in this particular measure because of the inclusion of two group-specific items. One aspect of this Bullied measure, described more fully below, is the perceived reasons for being bullied, including race and religion. Clearly race and religion may not be perceived as relevant for youth in some communities, given the context, and as we report below, is supported by the response patterns of diverse youth. This study includes an attempt to rescale the Bullied measure to accommodate the presence of DIF. The original and rescaled measures are then evaluated regarding mean differences among racial/ethnic communities and correlations with other developmental skills, supports, and challenges, again by racial/ethnic communities.

Methods

Minnesota Student Survey

The Minnesota Student Survey (MSS) is designed by an interagency team from the MN Departments of Education, Health & Human Services, Public Safety, and Corrections to monitor important trends and support planning efforts of the collaborating state agencies and local public school districts, as well as youth serving agencies and organizations. The MSS is administered every three years to students in grades 5, 8, 9, and 11. All operating public school districts are invited to participate. In 2013, the survey was administered to 162,034 students in 312 school districts, including all 87 Minnesota Counties. Students were asked to identify their race and ethnicity, including Hmong, Somali, and Latina/o heritage. A number of Developmental Assets and contextual challenges youth face were identified in subsets of items from the MSS, based on close attention to the Developmental Asset Framework of Search Institute and the more general ecological model of youth development described above. Components of the Developmental Asset Profile (DAP, from Search Institute, 2005) were introduced in 2013.

The measure of interest for this study is the extent to which students experience being bullied in school. *Bullied* measures student experiences as a victim of bullying, including being harassed because of race, religion, gender, sexual orientation, disabilities, physical appearance, through social media, or in person in relational or physical ways. The focus for these questions was on the prior 30 days of school from MSS administration (late-winter).

In the design of the MSS, the intent was to provide several items addressing aspects of bullying in schools. Twelve items addressed reasons for being bullied (because of race, ethnicity, or national origin; religion; gender; being gay or lesbian; physical or mental disability; weight or physical appearance), being bullied through email or social networks, and the frequency of experiencing a number of forms of physical and relational aggression (e.g., being pushed, shoved, slapped, hit, or kicked; threatened; mean rumors or lies; sexual jokes, comments, or gestures; or being excluded from friends, other students, or activities).

MSS Participants

In total, there were 162,034 students included in the 2013 administration of the MSS. For these analyses, only those students in grades 8, 9, and 11 who answered all Bullied items were included. This resulted in a sample size of 114,823 (see grade distribution in Table 1). This included 50% females, 9% with an IEP (receiving special education services), and 26% receiving free/reduced-price lunch.

Table 1
Participating Sample by Race and Grade

Race/Ethnicity	*Grade 8*	*Grade 9*	*Grade 11*	*Total*
American Indian	559	456	256	1271
Asian only (not Hmong)	1182	1096	1028	3306
Black only (not Somali)	1763	1648	1364	4775
Native Hawaiian PI	75	78	68	221
White	29069	29291	26952	85312
Multiple Race or Ethnicity	3278	3272	2101	8651
Latino	2854	2575	1963	7392
Somali	428	320	284	1032
Hmong	778	1032	1053	2863
Total	39986	39768	35069	114823

Confirmatory Factor Analysis (CFA) was conducted on the Bullied measure, which indicates the extent to which the proposed measure fits the observed data (responses). The CFA was completed with Mplus (Muthén & Muthén, 2012), resulting in two forms of evidence: (a) model-data fit information, regarding the consistency of the meaning and stability of the scale as defined by the MSS items and (b) item-factor loadings, which indicates the extent to which each item contributes to the intended measures.

Three measures of model fit provide different aspects of fit, including the root mean-squared error of approximation (RMSEA), the extent to which the model fits reasonably well in the population; comparative fit index (CFI), the relative fit to a more restricted baseline model; and the Tucker-Lewis index (TLI), which compensates for the effect of model complexity. It is generally agreed that multiple indicators of fit should be examined.

Winsteps (Linacre, 2015a) was used to complete DIF analyses of the items, a Rasch model software program. Winsteps employs the Rasch model. Since all of the Bullied items were 5-point rating scale items, a partial-credit model was used for calibration. This allows each item to have its own rating-scale structure, allowing the thresholds to vary in distance from adjacent thresholds, rather than fixing them across items (as in the rating-scale Rasch model).

To complete the DIF analysis, the DIF MEASURE was estimated with Winsteps, which is the difficulty of an item for a given class, with all else held constant. Differences in DIF MEASUREs between groups constitute the DIF CONTRAST and associated standard error, our measure of DIF. The statistical significance of the DIF CONTRAST is tested with a Rasch-Welch *t*-test and associated p-value. These values can be aligned with the commonly interpreted ETS DIF levels, such that B-level DIF is where $0.43 \leq |\text{DIF Contrast}| < 0.64$ indicating slight to

moderate DIF and C-level DIF is where |DIF Constrast| ≥ 0.64 indicating moderate to large DIF (Linacre, 2015b; Zwick, Thayer, & Lewis, 1999).

Once DIF was identified for two items, one regarding the role of Race (item 1) and one regarding the role of Religion (item 2), the measures were recalibrated, allowing these two items to be freely calibrated across racial and ethnic groups. The specific steps to complete that process are described here:

1. Estimate parameters for items 3-12 for all students, resulting in <all_3to12>
2. Estimate parameters for item 1 with White students anchored on <all_3to12>
3. Estimate parameters for item 1 with each race/ethnic group separately, anchored on <all_3to12>
4. Estimate parameters for item 2 with non-Somali students anchored on <all_3to12>
5. Estimate parameters for item 2 with Somali students anchored on <all_3to12>

Once all items were calibrated, where the common items <all_3to12> were calibrated on all students, item 1 (regarding race) was calibrated with each race and ethnic group independently with <all_3to12> fixed, and item 2 (regarding religion) was calibrated with Somali versus non-Somali students independently) with <all_3to12> fixed, persons were calibrated, fixed on the item calibrations relevant to their race/ethnicity. Essentially, the item parameters were estimated through the process described above, and then the measure for students in each group was scored using the items parameters relevant to each group. For comparison purposes, the Bullied measure was also scored through a single simultaneous calibration of all items with all students. This resulted in two version of the measure, Bullied based on full calibration and adjusted version, Bullied-a, based on freely calibrating the race and religion items.

Results

A confirmatory factor analysis model was evaluated with Mplus, employing a unidimensional factor structure for the Bullied measure. Results of the CFA indicated adequate model-data fit. Table 2 includes an abbreviated section of the Mplus output for this model.

Table 2
Mplus Output for a Unidimensional CFA for Bullied

MODEL FIT INFORMATION

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.053	
90 Percent C.I.	0.053	0.054

CFI/TLI

CFI	0.952
TLI	0.941

STANDARDIZED MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
BULLIED BY				
25A	0.600	0.004	159.418	0.000
25B	0.616	0.004	154.909	0.000
25C	0.701	0.004	195.104	0.000
25D	0.740	0.004	192.105	0.000
25E	0.711	0.004	189.866	0.000
25F	0.709	0.002	300.577	0.000
26	0.677	0.003	226.104	0.000
27A	0.724	0.002	293.454	0.000
27B	0.800	0.002	349.502	0.000
27C	0.805	0.002	465.005	0.000
27D	0.703	0.003	242.449	0.000
27E	0.719	0.002	331.082	0.000

Following the CFA, items were calibrated using Winsteps, where a DIF analyses was conducted based on race and ethnicity. All possible group differences were evaluated across five racial groups, three ethnic groups, and a multi-racial/ethnic group.

DIF

Two items in the Bullied measure regarding the role of race and religion exhibited DIF. The item regarding the role of race exhibited DIF between White students and students in all other racial/ethnic groups (Table 3). All DIF contrasts are statistically significant (p<.001) and at or larger than the ETS C-level DIF (|DIF Contrast| ≥ 0.64).

Compared to White students, the item location for the role of race (item 25a) is lower for all other students. When the item location is lower, it is more likely to be endorsed or more salient – more commonly experienced, given the same level of trait overall; that is, students of color identify race as a reason for being bullied at much lower levels of Bullied overall, compared to White students. This suggests that race is a more relevant issue in measuring Bullied for students of color than it is for White students.

Table 3
DIF Results for the Role of Race in the Bullied Measure

Group	DIF Measure	SE(Measure)	DIF Contrast	SE(Contrast)
White (reference group)	0.66	0.01		
American Indian	0.02	0.03	0.64	0.03
Asian	-0.69	0.02	1.36	0.02
Black	-0.48	0.02	1.14	0.02
Native Hawaiian Pacific Isl.	-0.28	0.07	0.95	0.07
Multiple Race/Ethnicity	-0.17	0.01	0.83	0.02
Latino	-0.37	0.01	1.03	0.02
Somali	-0.53	0.03	1.19	0.03
Hmong	-0.56	0.02	1.22	0.03

The item regarding the role of religion exhibited DIF between Somali students (who are Muslim) and students in all other race/ethnic groups (Table 4). All DIF Contrasts were statistically significant (larger than zero, $p<.001$), whereas six were larger than ETS C-level DIF and two were larger than ETS B-level DIF ($0.43 \leq \left| \text{DIF} \right| < 0.64$). Compared to other students, the item location for the role of religion (item 25b) is lower for Somali students. Somali students identify religion as a reason for being bullied at a much lower level of Bullied overall, compared to other students. This suggests that religion is a more relevant issue in measuring Bullied for Somali students that it is for others.

Table 4
DIF Results for the Role of Religion in the Bullied Measure

Group	DIF Measure	SE(Measure)	DIF Contrast	SE(Contrast)
Somali (reference group)	-0.50	0.03		
American Indian	0.44	0.04	-0.94	0.06
Asian	0.01	0.03	-0.51	0.05
Black	0.38	0.03	-0.88	0.04
Native Hawaiian Pacific Isl.	0.15	0.09	-0.65	0.09
White	0.47	0.01	-0.97	0.04
Multiple Race/Ethnicity	0.47	0.02	-0.97	0.04
Latino	0.41	0.03	-0.91	0.04
Hmong	0.07	0.04	-0.57	0.05

Item Responses

In order to investigate these DIF results, we reviewed the item responses for the role of race and religion by student group. Tables 5 (role of race) and 6 (role of religion) contain item response frequencies by group. In both tables, we observe one group with distinctly different response patterns, including White students regarding the role of race (96.1% report this has not been an issue in the last 30 days, compared to 69% to 81% for the other groups) and Somali

students regarding the role of religion (28% report this has been an issue in the past 30 days, compared to less than 10% for other groups; not including Native Hawaiians which is a small group of 221 students).

Table 5
Y25a: During the last 30 days, how often have other students harassed or bullied you for any of the following reasons: Your race, ethnicity or national origin?

	Never	*Once or twice*	*About once a week*	*Several times a week*	*Every day*
A Indian	81.1%	12.8%	2.2%	1.3%	2.6%
Asian	71.7%	19.8%	3.3%	2.8%	2.3%
Black	74.9%	15.9%	2.9%	3.3%	3.0%
N Hawaiian	72.4%	16.7%	5.9%	0.9%	4.1%
White	**96.1%**	**2.6%**	**0.5%**	**0.3%**	**0.4%**
Multiple	79.0%	13.1%	3.0%	2.4%	2.5%
Latino	77.0%	16.4%	2.8%	2.1%	1.8%
Somali	68.7%	20.4%	4.1%	2.5%	4.3%
Hmong	78.0%	17.5%	1.8%	1.3%	1.4%
Total	91.1%	6.0%	1.1%	0.8%	0.9%

Table 6
Y25b: During the last 30 days, how often have other students harassed or bullied you for any of the following reasons: Your religion?

	Never	*Once or twice*	*About once a week*	*Several times a week*	*Every day*
A Indian	90.2%	6.8%	1.5%	0.6%	0.9%
Asian	90.6%	6.5%	1.2%	1.0%	0.8%
Black	92.9%	4.4%	0.9%	0.9%	0.9%
N Hawaiian	88.2%	5.4%	3.6%	0.5%	2.3%
White	93.7%	4.8%	0.7%	0.4%	0.4%
Multiple	91.3%	5.7%	1.5%	0.7%	0.9%
Latino	93.4%	4.8%	0.8%	0.4%	0.6%
Somali	**71.9%**	**17.6%**	**3.0%**	**3.1%**	**4.4%**
Hmong	91.4%	6.7%	0.9%	0.6%	0.5%
Total	93.1%	5.1%	0.8%	0.5%	0.5%

Item Calibration

We then investigated item calibrations to more deeply understand the role of DIF in these two items. Table 7 includes the item locations (b-parameters) for each item based on the calibration of all 12 items versus the calibration of the 10 non-DIF items. We evaluated the stability of item locations (as a measure of the construct) upon removing the two DIF items,

since we hoped to use the remaining items as common items across all groups to fix the scale for Bullied. We observed that the item calibration differences between White students and all students were reduced for 6 of the 10 items after by removing the two DIF items from calibration; the resulting correlation between the locations for the 10 common items was .99 (Table 8). When this was done with all students, item locations did not shift much at all for most items.

Table 7
Item Calibrations (measures) including All Items and Items 3-12 Only, for White Students and All Students

	Items 1-12		Items 3-12 only	
Item	White students	All students	White students	All students
1	0.44	0.09		
2	0.47	0.36		
3	0.40	0.43	0.52	0.47
4	0.12	0.16	0.22	0.21
5	0.35	0.36	0.46	0.46
6	-0.43	-0.42	-0.35	-0.33
7	0.16	0.13	0.26	0.24
8	0.03	0.03	0.07	0.04
9	0.14	0.18	0.24	0.22
10	-0.46	-0.39	-0.38	-0.38
11	-0.65	-0.54	-0.58	-0.54
12	-0.53	-0.40	-0.45	-0.39

To provide an index of stability of item parameters across these different calibration groups, correlations of item parameters are summarized in Table 8. We find that when all items (1-12) are compared between White and all students, the lowest correlation results (.96); item 1 is likely creating trouble here. It appeared that the remaining 10 items could be safely used to identify the scale location for Bullied.

Table 8
Correlations among Item Parameters from Table 7

	White, items 1-12	*White, items 3-12*	*All, items 1-12*
White, items 3-12	.999		
All, items 1-12	.960	.996	
All, items 3-12	.993	.996	.994

We then reviewed the item calibrations for items 1 (race) and 2 (religion) by group, independently calibrated with items 3-12 fixed based on concurrent calibration with all students. Table 9 contains the item locations (Measure or logit value) and standard errors for both items by student group. These results differ slightly from the DIF Contrast results generated by the

simultaneous calibration in Winsteps, because these results consider the calibration of the two DIF items independently, one at a time (rather than all 12 items simultaneously).

Table 9
Items 1 and 2 measures (difficulties), anchored on items 3-12 from all students

Item 1 (race) and group	Measure	SE
White	0.54	0.01
American Indian	0.12	0.04
Asian	-0.65	0.03
Black	-0.38	0.02
Native Hawaiian Pacific Islander	-0.26	0.08
Multiple Race/Ethnicity	-0.07	0.02
Latino	-0.20	0.02
Somali	-0.57	0.04
Hmong	-0.37	0.03
Item 2 (religion) and group		
Somali	-0.58	0.04
NonSomali	0.50	0.01

Figure 1 contains the item map for the item location (item difficulty) of the 10 concurrently calibrated items (located in the boxes) and the specific student community location for the items freely calibrated including Race and Religion. Here we observe that being bullied because of weight or being bullied in ways regarding sexual jokes, rumors, or through social exclusion are the most common among all students – less severe, since they are located at the lower range of the logit scale. On the other hand, the two items that are the most severe or least common are being bullied because of a disability or because of gender, as they are located at the highest level on the logit scale. When we locate Race on the same logit scale based on the item's group-specific location, we see it is a rare (more severe) item for White students and a much more common item for students of color, particularly Black, Somali, and Asian students. Similarly, we see a stark differences in the severity of the Religion item for nonSomali students (for whom the item is much more rare or severe) compared to Somali students (for whom the item is much more common or less severe). It is interesting to note the location of both the Race and Religion items, as their group specific locations are similar for White students (regarding Race) and nonSomali students (regarding Religion), with a common location for Somali students on both items.

Another way to interpret these results from the item map is to say that being bullied because of Race for White students is comparable to being bullied because of Religion for nonSomali students, which is comparable to being bullied because of disability or gender for all students. In the same way, being bullied because of Race for Black, Somali, and Asian students is comparable to being bullied because of Religion for Somali students, which is also comparable

to being bullied because of weight, or through rumors, sexual jokes, or social exclusion for all students.

We may want to consider being bullied because of Race or Religion to be more severe, like being bullied because of disability or gender, but the measurement model does not support this intention. The severity of the reason for being bullied is a function of its frequency in the response patterns of students (more on this in the discussion).

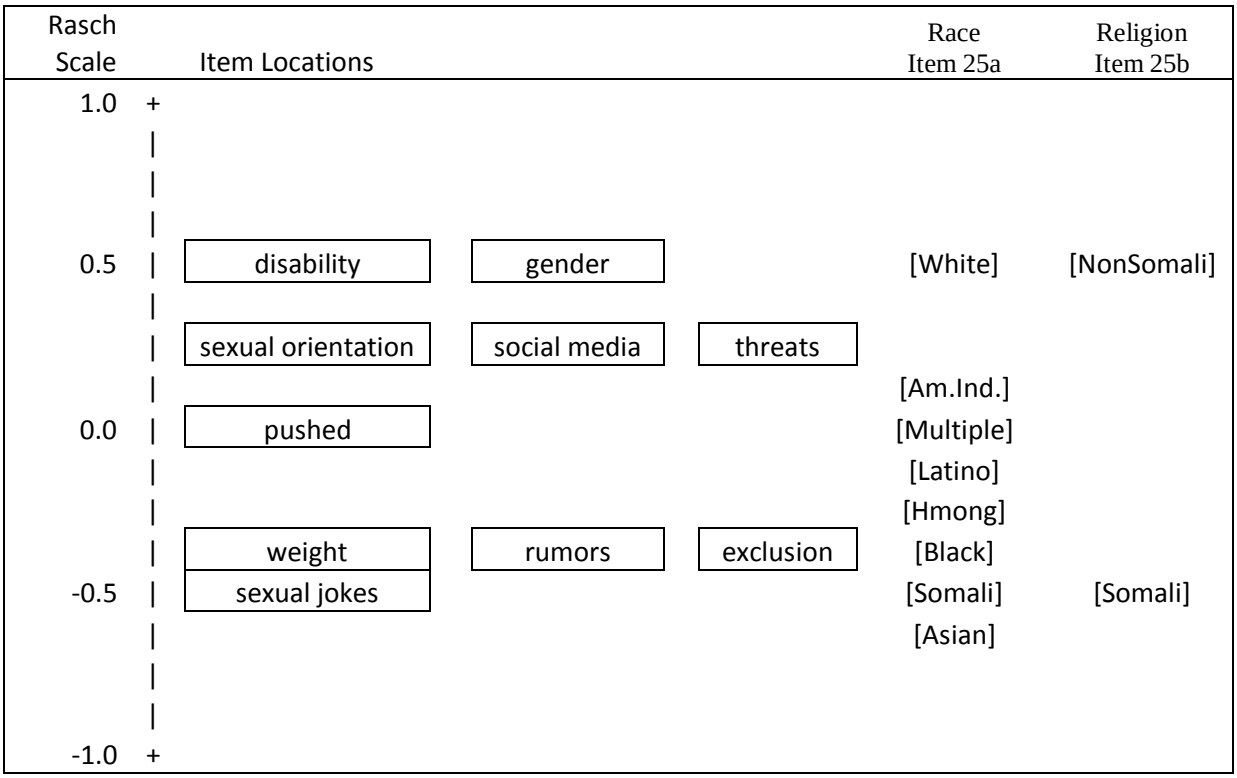

Figure 1. Item map of reasons for and methods of being bullied, with group-specific locations for Race (item 25a) and Religion (item 25b).

Once all of the item calibrations were obtained and used to score persons in each group, person scores could be compared between the full simultaneous calibration for Bullied and the fixed common-item calibration with the race and religion items freely calibrated for Bullied-a. Figure 2 contains the scatter plot for the person scores based on the two measures. We observe very little departures in person scores, except toward the lower end of the scale (persons bullied at lower levels).

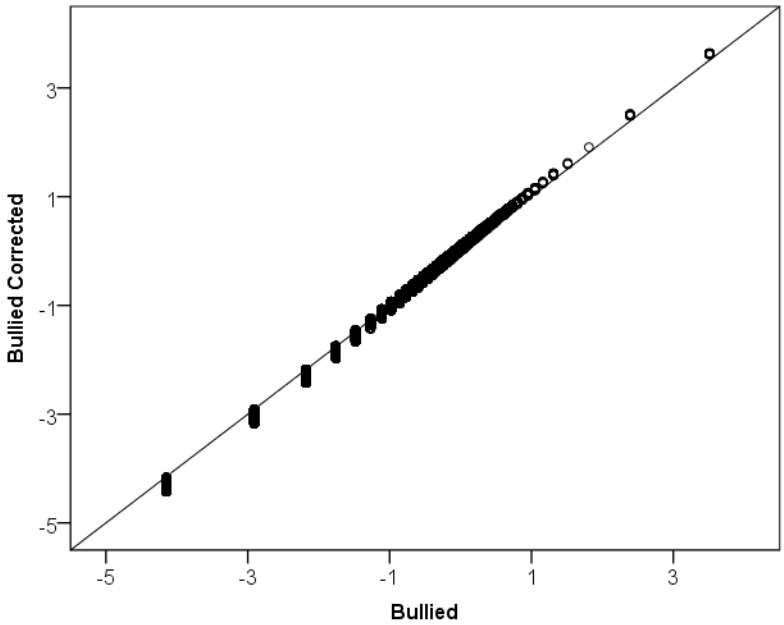

Figure 2. Association between original Bullied scale and adjusted Bullied-a scale.

Associated with each score for each version of the Bullied measure is the standard error of measurement. To evaluate the change in the SEM for each student between the two versions of the measure, we computed a difference in SEM, based on the Bullied original scale minus the Bullied-a adjusted scale. On average, the differences in SEM were negligible, all 0.01 or less.

Perhaps more relevant to our purposes is to evaluate the extent to which correlations with other variables are impacted, as a source of criterion-related validity evidence. We correlated the scores from the two versions of the Bullied measure with several other measures of developmental skills, supports, and challenges (all have been evaluated for DIF and exhibited no items with ETS C-level DIF) for each racial/ethnic group. As can be seen in Table 10, changes in correlations with other variables were small, but most of the 88 correlations were larger in magnitude (more negative or more positive); 72 of the 88 correlations increased in absolute magnitude by at least .001. The largest changes, approaching .01, were for correlations with School Violence.

Table 10a

Correlations between Bullied Original and Adjusted Scales with other Developmental Skills, Supports, and Challenges, by Race/Ethnicity

	AmericanIndian		Asian		Black		White	
Other Variables	Orig	Adj	Orig	Adj	Orig	Adj	Orig	Adj
Bullying	.435	.435	.466	.467	.447	.448	.457	.458
Commitment to learning	-.136	-.136	-.211	-.212	-.149	-.149	-.202	-.202
Empowerment	-.275	-.276	-.288	-.291	-.262	-.264	-.343	-.343
Positive identity	-.209	-.209	-.236	-.237	-.211	-.211	-.301	-.301
Family/community support	-.191	-.192	-.278	-.281	-.181	-.184	-.295	-.295
Social competence	-.135	-.135	-.219	-.221	-.166	-.167	-.237	-.237
Teacher/school support	-.200	-.201	-.277	-.282	-.193	-.196	-.270	-.270
Family violence	.309	.310	.292	.295	.276	.278	.314	.315
Mental distress	.392	.392	.401	.401	.345	.346	.456	.457
School violence	.379	.381	.340	.344	.361	.364	.355	.356
Grades on 4 pt scale	-.109	-.110	-.159	-.160	-.053	-.054	-.149	-.150

	Multiple		Latino		Somali		Hmong	
Other Variables	Orig	Adj	Orig	Adj	Orig	Adj	Orig	Adj
Bullying	.445	.446	.429	.430	.409	.410	.472	.474
Commitment to .earning	-.189	-.190	-.171	-.172	-.240	-.243	-.148	-.149
Empowerment	-.339	-.340	-.280	-.282	-.350	-.353	-.171	-.172
Positive identity	-.277	-.278	-.239	-.240	-.249	-.250	-.141	-.141
Family/community support	-.278	-.279	-.231	-.233	-.265	-.270	-.157	-.160
Social competence	-.221	-.222	-.174	-.175	-.216	-.217	-.117	-.118
Teacher/school support	-.277	-.278	-.225	-.228	-.274	-.281	-.177	-.181
Family violence	.338	.340	.317	.319	.271	.276	.290	.292
Mental distress	.432	.433	.417	.418	.398	.401	.354	.355
School violence	.398	.401	.353	.356	.439	.447	.389	.394
Grades on 4 pt scale	-.121	-.122	-.072	-.073	-.170	-.173	-.098	-.099

Note: Orig=Original Bullied measure; Adj=Adjusted Bullied measure.
All correlations are significant at $p<.001$.

Before leaving this comprehensive review of the effect of adjusting a scale for DIF, we examined the means and SDs for each version of the Bullied measure for each group of students. Table 11 contains summary statistics for both versions by group.

Table 11
Means and SDs for Bullied Original and Adjusted Scales

	Bullied		Bullied-a		
	M	*SD*	*M*	*SD*	*Difference (M)*
American Indian	-2.56	1.42	-2.59	1.45	0.03
Asian	-2.94	1.31	-3.15	1.37	0.21
Black	-2.79	1.36	-2.90	1.39	0.11
Native Hawaiian PI	-2.53	1.46	-2.61	1.49	0.08
White	-2.91	1.31	-2.91	1.34	0.00
Multiple	-2.50	1.41	-2.55	1.43	0.05
Latino	-2.75	1.35	-2.85	1.39	0.10
Somali	-2.78	1.47	-3.00	1.54	0.22
Hmong	-3.06	1.27	-3.22	1.32	0.16
Total	-2.86	1.33	-2.89	1.36	0.03

It is interesting to note that the adjustment to the Bullied measure (by freely calibrating the race and religion items) resulted in no mean change in Bullied for White students (the largest group with n = 85,312). However, for every other group, the Bullied-a measure resulted in lower scores for being bullied. The largest differences are observed for Asian, Black, Somali, and Hmong students. Overall, we observe that no change to White student scores is likely due to the limited reporting of being bullied because of race or religion by White students – no matter how these items are calibrated, very low response rates results in similar means with such a large group. Similarly, the higher scores of students of color on the original Bullied measure is due to the inflation due to the higher measure value of the two DIF items for White students, again, the largest group; whereas the lower scores on the adjusted Bullied-a scores is due to the elimination of this inflation factor when race and religion are freely calibrated.

Discussion

These results make conceptual and empirical sense, but perhaps are not intuitive; that is, the results are not consistent with the belief that being bullied because of race or religion are "severe" reasons, likely with dramatically negative effects on victims. Race is likely to only be a factor when one's race is different than that of a majority of others or different than the group with more privilege or power. Similarly, religion is more likely to be a factor when it differs substantially from the religion of others. Minnesota, as other states, continues to experience dramatic shifts in demographics. Because of the significant refugee resettlement efforts with the Hmong community (beginning in the 1990s) and more recently with the Somali community, and significant immigration from Latin America, the experiences and contexts for these communities are increasingly important for the success of not only their youth, but of all youth.

In creating measures of being bullied (victimization), information can be provided to schools and districts regarding variation in being bullied among different student communities.

This information can directly inform program and policy efforts to target bullying behaviors (reasons and methods). Thus, the information about the magnitude of bullying is clearly important and relevant. In utilizing information from the Minnesota Student Survey, we were able to create a strong and stable measure of being bullied (CFA results were supportive). But to provide evidence of measurement invariance across students from different communities, DIF analyses results indicated significant variation in item functioning for the two items regarding the role of Race and Religion, not entirely unexpected.

To accommodate DIF in the final measure, these two items were freely calibrated for groups expressing DIF, while fixing item parameters on all other items based on concurrent calibration across all groups. This resulted in very little differences in observed patterns of scores and correlations with other relevant variables.

However, resulting levels of being bullied were reduced for the groups experiencing these items. This is somewhat unintuitive. We observe that students of color are more likely to report to be bullied because of their Race than White students; Somali students are more likely to report being bullied because of Religion than other students (Tables 5 & 6). Because these reasons for being bullied are more common for these student communities, it reduces the severity or difficulty of the item in terms of the logit (probabilistic or log-odds) scale. The item becomes easier to endorse and thus indicates a lower-level of being bullied, particularly when compared to those students to whom the item is less likely to be relevant. For White students, few report to be bullied because of Race, making Race an item that requires a higher level of being bullied to be endorsed by White students – thus making it more severe or more difficult. So if Race is a reason for being bullied for White students, it indicates a higher level of being bullied; whereas if Race is a reason for being bullied for students of color, it indicates a lower level of being bullied.

This is the unintuitive interpretation. Most of us would agree that being bullied because of Race or Religion are particularly egregious reasons, particularly compared to reasons such as gender, weight, disability, or physical appearance (although no one should be bullied for any of these reasons). Some might argue that if a persons is bullied because of their Race, that should get more "weight" or count as a more severe form of being bullied. However, in the context of IRT (Rasch), reasons that are more common are scaled as less severe or likely to occur given lower levels of being bullied overall, even if they may be considered to be more egregious.

The presence of DIF for the Race and Religion items is not surprising. In more typical testing applications, if a knowledge test item displayed significant (C-level) DIF, it might be eliminated or replaced. But in the measure of being bullied, we would not want to eliminate important and relevant reasons for being bullied like Race and Religion because of DIF. So we freely calibrated them (while fixing all other items) so they can inform the Bullied measure for each group as the item is relevant (from group-specific calibrations).

The alternative, to treat each reason equally across all groups, appears to be inappropriate based on the Rasch measurement model; items are not located in the same position (Figure 1), indicating that some are indicative of a higher level of being bullied than others. In a simpler approach, we could simply count reasons or more accurately, add up the points in the 5-point rating scale for each item. In this way, every additional increase in frequency of being bullied is treated the same for each reason across each student community. Adding up points and counting reasons isn't measurement, since it ignores the latent variable information-value of each reason and each increment across points in the rating scale (thresholds). But it seems more intuitive to count each reason equally.

References

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

Benson, P.L., Scales, P.C., Hamilton, S.F., & Sesma, A. (2006). Positive youth development: Theory, research, and applications. In W. Damon & R.M. Lerner (Eds.), *Handbook of Child Psychology: Vol. 1* (6th ed., pp. 894-941). New York, NY: John Wiley & Sons.

Centers for Disease Control and Prevention. (2015). Injury Prevention & Control: Division of Violence Prevention. Atlanta, GA: Author. Retrieved from http://www.cdc.gov/violenceprevention/

Erikson, E.H. (1968). *Identity, youth and crisis*. New York, NY: Norton.

Farrington, C.A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T.S., Johnson, D.W., & Beechum, N.O. (2012). *Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review*. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hamburger M.E., Basile, K.C., Vivolo, A.M. (2011). *Measuring Bullying Victimization, Perpetration, and Bystander Experiences: A Compendium of Assessment Tools*. Atlanta, GA: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Retrieved from http://www.cdc.gov/violenceprevention/pdf/BullyCompendiumBk-a.pdf

Kyllonen, P.C. (2012). Measurement of 21st century skills in the Common Core State Standards. Princeton, NJ: K-12 Center, Educational Testing Service. Retrieved from http://www.k12center.org/rsc/pdf/session5-kyllonen-paper-tea2012.pdf

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Lerner, R.M., Almerigi, J.B., Theokas, C., & Lerner, J.V. (2005). Positive Youth Development. *Journal of Early Adolescence, 25*(1), 10–16.

Linacre, J.M. (2015a). Winsteps® (Version 3.90.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/

Linacre, J.M. (2015b). Winsteps® Rasch measurement computer program user's guide. Beaverton, Oregon: Winsteps.com.

Millsap, R.E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4*(1), 5-9.

Muthén, L.K., & Muthén, B.O. (2010). *Mplus 6*. Los Angeles, CA: Muthén and Muthén.

National Research Council. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: The National Academies Press.

National School Climate Center. (2015). *Measuring school climate*. New York, NY: Author. Retrieved from http://www.schoolclimate.org/climate/practice.php

Robinson, J.P., Shaver, P.R., & Wrightsman, L.S. (1991). *Measures of personality and social psychological attitudes*. San Diego, CA: Academic Press.

Rupp, A.A., & Zumbo, B.D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84.

Scales, P.C., Benson, P.L., & Mannes, M. (2006). The contribution to adolescent well-being made by nonfamily adults: An examination of developmental assets as contexts and processes. *Journal of Community Psychology, 34*(4), 401-413.

Search Institute. (2005). *Developmental Assets Profile technical manual*. Minneapolis, MN: Author.

Sesma, A., Jr., & Roehlkepartain, E.C. (2003). Unique strengths, shared strengths: developmental assets among youth of color. *Search Institute Insights & Evidence, 1*(2). Retrieved from http://www.search-institute.org/research/insightsevidence/november-2003

Zumbo, B.D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Zumbo, B.D., & Koh, K.H. (2005). Manifestation of differences in item-level characteristics in scale-level measurement invariance tests of multi-group confirmatory factor analyses. *Journal of Modern Applied Statistical Methods, 4*(1), 275-282.

Zwick, R., Thayer, D.T., Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. . *Journal of Educational Measurement, 36*(1), 1-28