# Assessing Differential Item Functioning in a Teacher Self-assessment of Cultural Responsiveness

Lindsay M. Fallon, Ph.D.[1], Sadie C. Cathcart, M.Ed.[1], Austin H. Johnson, Ph.D.[2]

[1]University of Massachusetts Boston

[2]University of California Riverside

Lindsay M. Fallon https://orcid.org/0000-0003-0813-3337

Sadie C. Cathcart https://orcid.org/0000-0003-0395-657X

Austin H. Johnson https://orcid.org/0000-0002-6349-0049

*Journal of Psychological Assessment* ©2021 Sage Publications



## Author Note

The U.S. Department of Education's Institute of Education Sciences supported this research through Grant R324B170010 to the University of Massachusetts Boston. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Address correspondence to Lindsay M. Fallon, Department of Counseling and School Psychology, 100 Morrissey Boulevard, Boston, MA 02125. Email: lindsay.fallon@umb.edu

**Abstract**

The *Assessment of Culturally and Contextually Relevant Supports* (ACCReS) was developed in response to the need for well-constructed instruments to measure teachers' cultural responsiveness and guide decision-making related to professional development needs. The current study sought to evaluate the presence of differential item functioning (DIF) in ACCReS items and the magnitude of DIF, if detected. With a national sample of 999 grade K-12 teachers in the U.S., we examined measurement invariance of ACCReS items in relation to responses from (a) racially and ethnically minoritized (REM) and White teachers (teacher race), (b) teachers in schools with 0-50% and 51-100% REM youth (student race), and (c) teachers with <1-5 years of teaching experience and teachers with >5 years of experience. Findings suggested that ACCReS items exhibited negligible levels of DIF. The lack of DIF found provides additional evidence for the validity of scores from the ACCReS to assess teachers' cultural responsiveness. Furthermore, descriptive analyses revealed that teachers were more likely to agree with items pertaining to their own classroom practice than items related to access to adequate training and support. Results inform implications for future educational and measurement research.

*Keywords:* cultural responsiveness, teachers, instrument development, classroom assessment

**Assessing Differential Item Functioning in a**

**Teacher Self-assessment of Cultural Responsiveness**

Over the past two decades, the population of U.S. public school students has become increasingly racially and ethnically heterogeneous, yet teachers have remained primarily White (U.S. Department of Education, 2019). Indeed, there has been a decrease in the number of Black teachers in the field, and Hispanic/Latinx teachers make up only about 9% of the teaching force (National Center for Education Statistics, 2016). For racially and ethnically minoritized (REM) youth (e.g., Black/African American, Hispanic/Latinx, American Indian/Alaskan Native; Malone & Ishmail, 2020), instruction from REM teachers is connected to a host of positive outcomes including better academic performance and social emotional wellness, as well as an increased likelihood to attend a college or university (Bates & Glick, 2013; Yarnell & Bohrnstedt, 2017). These findings imply the possibility of a disconnect between White teachers and REM students.

This disconnect (or "mismatch"; La Salle et al., 2020) aligns with teacher perceptions of being underprepared to engage in culturally responsive practices (e.g., Bergeron, 2008), especially teachers who are new to the field (e.g., within their first five years of teaching; Lee et al., 2012). Without targeted training and support, teachers may engage in actions that disadvantage REM youth. Specifically, there is extensive evidence to suggest that exclusionary disciplinary techniques (e.g., office discipline referrals, suspensions, expulsions) are disproportionately applied to Black/African American, Hispanic/Latinx, American Indian/Alaskan Native students (Girvan et al., 2017). Similarly, there has long been evidence of differences in achievement metrics between REM and White students (Hung et al., 2020). The "discipline and achievement gap" (Gregory et al., 2010), perhaps better understood as

opportunity gaps (Miretzky et al., 2016), evidence the need for educators to confront systemic racism in schools (Kohli et al., 2017) informed by comprehensive, ongoing training to establish more equitable and effective learning environments (García et al., 2010).

Assessment of teachers' cultural responsiveness may be an appropriate place to start in the process of identifying specific areas of need for staff intervention (such as training and professional development [PD]). Cultural responsiveness refers to the extent to which educators value students' individual differences (e.g., language, heritage, experiences) and align what and how they teach to students' cultures (Gay, 2018). Currently, few teacher self-assessments of cultural responsiveness exist. Of existing measures, some focus primarily on teachers' instruction, such as the *Culturally Responsive Teaching Self-Efficacy Scale* (α = 0.95; Siwatu, 2007) and the *Multicultural Efficacy Scale* (α = 0.80; Guyton & Wesche, 2005). Other scales focus on behavioral supports, such as the *Culturally Responsive Classroom Management Self-Efficacy Scale* (α = 0.97; Siwatu et al., 2017) and *Double Check Self-Reflection Tool* (α = 0.65; Hershfeldt et al., 2009).

The *Assessment of Culturally and Contextually Relevant Supports* (ACCReS) includes items pertaining to both culturally responsive teaching and behavior supports, as well as teachers' action to engage with students' culture (e.g., collaboration with families), and access to information and systems of support (e.g., relevant data, PD). It was created to be a comprehensive instrument targeting cultural responsiveness aligned with multi-tiered systems of support (MTSS; Authors, 2012b) and has undergone several initial validation procedures (Authors, 2020). ACCReS items and subscales were constructed to reveal teachers' perceptions (represented by scores) of their implementation of empirically-supported culturally relevant practices. Results are intended to be used by teachers, support staff and/or school leaders to

determine areas of relative strengths and weakness, and identify areas for growth to target with teacher professional development, coaching, and intervention efforts.

**Initial Validation of the ACCReS**

ACCReS items were originally derived from a systematic review of the literature related to culturally-relevant classroom practice (Authors, 2012a) and grounded in Vincent and colleagues' (2011) model of cultural responsiveness applied to a MTSS framework. Authors (2020) outline a multi-step process of content validation, and exploratory and confirmatory factor analyses with unique large teacher samples. Analyses resulted in a 35-item instrument assessing teachers' perceptions of their (a) use of *equitable classroom practices* (ECP; ω = .87), (b) *consideration of culture and context* (CCC; ω = .77) in the classroom, and (c) *access to information and support* (AIS; ω = .86) (Authors, 2020). The study also found significant correlations between teachers' responses on the ACCReS and their responses on the *Culturally Responsive Teaching Self-Efficacy Scale* (Siwatu, 2007) and *Culturally Responsive Classroom Management Self-Efficacy Scale* (Siwatu et al., 2017), providing initial evidence for the content validity of ACCReS scores.

**Purpose of Study**

Results of research have identified a mismatch between teacher and student identity, as well as new teachers reporting a lack of preparedness to provide culturally responsive supports, the purpose of this study was to conduct differential item functioning (DIF) analyses of ACCReS items. Specifically, DIF was conducted to determine if teachers' ratings on the ACCReS were invariant in relation to binary (a) racial/ethnic teacher identity (REM, White), (b) percentage of REM students in participants' schools (≤50% or >50% REM students), and (c) years of teaching experience (<1-5 years or 5 years of teaching experience). Detecting DIF might indicate

compromised validity of the ACCReS uses. Additionally, the absence of evidence for DIF might indicate that scores can be compared across variables of interest including teacher race/ethnicity, percentage of REM students in the school, and years of teaching experience in future research without underlying limitations to the instrument accounting for between-group differences (if detected). To date, no known study has assessed the presence of DIF on teachers' responses to a measure of cultural responsiveness in educational contexts. Therefore, we investigated the following three research questions:

1) Are ACCReS items invariant (i.e., do they function similarly) across teachers who identify as Black/African American, Hispanic/Latinx, American Indian/Alaska Native, Asian, Hawaiian Native/Pacific Islander or other compared to teachers who identify solely as White?
2) Are ACCReS items invariant across teachers in schools with 0-50% REM students compared to those in schools with 51-100% REM students?
3) Are ACCReS items invariant across teachers with <1-5 years of teaching experience compared to teachers with >5 years of teaching experience?

We were interested in results from DIF as well as analyzing descriptive statistics (e.g., means, standard deviations) for items and subscales to inform implications for future research and practice.

## Method

### Participants and Setting

The study presents a secondary analysis with an aggregate sample (N = 999) of teachers from three previous participant pools (n = 400, 500, and 100 teachers; Authors, 2020). One teacher's responses were removed as six items were left unanswered. No other instances of

missingness were observed. The three samples were recruited in the same calendar year (2018) using identical procedures. Specifically, Qualtrics Panel Management Services distributed study invitations to eligible teacher participants who had previously registered as panelists with Qualtrics. To participate, respondents had to be employed as an elementary, middle or high school teacher and were offered a $10 gift card for taking part in the study. Use of a paneling service for recruitment ensured data efficiency and quality in recruitment (e.g., national sample). All samples were unique. Participants were only recruited once.

The demographic makeup of the teacher sample is reflective of the teacher population in the U.S. (U.S. Department of Education, 2016). Specifically, as depicted in Table 1, the majority of participants identified as female and White. Many worked in public schools (81.66%) and general education classrooms (67.67%). Nearly all respondents (88.08%) indicated provisional or full teaching licensure/certification, and more than half had 11 or more years of teaching experience (52.01%). Respondents taught in elementary, middle, and high schools in large and small cities, as well as suburban and rural communities.

**Instrumentation**

The ACCReS includes 35 items and three subscales: ECP (13 items), CCC (11 items), and AIS (11 items). When completing the ACCReS, teachers indicated the extent to which they agree with items on a 6-point Likert-type scale (*strongly disagree* = 0, *disagree* = 1, *somewhat disagree* = 2, *somewhat agree* = 3, *agree* = 4, and *strongly agree* = 5). Items are phrased as statements corresponding with teachers' use of culturally responsive instruction and behavior support, consideration of students' culture, use of relevant data, and access to effective training and support (see Table 2).

**Analysis**

We examined descriptive statistics including item-level frequencies, means, standard deviations, and ranges. Reliability was assessed using McDonald's omega hierarchical due to its superiority to Cronbach's alpha in estimating internal consistency (Trizano-Hermosilla & Alvarado, 2016). To assess DIF, we used an iterative hybrid of ordinal logistic regression and item response theory (IRT; Choi et al., 2011) and included lordif software, similar to the procedure used in other studies of instruments producing polytomous data (e.g., the PROMIS scale; Paz et al., 2017; Reeve et al., 2007). IRT can be used to explain the relationship between latent constructs and their manifestations. Some advantages to IRT in scale development include control for confounding influences of sample characteristics, precision, and output that is easily graphed (Osterlind & Everson, 2009). Using an IRT framework to explore DIF allows for a more theoretically and procedurally rigorous examination of patterns than application of other approaches (e.g., classical test theory; Osterlind & Everson, 2009). The probability of a particular outcome, such as a score corresponding with an item on a questionnaire, should occur on a continuum according to the magnitude of the presence of a latent construct and not a separate characteristic (e.g., teacher race, student race). Applying an IRT framework makes this possible to evaluate. Item response graphs reflecting trait levels, inflection points, and other facets of respondent interactions with items can be examined when DIF is identified to assess level of impact on the instrument as a whole.

Evaluative procedures in the current study included the use of likelihood-test ratio, an approach that looks at the likelihood of a response pattern when reference and focal group responses are constrained to be invariant versus when they are permitted to vary organically (Osterlind & Everson, 2009). Although assessment of DIF according to an IRT framework assumes unidimensionality, there are some circumstances in which instruments with multiple

subscales are sufficiently unidimensional for IRT modeling. Reise et al. (2013) suggest that higher percentage of uncontaminated correlations (PUC) values can indicate fitness for IRT modeling, and when lower than .80, "researchers may consider ECV values greater than .60 and [McDonald's omega hierarchical] values greater than .70 as tentative benchmarks" (p. 22). In the absence of major violations to IRT assumptions, there are a variety of benefits to using an IRT framework for evaluation of DIF. Below, we describe how we evaluated IRT assumptions and parameters, and identified the presence of DIF. All statistical procedures were completed using R (version 1.1.423).

***IRT Assumptions***

The extent to which ACCReS response patterns satisfied IRT assumptions was evaluated across the areas of dimensionality, local independence, and monotonicity. Dimensionality refers to the presence of discrete constructs represented in a scale. In contrast to one underlying construct, the ACCReS is made up of three subscales, and assessment of dimensionality should theoretically align with the three-factor model proposed in Authors (2020). However, the underlying construct reflected by ACCReS items, teacher cultural responsiveness, was expected to make the instrument sufficiently unidimensional for assessment using an IRT framework. Local independence refers to the uniqueness of each individual item once controlling for a unifying, underlying trait. To investigate local independence, residual correlations (preferably <0.20) were examined through organization of a three-factor model using the lavaan (version 0.6-6; Rosseel, 2012) and psych (version 1.9.12; Revelle, 2019) packages in R to produce standardized residuals. Monotonicity refers to the connection between the presence of an underlying trait in relation to endorsement of an item (Sijtsma & Molenaar, 2002). The mokken

package (version 3.0.2; Andries van der Ark, 2012) was used to evaluate assumptions of monotonicity.

***Detecting the Presence of DIF***

As described above, DIF was evaluated in relation to two demographic binaries: teacher indicated race/ethnicity (REM teacher versus White teacher) and student racial/ethnic composition (<50% versus 51-100% REM students in school building). Although we would have preferred to analyze DIF across more than two categories, the sample size for certain racial and ethnic groups (e.g., Black/African American, n = 76; Hispanic/Latinx, n=66; Asian, n = 43) informed the decision to treat race/ethnicity as binary. It has been suggested that a minimum of 200 responses per variable of interest is needed when applying ordinal logistic regression (Scott et al., 2009). Therefore, responses from teachers indicating races or ethnicities other than White were aggregated (n = 214; 21.42%) and teachers who only identified as White (n = 785; 78.59%) were compared. The distribution of teachers working in schools with 0-50% and 51-100% REM students (58.07% and 34.51%, respectively) informed the decision to treat this variable as binary, as well.

To examine DIF, we used the lordif package (version 0.3-3; Choi et al., 2016) due to its strength with handling polytomous data which results from Likert-type response scales. Lordif integrates ordinal logistic regression with IRT-based trait scores, differing from the Rasch model. Lordif applies iterative purification of matching criterion by using group-specific IRT item parameter estimates for items for which DIF has been detected (Choi et al., 2011), identifies anchor items, then uses both to generate trait estimates. Lordif produces three logistic models across all items in an instrument. Model 1 includes the intercept plus an estimate of the trait. Model 2 includes Model 1 plus a group variable. Model 3 includes Model 2 plus the interaction

of trait and the group variable (Paz et al., 2017). According to Choi and colleagues (2011), this algorithm introduced by Crane et al. (2006) presents a favorable alternative to traditional purification methods because it reduces the occurrence of false positive identifications of DIF and can be more precise. Lordif integrates Samejima's graded response model (GRM; Samejima, 1969) to calibrate data to IRT assumptions. To identify the presence of DIF, the three models generated are compared according to the $\chi^2$ likelihood-ratio test. Significance (pseudo $R^2$ value $\geq$ 0.02; see Choi et al., 2011) in relationships with log likelihood values between Models 1 and 2 indicates uniform DIF, Models 1 and 3 indicates overall DIF, and Models 2 and 3 indicates nonuniform DIF. Within a logistic regression framework, identification of uniform DIF would represent DIF in which the effect was constant, whereas nonuniform DIF would be detected if effect varied according to trait level. Potential for the examination of uniform, nonuniform, and overall DIF is among the advantages to the application of lordif's hybrid model with elements of IRT and ordinal logistic regression, particularly helpful for clarifying the impact and magnitude of DIF when identified. There are various available strategies for interpreting output. The current analysis used McFadden's pseudo $R^2$ (see Lambert et al., 2018; Paz et al., 2017) to cross-compare each of the three models across items to indicate DIF (Menard, 2000).

## Results

### Assessment of IRT Assumptions

In a previous study (Authors, 2020), dimensionality was explored using exploratory factor analysis (including review of parallel analysis and factor loadings). Results from a national sample of teachers (n = 500) yielded the three factors described above. A confirmatory factor analysis conducted with a separate national sample of teachers (n = 400) produced acceptable internal consistency but mixed results with regard to adequacy of model fit (see Authors, 2020).

A PUC value of .68 (<.80) necessitated review of ECV (.58) and McDonald's omega hierarchical (.71), both indicating the absence of severe violations to IRT assumptions according to one tentative framework evaluating sufficiency for IRT modeling (Reise et al., 2013). Of note, four items were flagged for potential uncontrolled local dependence. Results revealed that the largest absolute residual correlation was > 0.20 for the following pairs of items: Items 2 and 18 (0.22), Items 3 and 18 (0.21), Items 4 and 18 (0.24), and Items 8 and 35 (0.22). These items were included nonetheless because correlations were close .20 and were considered to have minimal potential impact on results. Nonsignificant violations of manifest monotonicity were detected in 19 ACCReS items. Significant violations were identified corresponding with one item (Item 21), however this item was retained based on its potential clinical utility toward the overall purpose of the instrument.

**Identification of DIF and Assessment of Impact**

The lordif package collapses adjacent response categories when there are too few responses for an item reflecting a specific category (< 5 responses). Due to few respondents indicating *strongly disagree* for Items 3-7 and 13, the number of categories for these items was reduced from six to five by collapsing the categories *strongly disagree* and *disagree*. Additionally, as few indicated *strongly disagree* or *disagree* for Items 9 and 11, six response categories were reduced to four by combining *strongly disagree, disagree,* and *somewhat disagree*. All pseudo $R^2$ values across the three models compared across all ACCReS items were less than 0.02. Using the test value suggested by Choi and colleagues (2011) of 0.02, we therefore concluded that no items demonstrated significant DIF in relation to teacher race/ethnicity (Table 3), in relation to the percentage of REM students in the building (Table 4),

nor in relation to years of teaching experience (Table 5) according to McFadden's pseudo $R^2$ coefficient.

**Descriptive Statistics**

The mean response category across all ACCReS items was *somewhat agree* ($M = 3.73$). Response patterns across all ACCReS items were negatively skewed – such that teachers tended to favor response categories consistent with better-supported practices, and responses inconsistent with favorable practices (growth areas) represented outliers. Respondents agreed the most with items associated with the ECP factor (mean range 3.94 - 4.47), and least with items associated with the AIS factor (mean range 2.59 - 3.55). With the exception of items described previously, respondents interacted with the full range of ACCReS response options. Findings reflected acceptable to good internal consistency; McDonald's omega hierarchical coefficients for the ACCReS subscales were 0.82 (ECP), 0.79 (CCC) and 0.87 (AIS), and .71 for the complete instrument. This indicates that items comprising each subscale are sufficiently related to one another and provides evidence that the ACCReS items in each subscale reflect a single unifying construct (cultural responsiveness). This is important for its use as a measure of teachers' perception of their culturally responsive practice in the classroom for use in decision-making about professional development needs.

## Discussion

The aim of this study was to determine whether items on the ACCReS demonstrated measurement invariance across participants. Overall, findings indicated that items did not demonstrate significant DIF when comparing responses from (a) REM teachers and White teachers, (b) teachers in schools with 0-50% and 51-100% REM students, as well as (c) teachers with <1-5 and >5 years of teaching experience. These findings also provide additional evidence

toward the technical adequacy of the ACCReS as a teacher self-report measure of cultural responsiveness. The absence of evidence for DIF provides preliminary evidence that scores can be compared across variables of interest including teacher race/ethnicity, percentage of REM students in the school, and years of teaching experience in future research without underlying limitations to the instrument. In practice, the current analysis suggests there should not be systematic differences in scores based on the teacher and student demographic variables investigated in this study.

On average, items on the ECP subscale were rated higher than items on the CCC and AIS scales in the sample of 999 educators (see Table 2). This may be because items on the ECP primarily target foundational behavior management and instructional practices, whereas items on the CCC and AIS ask about explicit consideration for students' culture, and access to external data, support and training, respectively. Teachers may perceive items on the CCC and AIS scales as containing areas in which additional training or support might be beneficial. For instance, the item "I meet with support personnel (e.g., instructional coaches, lead teachers, consultants) to help me to find evidence of disproportionality (e.g., racial, gender) in my classroom data" had a lower average rating than other items on the ACCReS and AIS subscale in particular, indicating this may be an area of need for educators.

**Limitations and Future Research**

Findings should be interpreted with consideration of the study's limitations. First, the sample is comprised of participants from three prior studies, all of whom were educators preregistered as potential panelists with Qualtrics (Authors, 2020). Although this may introduce sampling bias, the large national sample of participants was representative of U.S. teacher demographic trends (U.S. Department of Education, 2016), taught in a variety of school settings

(urban, rural), and instructed youth across grade levels (elementary, secondary). Although the sample was sufficient for analyses, many participants identified as White and it was not possible to examine DIF beyond a binary operationalization of teacher race/ethnicity. There are significant disadvantages to bifurcating race in analyses. Ideally the sample size for specific race and ethnicity categories would have allowed for a more comprehensive analysis. In future research, this limitation should be addressed. The potential for inflation of Type I error rates in DIF analyses with unequal reference and focal groups (Herrera & Gómez, 2007) could also be better accounted for in additional, detailed analysis of DIF in relation to teacher race and ethnicity. The use of an IRT framework with a multidimensional instrument may be considered a limitation. However, this decision was carefully reasoned, aligned with the analyses of similarly structured instruments (Paz et al., 2017; Reeve et al., 2007). Further, the significant violation of monotonicity identified in one item is also a potential limitation. However, this item was considered critical to educators' understanding of underlying issues pertaining to the use of the ACCReS. Due to the purpose of the instrument (i.e., to help identify targets for teacher professional development and support), it was retained.

Additional research might also target DIF analyses with teachers of students from specific racial/ethnic groups, specifically those from groups disproportionately represented in disciplinary data (e.g., Black/African American, Hispanic/Latinx). It may also be important to conduct the analysis with more than the two groups reflecting percentage of REM students and teachers' years of experience as treating both variables as binary may have compromised the quality of these data in our analyses. Finally, future research might address comparing teachers' ratings on the ACCReS and their actual classroom practice, or distal outcomes that may result

from a culturally responsive approach (e.g., more time engaged in learning, less exclusionary disciplinary). This would provide evidence of concurrent and predictive validity, respectively.

## Conclusion

Findings from the current study indicate ACCReS items were invariant across REM and White teachers, as well as educators in buildings with 0-50% and 51-100% REM students, and teachers with <1-5 years and >5 years of teaching experience. These results support its continued validation for use in research and classroom-based practice. Teachers' assessment of their culturally responsiveness is a small piece of confronting systemic racism in schools to establish more equitable and effective learning environments.

**References**

Authors (2020). Blinded for review.

Authors (2018). Blinded for review.

Authors (2012a). Blinded for review.

Authors (2012b). Blinded for review.

Andries van der Ark, L. (2012). New developments in mokken scale analysis in R. *Journal of Statistical Software, 48*(5), 1-27. http://www.jstatsoft.org/v48/i05/.

Bates, L. A., & Glick, J. E. (2013). Does it matter if teachers and schools match the student? Racial and ethnic disparities in problem behaviors. *Social Science Research*, *42*(5), 1180-1190. https://doi.org/10.1016/j.ssresearch.2013.04.005

Bergeron, B. S. (2008). Enacting a culturally responsive curriculum in a novice teacher's classroom: Encountering disequilibrium. *Urban Education*, *43*(1), 4-28. https://doi.org/10.1177/0042085907309208

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1-30. http://www.jstatsoft.org/v39/i08/.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2016). *lordif: Logistic Ordinal Regression Differential Item Functioning using IRT*. R package version 0.3-3. https://CRAN.R-project.org/package=lordif

Crane, P. K., Gibbons, L. E., Jolley, L., & Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. *Medical Care, 44*(11), S115-S123. https://doi.org/10.1097/01.mlr.0000245183.28384.ed

García, E., Arias, M. B., Harris Murri, N. J., & Serna, C. (2010). Developing responsive teachers: A challenge for a demographic reality. *Journal of Teacher Education*, *61*(1-2), 132-142. https://doi.org/10.1177/0022487109347878

Gay, G. (2018). *Culturally responsive teaching: Theory, research, and practice*. Teachers College Press.

Girvan, E. J., Gion, C., McIntosh, K., & Smolkowski, K. (2017). The relative contribution of subjective office referrals to racial disproportionality in school discipline. *School Psychology Quarterly*, *32*(3), 392-404. https://doi.org/10.1037/spq0000178

Gregory, A., Skiba, R. J., & Noguera, P. A. (2010). The achievement gap and the discipline gap: Two sides of the same coin? *Educational Researcher*, *39*(1), 59-68. https://doi.org/10.3102/0013189X09357621

Guyton, E. M., & Wesche, M. V. (2005). The multicultural efficacy scale: Development, item selection, and reliability. *Multicultural Perspectives*, *7*(4), 21-29. https://doi.org/10.1207/s15327892mcp0704_4

Herrera, A., & Gómez, J. (2007). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality & Quantity, 42*(6), 739-755. https://doi.org/10.1007/s11135-006-9065-z

Hershfeldt, P. A., Sechrest, R., Pell, K. L., Rosenberg, M. S., Bradshaw, C. P., & Leaf, P. J. (2009). Double-Check: A framework of cultural responsiveness applied to classroom behavior. *Teaching Exceptional Children Plus*, *6*(2), 2-18.

Hung, M., Smith, W. A., Voss, M. W., Franklin, J. D., Gu, Y., & Bounsanga, J. (2020). Exploring student achievement gaps in school districts across the United

States. *Education and Urban Society*, *52*(2), 175-193. https://doi.org/10.1177/0013124519833442

Kohli, R., Pizarro, M., & Nevárez, A. (2017). The "new racism" of K–12 schools: Centering critical research on racism. *Review of Research in Education*, *41*(1), 182-202. https://doi.org/10.3102/0091732X16686949

La Salle, T. P., Wang, C., Wu, C., & Rocha Neves, J. (2020). Racial mismatch among minoritized students and white teachers: Implications and recommendations for moving forward. *Journal of Educational and Psychological Consultation*, *30*(3), 314-343. https://doi.org/10.1080/10474412.2019.1673759

Lee, J., Tice, K., Collins, D., Brown, A., Smith, C., & Fox, J. (2012). Assessing student teaching experiences: Teacher candidates' perceptions of preparedness. *Educational Research Quarterly*, *36*(2), 3-20.

Malone, C. M., & Ishmail, K. Z. (2020). A snapshot of multicultural training in school psychology. *Psychology in the Schools, 57*(7) 1022-1039. https://doi.org/10.1002/pits.22392

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*(1), 17–24. https://doi.org/10.1080/00031305.2000.10474502

Miretzky, D., Chennault, R. E., & Fraynd, D. J. (2016). Closing an opportunity gap: How a modest program made a difference. *Education and Urban Society*, *48*(1), 48-76. https://doi.org/10.1177/0013124513501320

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning.* SAGE.

Paz, S. H., Spritzer, K. L., Reise, S. P., & Hays, R. D. (2017). Differential item functioning of

the patient-reported outcomes information system (PROMIS) pain interference item bank by language (Spanish versus English). *Quality of Life Research, 26*(6), 1451–1462. https:// 10.1007/s11136-017-1499-3

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., … & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(5), S22-S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*(2), 129-140. https://www.ncbi.nlm.nih.gov/pubmed/23030794

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17* (4, Pt. 2).

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., Graeff, A. D., Groenvold, M., … Sprangers, M. A. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, *62*(3), 288–295. https://doi.org/10.1016/j.jclinepi.2008.06.003

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. SAGE.

Siwatu, K. O. (2007). Preservice teachers' culturally responsive teaching self-efficacy and outcome expectancy beliefs. *Teaching and Teacher Education, 23*(7), 1086–1101. https://doi.org/10.1016/j.tate.2006.07.011

Siwatu, K. O., Putman, S. M., Starker-Glass, T. V., & Lewis, C. W. (2017). The culturally responsive classroom management self-efficacy scale. *Urban Education*, *52(7)*, 862-888. https://doi.org/10.1177/0042085915602534

Revelle, W. (2019). psych: Procedures for personality and psychological research, Northwestern University. https://CRAN.Rproject.org/package=psychVersion=1.9.12.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. http://www.jstatsoft.org/v48/i02/

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions. *Frontiers in Psychology, 7,* 769-778. https://10.3389/fpsyg.2016.00769

U.S. Department of Education. (2016). *The state of racial diversity in the educator workforce*. Washington, D.C. https://rb.gy/8aqkav

U.S. Department of Education. (2019). *State nonfiscal survey of public elementary and secondary education, 2000–01 and 2017–18*. National Center for Education Statistics, Common Core of Data. https://nces.ed.gov/programs/coe/indicator_cge.asp#info

Vincent, C. G., Randall, C., Cartledge, G., Tobin, T. J. & Swain-Bradway, J. (2011). Toward a conceptual integration of cultural responsiveness and school-wide positive behavior support. *Journal of Positive Behavior Interventions, 13*(4), 219-229. https://doi.org/10.1177/1098300711399765

Yarnell, L. M., & Bohrnstedt, G. W. (2018). Student-teacher racial match and its association with Black student achievement: An exploration using multilevel structural equation modeling. *American Educational Research Journal*, *55*(2), 287-324. https://doi.org/10.3102/0002831217734804

**Table 1**

*Respondent Characteristics*

| | % | *n* |
|---|---|---|
| Respondent Gender | | |
| Female | 76.11 | 755 |
| Male | 23.79 | 236 |
| Other | 0.20 | 2 |
| Respondent Race or Ethnicity[1] | | |
| White | 82.08 | 820 |
| Black or African American | 7.61 | 76 |
| Hispanic or Latinx | 6.61 | 66 |
| American Indian, Alaska Native | 1.60 | 16 |
| Asian | 4.30 | 43 |
| Hawaiian Native. & Pacific Islander | 0.40 | 4 |
| Other | 1.60 | 16 |
| Status of Licensure | | |
| Licensed/certified | 79.80 | 790 |
| Not yet licensed/certified | 8.89 | 88 |
| Provisional license/certification | 8.28 | 82 |
| Other | 3.13 | 31 |
| Level of Certification | | |
| General education certification | 67.67 | 674 |
| Special education certification | 7.53 | 75 |
| Both | 12.85 | 128 |
| Neither | 12.05 | 120 |
| Years of Teaching Experience | | |
| < 1 Year | 2.31 | 23 |
| 1-5 Years | 24.04 | 239 |
| 6-10 Years | 21.73 | 216 |
| ≥ 11 Years | 52.01 | 517 |
| School Environment | | |
| Large City | 24.47 | 244 |
| Small City | 20.66 | 206 |
| Suburban | 35.80 | 357 |
| Rural | 19.16 | 191 |

| | | |
|---|---|---|
| School Type[1] | | |
| Public | 81.66 | 815 |
| Private | 13.13 | 131 |
| Charter | 4.91 | 49 |
| Regional, Alternative or Technical | 1.40 | 14 |
| Grades Taught[1] | | |
| Elementary (K – 5th grade) | 48.35 | 483 |
| Middle (6th – 8th grade) | 28.93 | 289 |
| High School (9th – 12th grade) | 37.54 | 375 |
| Percentage of racially and ethnically minoritized students in school | | |
| 0 - 25% | 38.61 | 385 |
| 26 - 50% | 19.46 | 194 |
| 51 - 75% | 18.56 | 185 |
| 76 - 100% | 15.95 | 159 |
| Not sure | 7.52 | 75 |
| Percentage of English learners | | |
| 0 - 25% | 65.93 | 658 |
| 26 - 50% | 12.32 | 123 |
| 51 - 75% | 8.12 | 81 |
| 76 - 100% | 6.61 | 66 |
| Not sure | 7.11 | 71 |
| Percentage of students receiving free or reduced lunch | | |
| 0 - 25% | 27.68 | 276 |
| 26 - 50% | 17.55 | 175 |
| 51 - 75% | 19.56 | 195 |
| 76 - 100% | 25.28 | 252 |
| Not sure | 10.03 | 100 |

*Note.* [1]Denotes questions for which respondents were asked to "Check all that apply." Percentages may exceed 100%. Participants were not required to answer all demographic items.

**Table 2**

*Item mean, standard deviation (SD), skew, kurtosis, and subscale internal consistency ($\omega_h$)*

| Subscale and Item Number | Mean | SD | Skew | Kurtosis | $\omega_h$ |
|---|---|---|---|---|---|
| **Equitable Classroom Practices (ECP)** | **4.26** | **0.88** | **-** | **-** | **0.82** |
| 1. I use explicit instruction when I teach (e.g., clearly describe, model, and practice content with students). | 4.24 | 0.92 | -1.58 | 3.45 | |
| 2. I differentiate instruction to support the different learners I teach. | 4.16 | 0.91 | -1.29 | 2.53 | |
| 3. I provide additional (or more intensive) academic support when a student needs it. | 4.31 | 0.84 | -1.51 | 3.46 | |
| 4. I plan lessons that are designed to actively engage all learners when I teach. | 4.24 | 0.86 | -1.38 | 2.98 | |
| 5. I listen actively to students when they express concerns. | 4.39 | 0.78 | -1.73 | 4.92 | |
| 6. I engage in more positive interactions with students than negative interactions. | 4.27 | 0.86 | -1.33 | 2.53 | |
| 7. I am consistent and fair when it comes to discipline. | 4.27 | 0.79 | -1.26 | 2.94 | |
| 8. I explicitly teach social skills (e.g., ways to ask for help appropriately). | 3.94 | 1.09 | -1.08 | 1.07 | |
| 9. I explicitly teach students about my expectations for classroom behavior. | 4.42 | 0.78 | -1.45 | 2.65 | |
| 10. Each day, I personally greet all of my students. | 4.14 | 1.06 | -1.4 | 1.88 | |
| 11. I work to build a positive relationship with each student I teach. | 4.47 | 0.73 | -1.66 | 4.39 | |
| 12. I deliver praise equitably in my classroom. | 4.30 | 0.80 | -1.33 | 2.96 | |
| 13. I actively monitor all parts of my classroom. | 4.24 | 0.82 | -1.27 | 2.49 | |
| **Consideration of Culture and Context (CCC)** | **3.61** | **1.15** | **-** | **-** | **0.79** |
| 14. Culturally and contextually relevant instruction is important to how I teach. | 3.73 | 1.12 | -0.93 | 0.94 | |
| 15. I know how to provide culturally and contextually relevant instruction. | 3.69 | 1.01 | -0.97 | 1.55 | |
| 16. I modify the curriculum to be culturally and contextually relevant, when appropriate. | 3.69 | 1.08 | -0.98 | 1.34 | |
| 17. I consider students' culture when I decide on the type of instructional support I will provide. | 3.41 | 1.23 | -0.92 | 0.64 | |
| 18. I understand that behavior may be context-specific (e.g., different behaviors may be more appropriate at home or school). | 4.08 | 0.89 | -1.18 | 2.41 | |
| 19. I consider a student's culture when selecting a research-based intervention strategy. | 3.38 | 1.21 | -0.77 | 0.43 | |
| 20. I self-assess my cultural biases regularly. | 3.39 | 1.18 | -0.76 | 0.48 | |
| 21. I understand that some students are at risk for being disproportionally excluded from the learning | 3.90 | 1.07 | -1.18 | 1.72 | |

| | | | | | |
|---|---|---|---|---|---|
| environment (e.g., sent to the office, suspended, expelled). | | | | | |
| 22. I gather information about my students' families (e.g., customs, languages spoken, cultural traditions). | 3.54 | 1.19 | -0.84 | 0.56 | |
| 23. I consider students' culture and language when I select assessment tools. | 3.45 | 1.27 | -0.85 | 0.32 | |
| 24. I know where to find information about culturally and contextually relevant behavior management practices. | 3.44 | 1.16 | -0.77 | 0.42 | |
| **Accessing Information and Support (AIS)** | **3.21** | **1.33** | **-** | **-** | **0.87** |
| 25. I ask families to help define my classroom expectations. | 2.59 | 1.43 | -0.10 | -0.88 | |
| 26. I collect classroom data to inform the equity of my interactions across students (e.g., frequency and distribution of positive interactions). | 3.24 | 1.30 | -0.62 | -0.26 | |
| 27. I collect classroom data to inform the equity of my disciplinary actions across students (e.g., evidence of consistent consequences administered). | 3.28 | 1.31 | -0.64 | -0.28 | |
| 28. I review academic data for trends that reflect disproportionality (e.g., students of a certain race not achieving in mathematics versus students from other groups). | 3.20 | 1.31 | -0.57 | -0.26 | |
| 29. I seek professional development opportunities (e.g., attend conferences, workshops, trainings) to learn about how to engage in culturally and contextually relevant practice. | 3.48 | 1.31 | -0.81 | 0.12 | |
| 30. I request the resources (e.g., time, staff, training) I need to implement culturally and contextually relevant instruction. | 3.28 | 1.25 | -0.71 | 0.11 | |
| 31. I request the resources (e.g., time, staff, training) I need to implement culturally and contextually relevant behavior support. | 3.35 | 1.20 | -0.64 | 0.15 | |
| 32. I request to meet with support personnel (e.g., instructional coaches, lead teachers, consultants) to help me consider cultural and contextual factors that might affect how I support students' behavior. | 3.12 | 1.34 | -0.49 | -0.45 | |
| 33. I meet with support personnel (e.g., instructional coaches, lead teachers, consultants) to help me to find evidence of disproportionality (e.g., racial, gender) in my classroom data. | 2.96 | 1.44 | -0.33 | -0.77 | |
| 34. I talk to administrators in my building about accessing the resources I need to provide culturally and contextually relevant academic supports. | 3.27 | 1.32 | -0.69 | -0.06 | |
| 35. I seek the resources (e.g., time, access, translators) I need to partner with families to support students. | 3.55 | 1.15 | -0.91 | 0.78 | |

*Note.* [1]Descriptives were calculated by coding participant responses as follows: 0 = strongly disagree, 1 = disagree, 2 = somewhat disagree, 3 = somewhat agree, 4 = agree, 5 = strongly agree

**Table 3**

*Differential Item Functioning (DIF) Results for Racially and Ethnically Minoritized Teachers and White Teachers*

| Subscale and Item Number | Number of Categories | [a]Test for Overall DIF | [b]Test for Uniform DIF | [c]Non-Uniform DIF |
|---|---|---|---|---|
| Equitable Classroom Practices (ECP) | | | | |
| 1 | 4 | 0.01 | 0.01 | 0 |
| 2 | 4 | 0.01 | 0.01 | 0 |
| 3 | 3 | 0 | 0 | 0 |
| 4 | 3 | 0 | 0 | 0 |
| 5 | 3 | 0 | 0 | 0 |
| 6 | 4 | 0 | 0 | 0 |
| 7 | 3 | 0 | 0 | 0 |
| 8 | 5 | 0 | 0 | 0 |
| 9 | 4 | 0 | 0 | 0 |
| 10 | 5 | 0 | 0 | 0 |
| 11 | 3 | 0.01 | 0.01 | 0 |
| 12 | 4 | 0 | 0 | 0 |
| 13 | 4 | 0 | 0 | 0 |
| Consideration of Culture and Context (CCC) | | | | |
| 14 | 6 | 0 | 0 | 0 |
| 15 | 4 | 0 | 0 | 0 |
| 16 | 4 | 0 | 0 | 0 |
| 17 | 6 | 0 | 0 | 0 |
| 18 | 3 | 0 | 0 | 0 |
| 19 | 6 | 0 | 0 | 0 |
| 20 | 6 | 0 | 0 | 0 |
| 21 | 5 | 0 | 0 | 0 |
| 22 | 6 | 0 | 0 | 0 |
| 23 | 6 | 0 | 0 | 0 |
| 24 | 6 | 0 | 0 | 0 |
| Accessing Information and Support (AIS) | | | | |
| 25 | 6 | 0 | 0 | 0 |
| 26 | 6 | 0 | 0 | 0 |
| 27 | 6 | 0 | 0 | 0 |
| 28 | 6 | 0 | 0 | 0 |
| 29 | 6 | 0 | 0 | 0 |
| 30 | 6 | 0 | 0 | 0 |
| 31 | 6 | 0 | 0 | 0 |
| 32 | 6 | 0 | 0 | 0 |
| 33 | 6 | 0 | 0 | 0 |
| 34 | 6 | 0 | 0 | 0 |
| 35 | 6 | 0 | 0 | 0 |

*Note.* Bold shows Pseudo $R^2p$ value ≥ 0.02 indicating statistically significant DIF
[a]Model 1 (intercept + rating) versus Model 3 (Model 2 + rating * group)
[b]Model 1 (intercept + ability) versus Model 2 (Model 1 + group)
[c]Model 2 (Model 1 + group) versus Model 3 (Model 2 + rating * group)

**Table 4**

*Differential Item Functioning (DIF) Results for Teachers in Schools with 0-50% REM Students and Teachers in Schools with 51-100% REM Students*

| Subscale and Item Number | Number of Categories | [a]Test for Overall DIF | [b]Test for Uniform DIF | [c]Non-Uniform DIF |
|---|---|---|---|---|
| Equitable Classroom Practices (ECP) | | | | |
| 1 | 4 | 0 | 0 | 0 |
| 2 | 4 | 0 | 0 | 0 |
| 3 | 3 | 0 | 0 | 0 |
| 4 | 4 | 0 | 0 | 0 |
| 5 | 3 | 0.01 | 0 | 0 |
| 6 | 4 | 0 | 0 | 0 |
| 7 | 3 | 0 | 0 | 0 |
| 8 | 5 | 0 | 0 | 0 |
| 9 | 4 | 0 | 0 | 0 |
| 10 | 5 | 0 | 0 | 0 |
| 11 | 3 | 0 | 0 | 0 |
| 12 | 4 | 0 | 0 | 0 |
| 13 | 4 | 0 | 0 | 0 |
| Consideration of Culture and Context (CCC) | | | | |
| 14 | 6 | 0 | 0 | 0 |
| 15 | 5 | 0 | 0 | 0 |
| 16 | 6 | 0 | 0 | 0 |
| 17 | 6 | 0 | 0 | 0 |
| 18 | 4 | 0 | 0 | 0 |
| 19 | 6 | 0 | 0 | 0 |
| 20 | 6 | 0 | 0 | 0 |
| 21 | 6 | 0 | 0 | 0 |
| 22 | 6 | 0 | 0 | 0 |
| 23 | 6 | 0 | 0 | 0 |
| 24 | 6 | 0 | 0 | 0 |
| Accessing Information and Support (AIS) | | | | |
| 25 | 6 | 0 | 0 | 0 |
| 26 | 6 | 0 | 0 | 0 |
| 27 | 6 | 0 | 0 | 0 |
| 28 | 6 | 0 | 0 | 0 |
| 29 | 6 | 0 | 0 | 0 |
| 30 | 6 | 0 | 0 | 0 |
| 31 | 6 | 0 | 0 | 0 |
| 32 | 6 | 0 | 0 | 0 |
| 33 | 6 | 0 | 0 | 0 |
| 34 | 6 | 0 | 0 | 0 |
| 35 | 6 | 0 | 0 | 0 |

*Note.* Bold shows Pseudo $R^2p$ value $\geq 0.02$ indicating statistically significant DIF
[a]Model 1 (intercept + rating) versus Model 3 (Model 2 + rating*group)
[b]Model 1 (intercept + ability) versus Model 2 (Model 1 + group)
[c]Model 2 (Model 1 + group) versus Model 3 (Model 2 + rating*group)

**Table 5**

*Differential Item Functioning (DIF) Results for Teachers with <1-5 Years of Teaching Experience and Teachers with >5 Years of Teaching Experience*

| Subscale and Item Number | Number of Categories | [a]Test for Overall DIF | [b]Test for Uniform DIF | [c]Non-Uniform DIF |
|---|---|---|---|---|
| Equitable Classroom Practices (ECP) | | | | |
| 1 | 4 | 0.01 | 0.01 | 0 |
| 2 | 4 | 0.01 | 0.01 | 0 |
| 3 | 3 | 0.01 | 0 | 0 |
| 4 | 5 | 0.01 | 0.01 | 0 |
| 5 | 3 | 0.01 | 0.01 | 0 |
| 6 | 4 | 0 | 0 | 0 |
| 7 | 4 | 0.01 | 0.01 | 0 |
| 8 | 4 | 0 | 0 | 0 |
| 9 | 4 | 0.01 | 0.01 | 0 |
| 10 | 5 | 0 | 0 | 0 |
| 11 | 3 | 0 | 0 | 0 |
| 12 | 4 | 0 | 0 | 0 |
| 13 | 4 | 0.01 | 0 | 0.01 |
| Consideration of Culture and Context (CCC) | | | | |
| 14 | 6 | 0 | 0 | 0 |
| 15 | 5 | 0 | 0 | 0 |
| 16 | 5 | 0 | 0 | 0 |
| 17 | 6 | 0 | 0 | 0 |
| 18 | 4 | 0 | 0 | 0 |
| 19 | 6 | 0 | 0 | 0 |
| 20 | 6 | 0.01 | 0 | 0 |
| 21 | 5 | 0 | 0 | 0 |
| 22 | 6 | 0 | 0 | 0 |
| 23 | 6 | 0 | 0 | 0 |
| 24 | 5 | 0 | 0 | 0 |
| Accessing Information and Support (AIS) | | | | |
| 25 | 6 | 0 | 0 | 0 |
| 26 | 6 | 0 | 0 | 0 |
| 27 | 6 | 0 | 0 | 0 |
| 28 | 6 | 0 | 0 | 0 |
| 29 | 6 | 0 | 0 | 0 |
| 30 | 6 | 0 | 0 | 0 |
| 31 | 6 | 0 | 0 | 0 |
| 32 | 6 | 0 | 0 | 0 |
| 33 | 6 | 0 | 0 | 0 |
| 34 | 6 | 0 | 0 | 0 |
| 35 | 6 | 0 | 0 | 0 |

*Note.* Bold shows Pseudo $R^2p$ value $\geq 0.02$ indicating statistically significant DIF

[a]Model 1 (intercept + rating) versus Model 3 (Model 2 + rating*group)

[b]Model 1 (intercept + ability) versus Model 2 (Model 1 + group)

[c]Model 2 (Model 1 + group) versus Model 3 (Model 2 + rating*group)