Article Goal

The primary goal of this article is to describe a psychometric analysis of the School Success Profile 2020 (SSP 2020) Microthreat Scale using advanced methods based on modern measurement theory (MMT). Evidence-based school social work practice requires measurement tools that are, themselves, evidence-based (LeCroy, 2019; Thompson & Frey, 2019; Unick & Stone, 2010). MMT methods provide a comprehensive framework for establishing the psychometric viability of a scale. These methods facilitate a detailed examination of a measurement instrument at both the item level and the scale level and help to substantively add to the argument-based measurement validity of a test or scale (American Educational Research Association et al., 2014).

Introduction

We live in contentious times—many of us see and hear a steady stream of divisive political and social discourse daily. While there are many adverse consequences of this divisiveness, perhaps one of the most unfortunate of these is the impact it has on our young people. Adults serve as role models for children and adolescents. The more adults act on their negative opinions and beliefs, the more likely it is that children will see that behavior and think it is acceptable. Interestingly, not all of this bad behavior is necessarily overt or obvious—it can manifest itself in quietly insidious ways in the course of social interactions. One manifestation is through everyday slights, insults, and offensive behavior people—often people from marginalized groups—regularly experience that can cause real psychological damage. Sue (2021) coined an apt descriptor of these negative behaviors. He calls them microaggressions or "death by a thousand cuts" (p.1).

Bowen and Stewart (2021) discussed the critical need for school social work practitioners to have access to measurement tools that can be used to assess and understand microaggressions. They noted microaggressive behaviors can accumulate in a school setting leading to an array of negative impacts on students, including emotional and behavioral difficulties (Banks et al., 2022). Others have also identified negative impacts of microaggressive behavior on school climate, which ultimately results in an overall challenging learning environment for students and teachers (Allen, 2010; Banks et al., 2022; Henfield, 2011). The Microthreat Scale described in this article was designed to provide school social workers and other school staff with a tool that can be used to assess the extent to which microthreats prevail in a school setting. Bowen and Stewart (2021) described and discussed the importance of the Microthreat Scale for use in assessing school settings given the growing prevalence of microthreats in both schools and larger community contexts (see also Keels et al., 2017; Sue, 2021), and the well-established negative effects of experiencing microthreats on such things as self-esteem, depression, anxiety, distress, and physical complaints (Bowen & Stewart, 2021).

The use of the term microthreat for the scale was an intentional shift away from the term microaggression (Bowen & Stewart, 2021) because of the extensive literature and very specific multi-dimensional structure of the microaggression construct and its focus on the school setting. Microthreats are interpersonal words and deeds that are demeaning, insulting, unfair, or physically threatening. They note that "micro" refers to the direct interpersonal nature of the exchange; the term "threat" refers to the challenge the words and deeds present to student's psychological well-being and school success (p. 4072). Further, a strength of the scale is it is not linked to any specific student identity, allowing it to measure the construct as experienced by individuals with different characteristics, identities, or combinations thereof (p. 4073). This is an important assumption—experiencing a microthreat can happen to anyone.

Modern Measurement Theory

School social workers deal with constructs (e.g., school connectedness, academic self-efficacy, bullying, depression, anxiety, etc.). Typically, constructs are defined as abstract concepts that cannot be directly observed or measured; they are usually inferred from responses to multiple survey or interview items designed to be representations of a construct of interest. For example, in this study we considered perceived microthreats to be the focal construct, with student responses to each of the nine scale items to be empirical manifestations of those perceived microthreats.

In the measurement world, constructs are referred to as latent traits. Latent traits are thought to be stable and persistent characteristics that will always be present if a person has them. Latent refers to the notion that a trait exists, is underlying, and can be inferred from behavioral manifestations such as, as noted above, responses to a set of items designed to measure the presence of the trait. In our study, we assumed all students had some measure of the latent trait of perceived microthreats. For some students, the amount of perceived microthreats was small—they might not have experienced many microthreatening acts—whereas other students who experienced numerous incidents had larger amounts of perceived microthreats. One important detail to keep in mind is that when we refer to theta (θ) in the following discussions, we are referring to the latent trait of perceived microthreats (these terms are used interchangeably).

Modern measurement theory evolved as a collection of models designed to study the relationship between measured variables and latent traits. It was specifically developed as a response to the limitations of classical test theory (CTT) in the study of academic testing data and has been widely adapted for use in the examination of psychological and behavioral measurement data (Edwards, 2009; Reeve, n.d.). Houts et al. (2022) note that MMT is particularly helpful in modeling the item-trait relationship where item responses are dichotomous (e.g., yes/no, correct/incorrect, true/false) or polytomous (e.g., such as a five-response ordinal Likert scale). Because most of the data collected for use in school social work measurement studies is either dichotomous or polytomous, MMT methods seem particularly relevant to these studies.

Probably the most dominant MMT methodological framework for studying items with categorical responses (dichotomous or polytomous) is item response theory (IRT). In fact, IRT is frequently referred to as modern measurement theory (Reeve, n.d., Willoughby et. al., 2011). While IRT is considered the standard-bearer for MMT, other latent trait methods are popular and useful. For example, structural equation modeling (SEM) and confirmatory factor analysis (CFA) deal with latent traits (Wirth & Edwards, 2009). There has been an unfortunate tendency to pit methods against each other (e.g., CFA vs. IRT) when, in fact, they are often useful when used in tandem in measurement studies. For example, Bean and Bowen (2021) describe how CFA and IRT can be used as complementary approaches in the analysis of a school social work scale. We use the Bean and Bowen perspective in this paper.

Method

Sample

The data used in this article came from a dataset made available by Natasha. K. Bowen at The Ohio State University College of Social Work. The dataset contains responses collected from sixth through ninth grade students in three schools in 2021–22 (N = 785). Students attended three schools in the southeastern region of the United States. Schools were using the SSP 2020 to gain a better understanding of their students' strengths and needs and to guide intervention choices. School 1 was in Tennessee; Schools 2 and 3 were in North Carolina. Sample characteristics for the three schools are presented in Table 1.

For this analysis, we used cases with non-missing responses to the Microthreat Scale items discussed below (N = 680). A small number of cases (< 1%) were missing data on demographic characteristics but were retained for the analysis. As shown in Table 1, there was a slightly larger percentage of boys across the three school, and 6% of students chose a response other than “boy” or “girl” for the gender question. Because one school had students in grades seven through nine, there were fewer students in grades six and nine than in grades seven and eight. Almost all students in all three schools took part in the free and reduced lunch program in their schools. Almost 60% of students were African American; almost 20% were Latino, and about 7% each were White, Other, or More than one race/ethnicity.

Instrument

The Microthreat Scale is composed of nine items listed in Table 2. The scale uses a simple, three option ordinal response set: *Never* (1), *Once or twice* (2), and *More than twice* (3). The summary scale score is a total of the nine items yielding a score of 9 to 27 with higher scores corresponding to higher levels of perceived levels of microthreats.

Table 1

Demographic Characteristics of Students in Three Schools and Overall

	School 1		School 2		School 3		All Students	
	n = 206		*n* = 242		*n* = 232		N = 680	
	n	%	*n*	%	*n*	%	N	%
Gender								
Girl	96	46.8	110	45.6	104	44.8	310	45.7
Boy	91	44.4	117	48.5	119	51.3	327	48.2
Other^	18	8.8	14	5.8	9	3.9	41	6.0
Total	205	100	241	100	232	100	678	100
Grade level								
6	60	29.3	73	30.7	64	27.6	197	29.4
7	61	29.8	78	32.8	98	43.4	237	35.4
8	84	40.5	87	36.1	64	28.3	235[&]	34.8
Total	205	100	238	100	226	100	669	100
Race/ethnicity								
Black only	158	77.1	106	43.8	136	59.1	400	59.1
Latino only	24	11.7	48	19.8	63	27.4	135	19.9
Other[#]	23	11.2	88	36.4	31	13.5	142	21.0
Total	205	100	242	100	230	100	677	100

^Other genders included transgender or non-binary, other, not sure, or prefer not to answer (two schools had only Female, Male, and Prefer not to answer options)

[&] Includes 3 students who specified a higher grade

[#] Other included American Indian, Asian, Native Hawaiian/Pacific Islander, More than one race, and White

Data Analysis Plan

We conducted all MMT analyses using the R statistical computing environment (R Development Core Team, 2024). We used the RStudio integrated development environment (RStudio Team, 2024) to create and run R scripts. The R packages we used for specific analyses are discussed below.

Step 1. Assessment of IRT Assumptions

IRT model testing requires attention to three assumptions: unidimensionality, local independence, and fitting an appropriate model (Paek & Cole, 2020, p. 4).

Unidimensionality refers to the assumption that the items which compose a scale measure something in common, that is, they are influenced by a single, underlying latent trait. We tested the unidimensionality assumption of our scale items using confirmatory factor analysis (CFA) methods as implemented in the R *lavaan* package (Rosseel, 2012). Following recommendations by Shi, Maydeu-Olivares, and Rosseel (2020), we specified an ordinal factor analysis using a polychoric correlation matrix and a robust unweighted least squares fitting function. We assessed model fit using the root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR) fit indexes.

Local independence refers to the assumption items should be uncorrelated after controlling for the effect of the latent trait. Large residuals correlations mean something other than the latent trait is influencing relationships between items. For this analysis, we used the R *mirt* package (Chalmers, 2012) to compute a set of pairwise Cramer's V residual correlation coefficients. These correlations are a product of fitting an appropriate IRT model and using model parameters to calculate a predicted score for each respondent. Residuals are computed as the difference between actual respondent item scores and the item scores predicted by the model. The residuals are then used in the calculations of pairwise item Cramer's V correlations. Thus, the resulting correlations are measures of pairwise relationships conditioned on or controlling for the effects of the underlying trait with large correlations indicating factors other than the trait of interest are influencing the relationship.

Step 2. IRT Model Fitting and Assessment

Once assumptions have been examined, the next step in the process is to fit and assess the appropriate IRT model. For this analysis, we used the R *mirt* package (Chalmers, 2012) specifying a graded response model (GRM), the recommended model for ordered polytomous response data (Hambleton et al., 2010). We examined various indexes to assess model adequacy. For overall model fit, we used M_2*, which is a limited information fit index specifically designed to assess the fit of IRT models for ordinal data (Cai & Henson, 2013). In addition, we used the SRMR to assess adequacy of model fit based on suggestions made by Maydeu-Olivares and Joe (2014). Note that although the model fit indexes discussed here are similar to those presented for the CFA analysis reported above, the M_2*- RMSEA, and SRMR used in this part of the analysis were developed specifically for assessing ordinal IRT models (Cai & Henson, 2013).

Step 3. Scoring

The primary purpose for using a scale is to estimate a score that represents where a respondent is located on the continuum of the measured construct. For example, if conventional scoring is used, a student's summed scale score will fall somewhere between 9 and 27 with a higher score corresponding to a higher level of perceived microthreats. In classical test theory terms, this score is referred to as an observed score or estimated true score for that person. That student also will have an IRT estimated score. This score will be expressed in the θ metric and will usually fall somewhere in the $-3 \leq \theta \leq +3$ range. Like traditional scoring, higher scores on this scale correspond to higher levels of perceived microthreat experiences.

IRT scoring is model-based, meaning that model parameters (slopes and intercepts discussed below) are used to generate estimates of student θ scores (referred to as person parameters). Thus, these scores are weighted, which allows for items that have stronger relationships with θ to have a greater influence on a score than items with weaker relationships. We used a scoring procedure implemented in the R *mirt* package called *expected a posteriori* (EAP) estimation.

A significant strength of IRT modeling is that it provides substantive detail about the precision of θ score estimates through two related components: information and conditional standard errors. Information is a statistical concept that refers to the ability of an individual item and a composite of items to accurately estimate scores on θ (Baker & Kim, 2017, p. 89). Higher levels of information lead to more accurate score estimates. Information is computed at both the item level and at the test or scale level. Item level information clarifies how well each item contributes to score estimation precision. Scale information is the total of the individual information values of items used to form the scale (Baker & Kim, 2017).

Conditional standard errors are measures of the precision of estimates expressed in the θ metric. Conditional refers to the fact that each estimate across the θ scale has an associated standard error term. Information and conditional standard error are mathematical functions of each other. Specifically, a standard error is defined as the reciprocal of the square root of information for any value of θ. Thus, the more information available at any given θ level, the smaller the standard error will be for that score estimate (Baker & Kim, 2017, p. 98).

Finally, DeMars (2010) suggested that scale information and conditional standard errors are more helpful than a traditional reliability index and standard error term typically used for scale analysis because of the detail they provide about estimate precision across a range of θ values. For example, using IRT it is possible to compute conditional reliability function which shows where on θ score estimates are most reliable. It is important to note conditional standard errors are mathematically related to conditional reliability—lower conditional standard errors correspond to higher conditional reliability. This approach to reliability stands in contrast to classical test theory reliability where one measure—Cronbach’s coefficient alpha, for example—is computed and interpreted under the assumption it covers the entire scale score range.

Additional Validity Argument Analysis

As a final step in the scale testing sequence, we examined the relationship between IRT scores on the microthreats measure and scores on selected social environmental and individual adaptation scales from the SSP 2020. To the extent that Microthreat scores are distinct from other social environmental constructs and are predictive of academic and individual adaptation outcomes established in the literature, this step contributes to the validity argument of the scale. We examined bivariate correlations of the IRT microthreat scores with the following scales from the SSP 2020: School Climate, Teacher Support, School Safety, School Behavior, School Engagement, School Belonging, Self-Esteem, and Social Isolation. Structural equation modeling with Mplus 8.6 (Muthén & Muthén, 2017) was used for the analyses. Because of the ordinal response options, a mean and variance adjusted weighted least squares estimator (WLSMV) was used with a polychoric correlation matrix.

Results

Descriptive Statistics

Item response percentages are shown in Table 2. These percentages reflect wide variability in the endorsement of the items. For example, the three most common microthreats experienced by students at least once were Item 2 (“Ignored you when asked a question”) (43.7%), Item 4 (“Thought you did something wrong when you didn’t”) (48.8%), or Item 6 (“Acted surprised when you did something well”) (44.6%). The three least commonly experienced microthreats were Item 3 (“Kept you from taking part in an activity”) (17.6%), Item 8 (“Threatened to hurt you”) (18.7%), or Item 9 (“Pushed, shoved, or hit you”) (27.8%). Of the items in the scale, Item 8 and Item 9 measure more physical microthreats. Most students (N = 79.0%) reported that they experienced at least one microthreat. Cronbach’s coefficient alpha for the scale was acceptable (α = .85).

Table 2

Item Response Percentages (N = 680)

Item	Never	Once or twice	More than twice
How often during the past month did the things below happen? Someone:			
1. Insulted you	67.6	24.6	7.8
2. Ignored you when you asked a question.	56.3	31.0	12.7
3. Kept you from taking part in an activity	82.4	13.2	4.4
4. Thought you did something wrong when you didn’t	51,2	33.4	15.4
5. Treated you unfairly	68.8	21.5	9.7

6. Acted surprised when you did something well	55.4	30.3	14.3
7. Made fun of you or picked on you (in a mean way)	69.9	30.0	10.1
8. Threatened to hurt you	81.3	12.1	6.6
9. Pushed, shoved, or hit you	72.2	19.3	8.5

Testing IRT Assumptions

Dimensionality

Results from our CFA supported a one-factor solution. Values for model fit measures were: χ^2 = 77.244(27), p < .000, RMSEA = .052 (90 % CI [.039, .066]), and unbiased SRMR = .038 (90 % CI [.027, .048]). We assessed a close fit hypothesis for both effect sizes (e.g., H_0: RMSEA ≤ .05 vs. H_1: RMSEA > .05 and SRMR H_0: SRMR ≤ .05 vs. H_1: SRMR > .05 where .05 is the cutoff value for the population RMSEA and .05 is the cutoff value for the population unbiased SRMR, respectively). Following suggestions by Shi et al. (2019), our obtained p = .367 supported the RMSEA close fit hypothesis; the obtained p = .896 for the SRMR also supported a close fit hypothesis (Note: the goal in assessing a close-fitting index is to retain (or fail to reject) the null (H_0) hypothesis). Based on these results, we concluded that the microthreat scale was sufficiently unidimensional.

Table 3 presents unstandardized and standardized factor loadings, 95% confidence intervals for unstandardized loadings, and R^2 values computed from standardized loadings for each item. All loadings were statistically significant at the $p <$ .001 level.

Table 3

CFA parameter estimates

Item	^Factor loadings (*SE*)	95% *CI*s	^^Factor loadings	Item R^2
Someone:				
1. Insulted you	1.00 (.00)	—	.80	.68
2. Ignored you when you asked a question.	.84 (.04)	[.76, .93]	.68	.46
3. Kept you from taking part in an activity	.81 (.05)	[.70, .91]	.65	.42
4. Thought you did something wrong when you didn’t	.90 (.04)	[.82, .98]	.72	.52
5. Treated you unfairly	1.08 (.04)	[1.00, 1.15]	.87	.75
6. Acted surprised when you did something well	.45 (.06)	[.33, .56]	.36	.13

7. Made fun of you or picked on you (in a mean way)	1.10 (.04)	[1.02, 1.18]	.88	.78
8. Threatened to hurt you	1.04 (.04)	[.95, 1.21]	.83	.70
9. Pushed, shoved, or hit you	.93 (.05)	[.84, 1.02]	.75	.56

Note: ^ = unstandardized; ^^ = standardized; *SE* = standard error; *CI* = confidence interval

Local Independence

As noted above, we assessed local independence by examining conditional Cramer's V residual correlations. Summary measures for the 36 pairwise residual correlations were as follows: *min* = .018, *max* = .151, *Mdn* = .071, and *M* = .070. Measurement researchers suggest residual correlations greater than .2 should be flagged as possible violations of local dependence. Using that rule, we determined there was no evidence of violations of local independence.

IRT Model Fitting and Interpretation

We fit the GRM using a full-information marginal maximum likelihood with an expectation-maximization algorithm fitting function (Chalmers, 2012). The obtained fit indexes were as follows: M_2* = 40.015(18), p < .002; M_2*-RMSEA = .042 (95% *CI* [.025, .060]) and SRMR = .044. The M_2*-RMSEA fell below a recommended cutoff of ≤ .06 for an acceptable GRM model fit and the SMSR value fell below the recommended cutoff of ≤ .05 (Maydeu-Olivares & Joe, 2014). Taken together, our fit indexes indicated that the GRM was a plausible model.

The estimated parameters for the model and their standard errors are shown in Table 4. Two types of parameters are estimated for a GRM. The slope parameter (*a*) is a measure of how well an item differentiates respondents with different levels of θ. Larger values, or steeper slopes, are better at differentiating θ values. Slopes are also measures of the strength of the relation between an item and θ with larger slopes indicating stronger relationships. Location parameters (*b*) are interpreted as the value of θ corresponding to a .5 probability of responding at or above that location on an item. There are *m*-1 location parameters where *m* refers to the number of response categories on the response scale. The microthreat scale has three possible responses so there are two location parameters (b_1, b_2) for each item.

For our GRM, the values of the slope parameters ranged from .65 (Item 6) to 3.75 (Item 7). Using the framework for interpreting slope parameters recommended by Baker and Kim (2017, p. 26), we considered our estimates to be moderate to very high in their ability to differentiate respondents with different levels of θ. (Note: The labels are as follows: 0 = *No ability*; .01-.34 = *Very low*; .35-.64 = *Low*; .65-1.34 = *Moderate*; 1.35-1.69 = *High*; > 1.70 = *Very high*).

Table 4
GRM Item Parameter Estimates

Item	*a*	b_1	b_2	*Item Information*
Someone:				
1. Insulted you	2.48 (.24)	.55 (.06)	1.71 (.11)	4.41
2. Ignored you when you asked a question.	1.60 (.14)	.24 (.07)	1.67 (.13)	2.63
3. Kept you from taking part in an activity	1.54 (.17)	1.37 (.12)	2.59 (.24)	2.37
4. Thought you did something wrong when you didn't	1.75 (.15)	.05 (.07)	1.44 (.11)	2.92
5. Treated you unfairly	3.23 (.32)	.56 (.06)	1.46 (.09)	5.65
6. Acted surprised when you did something well	.65 (.10)	.37 (.14)	2.96 (.42)	.90
7. Made fun of you or picked on you (in a mean way)	3.75 (.39)	.58 (.06)	1.39 (.08)	6.59
8. Threatened to hurt you	2.78 (.29)	1.04 (.07)	1.76 (.11)	4.52
9. Pushed, shoved, or hit you	2.08 (.20)	.77 (.07)	1.78 (.113)	3.38

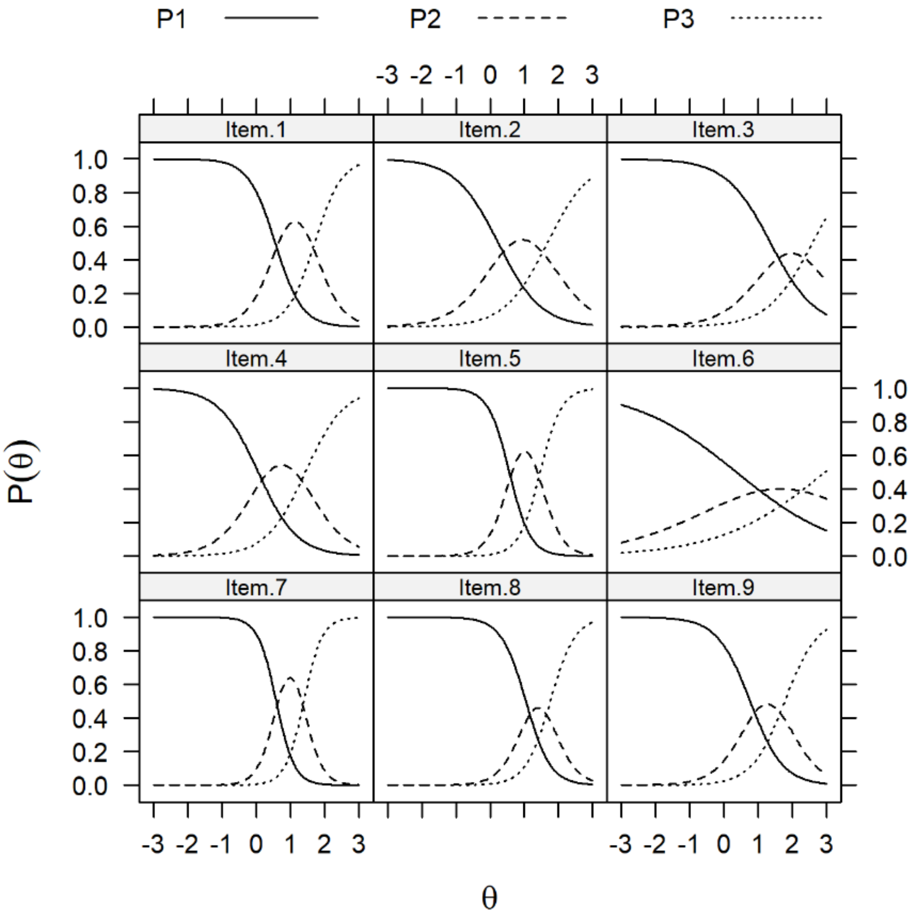
Note: a = item slope; b_i = item location

Estimates of location parameters (b_1, b_2) are also listed for each item in Table 4. Location parameters for Item 1 can be interpreted as follows: b_1 = .51 represents the point on θ where a respondent has a 50% chance of endorsing either the "Once or twice" or the "More than twice" categories; the b_2 = 1.71 parameter represents the point on θ where a respondent has a 50% chance of responding to "More than twice" category. Location parameters provide insights into how levels of the latent trait of perceived microaggression influence responses. For example, it takes a fairly high level of perceived microthreats to have a higher chance of responding "Once or twice" and "More than twice" to Item 3 ("Kept you from taking part in an activity") compared to Item 4 ("Thought you did something wrong when you didn't") as indicated by the lower b_1 = .05 and b_2 = 1.44 values for Item 4.

It also is instructive to visually examine the probabilities of responding to specific categories in an item's response scale. These probabilities are graphically displayed in the category response curves (CRCs) shown in Figure 1. Each trace line represents the probability of endorsing a response category (P1 = *"Never*," P2 = *"Once or twice,"* P3 = *"More than twice*." These curves have a functional relationship with θ: As θ increases, the probability of endorsing a specific category starts to increase and then to decrease as responses transition to the next higher category. For example, at low levels of θ, the P1 trace lines show that the probability of responding "*Never*" was very high. The CRCs then indicated that it took fairly high levels of θ for a respondent to endorse "*Once or* *twice*” on the response scale for most items. Finally, it took very high levels of θ for a respondent to endorse the “*More than twice*” response category. These types of CRC trace lines are not uncommon when items measure low frequency or relatively rare events as illustrated in the frequency distributions shown in Table 2. In addition, the impact of slope parameters was evident in each CRC. For example, the curves for Item 6 (slope = .65) were relatively flat and more spread out horizontally compared to the steeper curves for Item 7 (slope = 3.75).

Figure 1

Category Response Curves



Note: P(θ) = conditional response probability; P1, P2, and P3 are trace lines of category response probabilities for “*Never*”, “*Once or twice*”, “*More than twice*”, respectively.

Scoring

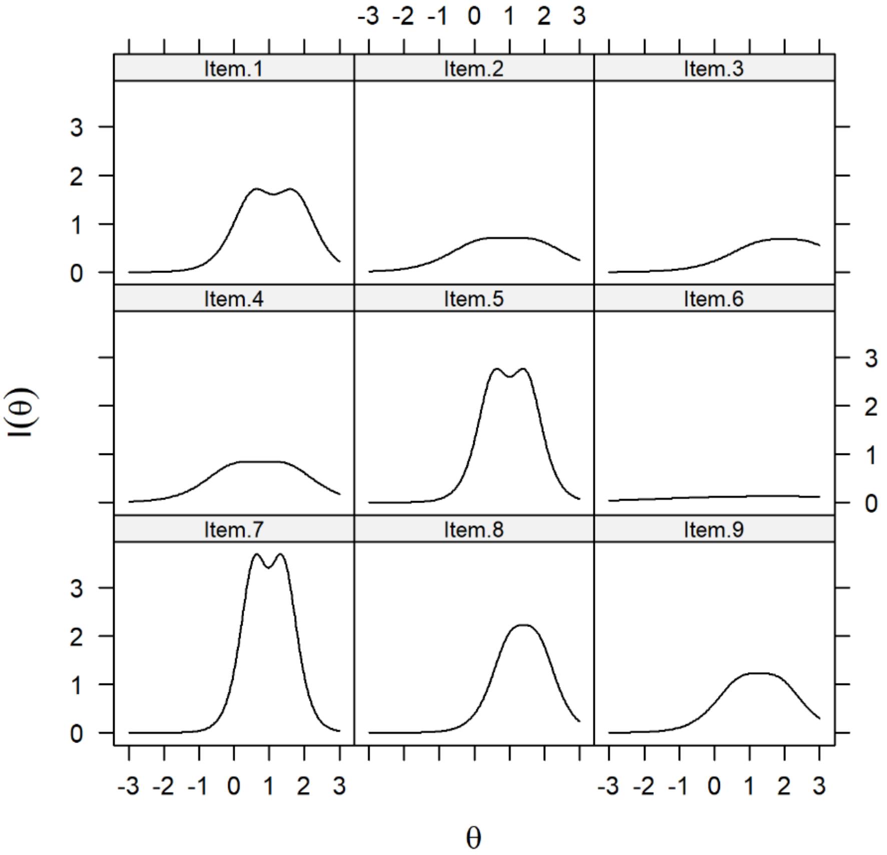
Results for the EAP θ score estimates were as follows: *M* = .00, *SD* = .91, Range = -1.24–2.89. Keep in mind these score estimates are expressed in θ metric (z-scores) so the expected value of the mean of the distribution should equal zero the standard deviation should equal one. However, the EAP estimation process results in a slightly adjusted estimated score distribution with a smaller than expected standard deviation (Brown & Croudace, 2015).

Details about the precision of these estimates are best understood by using a set of plots of item and scale level information, conditional standard errors, and conditional reliability. Starting first at the item level, item information is a function of both item slope and information derived from each response category (DeMars, 2010). In general, the larger the slope and the more distributed the response categories, the higher information is for an item. The relationship between item slopes and item information is shown in Table 4. For example, Item 7 had the highest slope or strongest relationship with θ (slope = 3.75) and had the highest information value (information = 6.59). Item 6, on the other hand, had a low slope value or a weak relationship with θ (slope = .65) and had the lowest information value (information = .90). Model based EAP scoring means that Item 7 was more heavily weighted in the scoring process than Item 6.

Item information functions are graphically displayed using the item information curves (IICs) presented in Figure 2 which are designed to visually display item information along the θ continuum. The curves for Items 5 and 7 indicate they were the most informative items (tall items); the curves for Items 3 and 6 indicated they were the least informative items (shallow items). In addition to visually displaying item information values, IICs display where on the θ continuum items are most informative. Being able to discern where information is highest on θ is one of the strengths of IRT. For example, each item was most informative in the $0 \leq \theta \leq +2$ range.

Figure 2

Item information curves

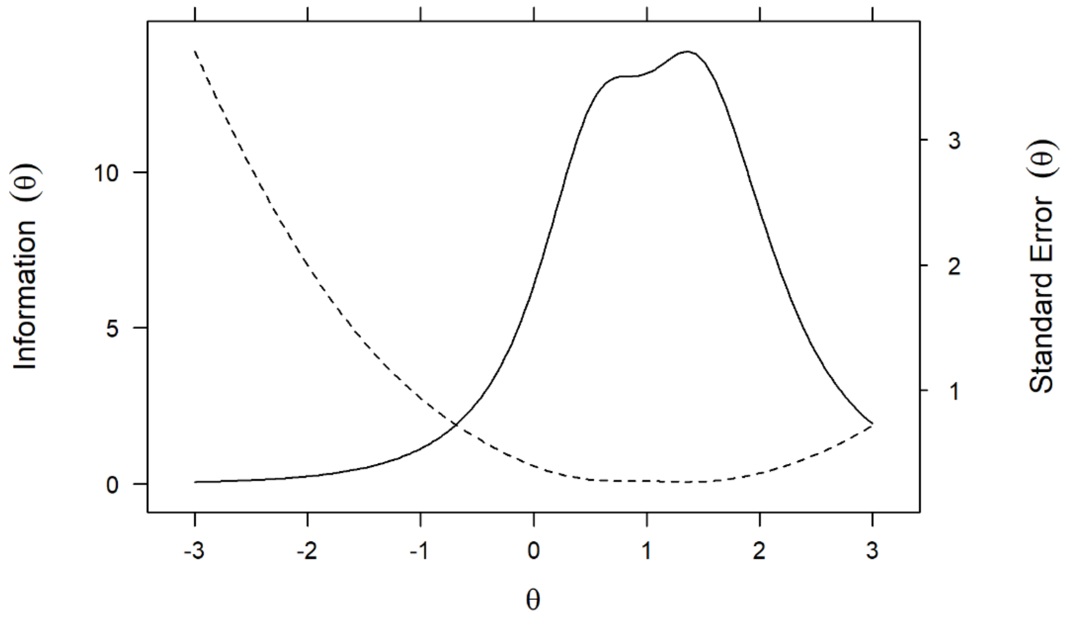


Note: I(θ) = conditional item information; Item 7 ("Someone picked on you or made fun of you") is the most statistically informative item; Item 6 ("Someone acted surprised when you did well") is the least informative.

As noted above, an important feature of IRT is that information for individual items can be summed to form a test (scale) information function (TIF). A TIF provides detail about where on θ the scale works best. Further, also as briefly discussed above, scale information and conditional standard errors are mathematically linked and together provide a picture about how precisely a θ score can be estimated. This relationship is illustrated in Figure 3. The solid line represents the scale information function. Consistent with the item level patterns, the overall scale provided the most information in the range $0 \leq \theta \leq +2$. The dotted line represents the conditional standard errors. This line provides a visual about how estimate precision varies across θ with smaller values corresponding to better estimate precision. Because conditional standard errors mirror the TIF, estimated score precision was best in the $0 \leq \theta \leq +2.0$ range.

Figure 3

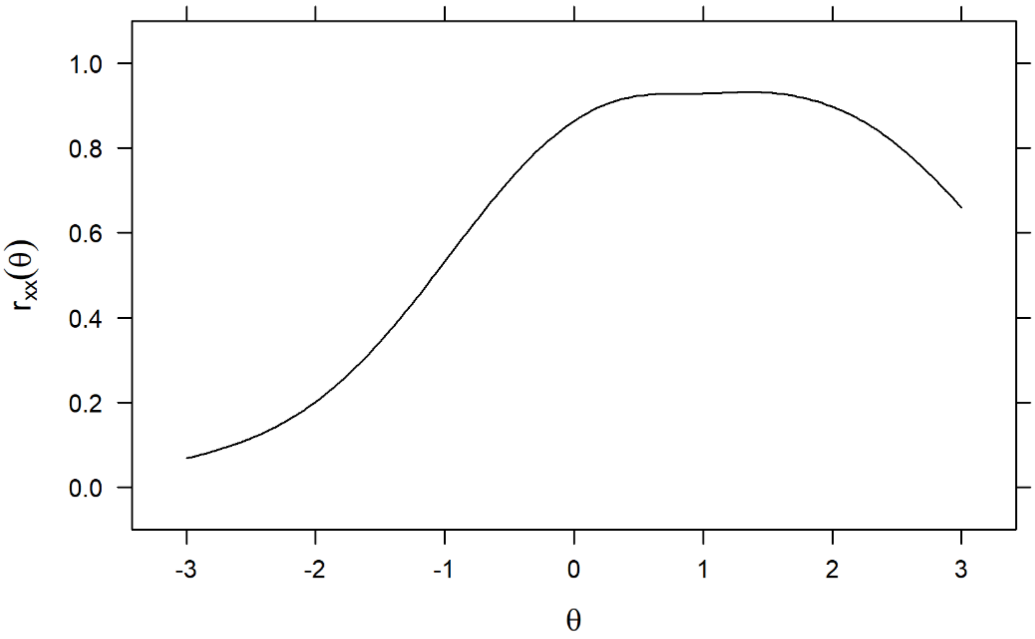
Scale information curve and conditional standard error curve



Note: Solid line Information (θ) = conditional scale information function; dotted line Standard error (θ) = conditional scale standard errors; the curves are mathematical functions of each other where SE(θ) = 1 - √I(θ). Optimal scale precision is in the 0 ≤ θ ≤ 2 range.

In addition to item and scale information and conditional standard errors, it also is possible to compute a conditional reliability function. This function is mathematically related to conditional standard errors where lower standard errors reflect higher precision in score estimates which, in turn, correspond to more reliable estimates. The conditional reliability curve is shown in Figure 4. The curve indicates the scale was quite reliable in estimating scores across the 0 ≤ θ ≤ +2.0 range. Conditional reliabilities in this θ range went from a low of .85 to a high of .89. This is not to say the scores outside this range are not useful, but that scale users should be aware that low scores (θ ≤ 0) and higher scores (θ ≥ +2) may not be estimated with the same reliability.

Figure 4
Conditional Reliability



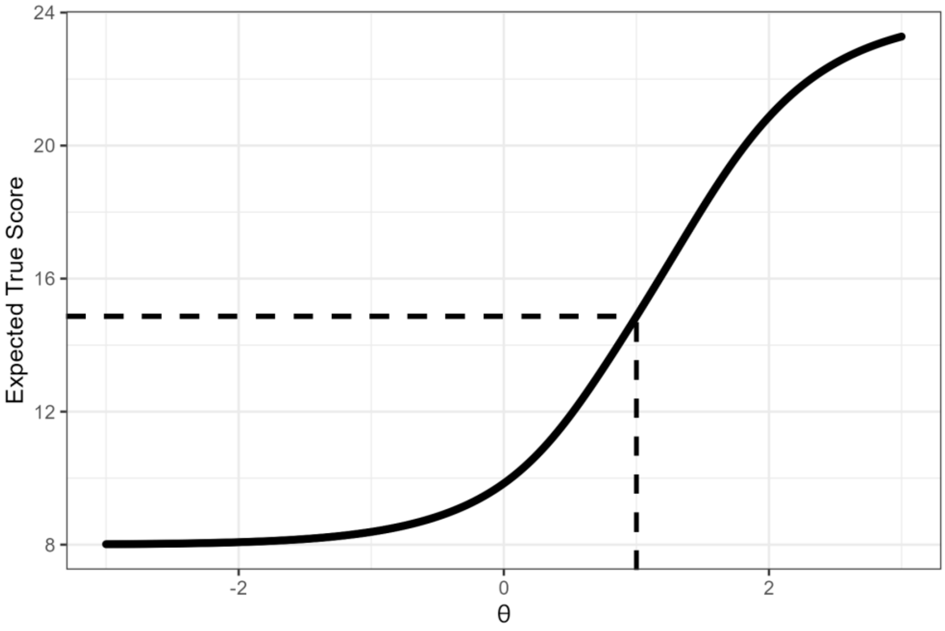
Note: This curve presents scale reliability as a function of θ. It illustrates that the scale is most reliable in estimating scores across the $0 \leq \theta \leq +2$ range.

This approach to reliability stands in contrast to classical test theory reliability where one measure—Cronbach's alpha, for example—is computed and interpreted under the assumption it covers the entire scale score range. However, it is possible to compute a single reliability index in IRT. We computed a version of IRT reliability called empirical reliability, which is recommended for unidimensional models (Brown & Coudace, 2015). The obtained value for the empirical reliability of the scale was .79. just below the obtained coefficient alpha of .85. Brown and Coudace (2015) make the important point that single reliability indices are good summaries of measurement precision only for instruments with relatively uniform estimated θ score distributions (p. 11). For our scale, a single point index of reliability—either empirical reliability or Cronbach's alpha—was not an accurate representation of scale reliability because score estimates were not uniformly distributed— they concentrated and peaked in the upper portion of θ. Our obtained conditional standard errors and conditional reliability were much better indicators of estimate precision and scale reliability.

Finally, once model based θ score estimates are made, it often is of interest to transform those estimates into the original scale metric (a score ranging from 9 to 27). A test characteristic function does just that; it provides a means of transforming θ scores to expected true scores (Baker & Kim, 2017, p. 59; Nugent, 2017, p. 307). The test characteristic function is graphically illustrated in the test characteristic curve (TCC) shown in Figure 5. It has a straightforward use. For any given estimated θ score we can easily find a corresponding expected true score. For example, a θ value of 1 gives rise to expected true scores of 14 (this is expressed in the original scale metric). The TCC also illustrates the fact that it takes fairly high levels of perceived microthreats to generate high expected true scores.

Figure 5

Scale Characteristic Curve Linking Estimated θ Scores and Expected True Scores



Note: The scale characteristic curve linking expected true scores (θ) and estimates of θ. For example, a θ estimate = 1 corresponds to an expected true score = 14. Expected true scores are expressed in the metric of the original scale.

Additional validity argument: Relationship with other SSP 2020 scales

As shown in Table 5, correlations between the IRT Microthreat scores and latent variable scores for selected social environmental and individual adaptation scales from the SSP 2020 ranged from nonsignificant to medium-large (Cohen, 1988). The pattern of correlations indicated the experiences captured with the Microthreat Scale had more of a personal effect on students (i.e., on their behavior and emotional well-being) than on their perceptions of the school environment in general. The SSP 2020 Positive School Climate scale comprises six items about whether respondents believe students are cared about and listened to, and whether students get a good education at the school. Its small negative correlation with school climate highlights that the Microthreat Scale captures a different facet of the school environment than general school climate. Although some authors have noted that microaggressions at school may be associated with negative perceptions of school climate, research on the relationship in secondary schools is minimal and school climate has been measured in different ways (Banks et al., 2022). Two other school environment variables—Teacher Support and School Safety—had nonsignificant relationships with microthreats. Teacher support is a 6-item scale with items about whether teachers care about and respect the responding student. The school safety scale comprises nine items ranging from whether students from different races argue to whether students hit or push teachers. These null findings further suggest students may see the threats as a personal issue, distinct from any failing of the school or teachers. Assessments that focus only on general school climate and teacher-student relationships, therefore, may miss critical information about harmful experiences students at the school may be experiencing.

Table 5

Relationships of IRT Microthreat Scale Scores with Selected SSP 2020 CFA Scale Scores

Construct	Covariance with IRT Scores	Correlation with IRT Scores
School Climate	-.069(.025), *p* = .006	-.**115**(.041), *p* = .005
Teacher Support	-.031(.030), *p* = .303	-.042(.041), *p* = .302
School Safety	-.033(.026), *p* = .207	-.050(.040), *p* = .208
School Behavior	.121(.022), *p* < .000	.**272**(.041), *p* < .001
School Engagement	-.007(.029), *p* = .820	-.010(.046), *p* = .820
School Belonging	-.096(.024), *p* < .001	-.**179**(.041), *p* < .001
Self-Esteem	-.173(.036), *p* < .001	-.**208**(.041), *p* < .001
Social Isolation	.274(.031), *p* < .001	.**416**(.038), *p* < .001

The five additional constructs used to explore the validity and utility of Microthreat Scale scores were measures of individual adaptation: School Behavior, School Engagement, School Belonging, Self-Esteem, and Social Isolation. School Behavior is a 10-item scale with items ranging from skipping a class to getting into a physical fight. School Engagement is a 3-item scale with items about finding school fun and exciting and looking forward to learning new things. Sense of School Belonging is a 5-item scale about how well a student gets along with others and feels like they belong at the school. Self-Esteem is a 5-item scale about self-perceptions of good qualities and satisfaction with self. The Social Isolation scale (previously labeled maladjustment) has six items related to emotional distress. Although the correlation with school engagement was nonsignificant, absolute values of the correlations between microthreat scores and the other four constructs ranged from .179 to .416 and in the expected directions. The largest correlations were with negative behaviors at school (.272) and a sense of social isolation (.416). Our correlational data do not permit a determination of the direction of effects. For example, we cannot determine if a sense of social isolation makes a student more vulnerable to microthreats, or if experiencing microthreats contributes to a sense of social isolation. Regardless of the direction of the effects, the relationship underscores the importance of identifying students with high scores on the Microthreat Scale and finding ways to promote their well-being.

Discussion

Our overall conclusion is the SSP 2020 Microthreat Scale is a potentially valuable measurement tool for use in schools and other settings. The important take-aways from our analyses are summarized as follows:

- Using CFA methods, we determined that the nine Microthreat Scale items measured a latent trait—perceived microthreats—in common.
- Using IRT methods, we determined that a GRM adequately fit the data; item slope parameters which measure the strength of the relationship between an item and the latent trait were acceptable.
- Item-level information varied by item but, in general, all items contributed non-trivial information to the overall scale information function; Item 6 (“Someone acted surprised when you did something well”) was least informative, Item 7 (“Someone made fun of you or picked on you (in a mean way)”) was most informative.
- Scale information and conditional standard errors indicated the scale operated optimally over the $0 \leq \theta \leq +2$ range of the latent trait; information in this range was high, conditional standard errors were low, and θ estimates in this range were most reliable.
- Using estimated θ values, we computed expected true scores which are expressed in the original score metric; these scores are optimal scores from the standpoint that they reflect model-based (and therefore weighted) latent trait estimates; expected true scores improve on standard summed scores.

These findings substantively contribute to the validity argument supporting the use of Microthreat Scale in schools (and other applied settings). In addition, we examined scale correlational relationships with other scales measuring various aspects of student school experiences. Although most of these relationships were modest, they did provide insights into how perceived microthreats covaried with scales measuring school behavior, school belonging, self-esteem, and social isolation. In argument-based validity, the correlations address evidence that microthreats should be considered in future research and efforts to support students’ well-being (American Educational Research Association et al., 2014).

Limitations

As with all studies, our conclusions must be tempered by a few cautions and limitations. First, the study used a convenience sample of 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 that 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 responding 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). For example, in our study, we assumed a student's response to each item was a manifestation of that student's underlying latent trait of perceived microthreats. Students with high perceptions of perceived microthreats were more likely to endorse the "*More than twice*" response category than students with low perceptions of microthreats.

Finally, Nugent (2107) stressed the need to examine differential item functioning and differential test function in social work measurement. He discussed the implications of ignoring item and test differential functioning, the most egregious of which is that a relationship between measures may be more a function of measurement nonequivalence than of substantive differences if differential functioning is present (Nugent, 2017, p. 306). There is a large set of factors and characteristics that potentially could drive differences in how students perceive microthreats: gender, race, socioeconomic status, and cultural heritage are just a few examples (Banks, et al., 2020; Keels et al., 2017). Given the importance of perceived microthreats as a significant contributor to school climate and academic success, this would be an important area for future research (see Bowen & Stuart [2021] for a detailed discussion about microthreats and race). Understanding differential item and scale functioning are key components to establishing strong argument-based validity for the Microthreat Scale (American Educational Research Association et al., 2014).

A Note on Applied Scoring

There are three sets of Microthreat Scale scores that might be of use to an applied practitioner or school social work measurement researcher. The most straightforward scoring procedure is to simply sum the items. As noted, the scale uses a simple, three option ordinal response set: *Never* (1), *Once or twice* (2), or *More than twice* (3). The summary scale score is a total of the nine items yielding a score of 9 to 27 with higher scores corresponding to higher levels of perceived levels of perceived microthreats.

Although we think a sum score approach is viable, scale users should be aware using this approach assumes items are equally weighted, that is, each equally contributes to the total score. We know this assumption for the Microthreat Scale is not strictly met as evidenced by both our CFA and IRT analyses.

IRT model-based scoring makes substantive improvements over summed scoring. In this scoring process (EAP estimation, for example), items are weighted such that items with stronger relationships are given more weight in scoring than items with weaker relationships. Using this procedure, respondents with the same sum score would likely have different estimated EAP scores. This is illustrated in Table 6 where we present five student score patterns. Note the patterns are quite different in item endorsements, but all sum to a score of 12. In conventional sum scoring, these students would have the same score. However, further note that the model-based EAP scores are different for each student (remember the scores are in the θ metric) ranging from -.32 to .50. These score estimates are considerably more variable than sum scores, an attribute that can be beneficial in working with the scores statistically (e.g., exploring relationships with other variables, model-building, etc.). Edwards (2009) noted that IRT model-based scores have another favorable property that improves on a summed score approach—scores are expressed in a standard normal metric. Thus, we can use our knowledge of the standard normal distribution to make comparative decisions. For example, someone with a θ score of one is one standard deviation above average and we can expect that 84% of the sample to have lower scores and 16% to have higher scores (Edwards, 2009, p. 519). Other comparisons of interest based on standard normal characteristics are possible.

Table 6

Variations on Scoring

Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Sum Score	EAP Score	EAP Score SE	EAP Score r_{xx}	Estimated True Score
2	1	1	1	2	1	2	1	1	12	.50	.28	.92	13.6
1	3	1	1	1	2	1	1	1	12	-.32	.47	.78	10.6
1	1	1	2	1	2	1	2	1	12	.07	.35	.88	11.7
2	2	1	2	1	1	1	1	1	12	.33	.33	.89	12.8
1	1	1	3	1	2	1	1	1	12	-.26	.46	.79	10.8

Finally, because not all users will be familiar with or comfortable using a θ score frame of reference, it is possible to transform θ scores into expected scores using a test (scale) characteristic function (Baker & Kim, 2017). As noted above, expected scores are re-expressions of θ into the original summed score metric. They are more precise than conventional summed scores from the standpoint that they use model-based θ scores which are optimally computed (Baker & Kim, 2017). The variability of these scores is shown in the last column in Table 7. These scores could be quite useful in applied use.

There are a few challenges however, when conducting IRT model-based model fitting and scoring. Foremost among the challenges is the fact that specialized software is required. However, for readers interested the open-source R statistical computing environment and a wide array of specialized R packages are freely available (see below for a link to the software used in this analysis). As more free resources are developed by social workers and for social workers (see, for example, https://bookdown.org/bean_jerry/using_r_for_social_work_research/), the training barrier can be reduced.

Replication

For readers interested in replicating the analyses presented in this article, data and R files can be downloaded from: https://github.com/JerryBean46/Microthreat-Scale-Analysis. Follow the directions listed in the README.md document. If you require assistance, contact the first author. For readers interested in using the SSP 2020 in research or in schools, contact the second author.

References

Allen, Q. (2010). Racial microaggressions: The schooling experiences of Black middle-class males in Arizona's secondary schools. *Journal of African American Males in Education*, *1(2)*, 125-143.

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

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

Banks, B.M., Cicciarelli, K.S., & Pavon, J. (2022). It offends us too! An exploratory analysis of high school-based microaggression. *Contemporary School Psychology*, *26*(2), 182–194. doi:10.1007/s40688-020-00300-1

Bean, G.J. & Bowen, N.K. (2021). Item response theory and confirmatory factor analysis: Complementary approaches for scale development. *Journal of Evidence-Based Social Work*, *18(6)*, 597-618. Doi: 10.1080/26408066.2021.1906813

Bowen, N.K., Lucio, R., Patak-Pietrafesa, M. & Bowen, G.L. (2020). The SSP 2020: The revised school success profile. Children & Schools, *42(1)*, 19-28. doi:10.1093/cs/cdaa002

Bowen, N.K. & Stewart, A.E. (2021). Measuring microthreats in middle and high school: A first step toward making schools safe for all students. *Current Psychology*, *40(8)*, 4072-4085. doi:10.1007/s12144-019-00345-3

Brown, A. & Croudace, T. (2015). Scoring and estimating score precision using multidimensional IRT. In Reise, S. P. & Revicki, D. A. (Eds.). *Handbook of item response modeling: applications to typical performance assessment*. New York: Routledge/Taylor & Francis Group.

Cai, L. & Henson, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*, 245-276. doi:10.111/j.2044-8317.2012.02050.x

Chalmers, R. Phillip. (2012). *mirt*: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*, 1-29. URL http://www.jstatsoft.org/v48/i06/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* New York, NY: Lawrence Erlbaum Associates.

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

Edwards, M. C. (2009). An introduction to item response theory using the need for cognition scale. *Social and Personality Psychology Compass*, *3(4)*. https://doi.org/10.1111/j.1751-9004.2009.00194.x

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

Hambleton, R.K., van der Linden, W.J., & Wells, C.S. (2010). IRT models for the analysis of polytomously scored data: Brief and selected history of model building advances. In Nering, M.L. & Ostini, R. (Eds.). *Handbook of Polytomous Item Response Models* (pp. 21-42). New York, NY: Routledge.

Henfield, M. S. (2011). Black male adolescents navigating microaggressions in a traditionally white middle school: A qualitative study. *JMCD Journal of Multicultural Counseling and Development, 39(3)*, 141-155.

Houts, C.R., Savord, A. & Wirth, R.J. (2022). Overview of modern measurement theory and examples of its use to measure execution function in children. *Journal of Pediatric Neuropsychology*, *8*, 1-14. doi:10.1007/s40817-021-00117-7

Keels, M., Durkee, M. & Hope, E. (2017). The psychological and academic costs of school-based racial and ethnic microaggressions. *American Educational Research Journal*, *54(6)*, 1316-1344. doi:10.3102/0002831772220

LeCroy, C. W. (2019). Mismeasurement in social work practice: Building evidence-based practice one measure at a time. *Journal of the Society for Social Work and Research*, *10*(3), 301–318. doi:10.1086/704363

Maydeu-Olivares, A., & Joe, H. (2014). Assessing Approximate Fit in Categorical Data Analysis. *Multivariate Behavioral Research*, *49*, 305-328.   doi: 10.1080/00273171.2014.911075

Muthén, L. K., & Muthén, B. O. (2017). Mplus: Statistical Analysis with Latent Variables: User's Guide (Version 8). Los Angeles, CA: Authors.

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

Paek, I., & Cole, K. (2020). *Using R for item response theory model applications*. New York, NY: Routledge

R Development Core Team. (2024). R: A language and environment for statistical computing, reference index version 4.3.1 [Computer software]. R Foundation for Statistical Computing.

Reeve, B. B. (n.d.). An introduction to modern measurement theory. National Cancer Institute. http://citeseerx.ist.psu.edu.proxy.lib.ohio-state.edu/viewdoc/download?doi=10.1.1.207.3244&rep=rep1&type=pdf. Accessed 29 Nov 2021.

RStudio Team. (2024). *RStudio: Integrated Development Environment for R*. Boston, MA. Retrieved from http://www.rstudio.com/

Shi, D., Maydeu-Olivares, A. & Rosseel, Y. (2020) Assessing fit in ordinal factor analysis models: SRMR vs. RMSEA. *Structural Equation Modeling: A Multidisciplinary Journal, 27(1),* 1-15, doi:10.1080/10705511.2019.1611434

Sue, D.W. (2021). Microaggressions: death by a thousand cuts. *Scientific American.* https://www.scientificamerican.com/article/microaggressions-death-by-a-thousand-cuts/. Retrieved 10/1/23.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48(2)*, 1-36. doi:10.18637/jss.v048.i02

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

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

Willoughby, M.T., Wirth, R.J. & Blair, C.D. (2011). Contributions of modern measurement theory to measuring executive functions in early childhood; An empirical demonstration. *Journal of Experimental Child Psychology*, *108(3)*, 414-435. doi:10.1016/j.jeep.2010.04.007

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12(1)*, 58–79. doi:10.1037/1082-989X.12.1.58