Multilevel IRT Analysis of the Everyday Discrimination Scale and the Racial/Ethnic Discrimination Index

Ye Feng[1], **Yuen Mi Cheon**[2], **Tiffany Yip**[1], **Heining Cham**[1]

[1]Department of Psychology, Fordham University

[2]Department of Child Development and Education, Myongji University

Abstract

Unfair treatment based on race is an unfortunate reality. While there is increasing interest in mapping the daily and longer-term impact of discrimination in psychology, studies that examine the psychometric properties of indicators spanning these timeframes is limited. Item response analysis examined the measurement characteristics of two daily measures of ethnic/racial discrimination: 1) the six-item Racial/Ethnic Discrimination Index (REDI), and 2) the modified five-item Everyday Discrimination Scale (EDS; Williams, Yu, Jackson, & Anderson, 1997). This study investigated whether the two scales can be appropriately adapted to access adolescents' daily-level ethnic/racial discrimination experiences. Both measures were administered for 14 consecutive days in a sample of 350 adolescents attending public schools in a large, urban area. Results suggest that the REDI has high loading and high difficulty. All REDI items functioned similarly at daily and person levels, suggesting that any single REDI item measured on a single day is sufficient for measuring daily ethnic/racial discrimination experiences. The EDS also shows high loading and high difficulty. However, EDS items functioned differently at the daily- and person-levels. REDI items were invariant across gender and race/ethnicity (African Americans, Asians, and Latinx). Recommendations for measuring daily ethnic/racial discrimination are discussed.

Keywords

racial/ethnic discrimination; daily diary; multilevel item response theory

Racial/ethnic discrimination (RED) is considered to be differential or unfair treatment on the basis of ethnicity or race, and is a normative experience for racial/ethnic minorities in the United States (Williams & Williams-Morris, 2000). A large and growing body of systematic reviews and meta-analyses have found that RED is associated with negative mental health outcomes such as psychological distress, depression and anxiety, academic

outcomes, and negative physical health outcomes, such as increased blood pressure and risky health behaviors at various points along the lifespan (Benner et al., 2018; Paradies, 2006; Paradies et al., 2015; Schmitt, Postmes, Branscombe, & Garcia, 2014; Williams & Mohammed, 2009).

Studies of RED began with a focus on lifetime discrimination and major discriminatory events such as housing or employment discrimination (Collins & Essed, 1992; Pager & Shepherd, 2008; Ziegert & Hanges, 2005). At the same time, researchers recognized that RED also exists in everyday interactions in more subtle, but no less harmful ways (Sue, 2010). Research finds that "everyday" experiences of RED comprise a larger percentage of exposure compared to major life events (Kessler, Mickelson, & Williams, 1999). Contemporary researchers have come to understand that RED takes many forms and encompasses both major and everyday events. In order to develop a holistic picture of how RED is implicated in mental and physical health, it is necessary to study the discrimination at the person- (i.e., how it varies from one person to the next) and the daily- (i.e., how it varies from one day to the next for the same person) levels, respectively capturing major and everyday discrimination. Though there has been interest in studying the RED at both daily- and person-level, only a few studies have distinguished the source of variability between daily-level RED and person-level RED (Seaton & Iida, 2019; Torres & Ong, 2010), and no study has examined the daily RED measure systematically. The current study addresses this gap in the literature with an item response analyses of two measures of RED assessed at both the person and the daily levels.

Measuring Racial/Ethnic Discrimination

The growing interest in measuring the impact of RED is reflected in the number of measures that have been used in the literature. Recent meta-analyses and systematic reviews find that there are 36 to 123 measures of RED (Benner et al., 2018; Carter et al., 2019; Yip, Wang, Mootoo, & Mirpuri, 2019), reflecting the considerable variability in measuring the construct of racial/ethnic discrimination. Four sources of variability can be identified. First, scales employ different time frames for retrospective responses (e.g., past month, past year; Mouzon, Taylor, Woodward, & Chatters, 2017). Second, some measures focus on the frequency of racial/ethnic discriminatory events in a checklist format (Williams & Mohammed, 2009), while other measures incorporate an assessment of impact (Fisher, Wallace, & Fenton, 2000; Harrell, 2000). Third, there is variability in the number of items or measures assessing RED ranging from single items (Gee, Spencer, Chen, Yip, & Takeuchi, 2007; Kessler et al., 1999; Krieger, Smith, Naishadham, Hartman, & Barbeau, 2005) to multiple items or multidimensional scales (Lewis, Yang, Jacobs, & Fitchett, 2012; Schulz et al., 2006; Shariff-Marco et al., 2011). Fourth, there is variability in the extent to which measures interrogate the form and the perpetrator of discrimination, for example, in-person discrimination vs. online discrimination, overt discrimination vs. subtle discrimination, and vicarious vs. personal experiences (Benner et al., 2018; Tynes, Giang, Williams, & Thompson, 2008). To this list, there is a dearth of evidence-based guidelines about measuring RED across days and individuals.

One of the most widely used racial/ethnic discrimination measures is the Everyday Discrimination Scale (EDS; Williams, Yu, Jackson, & Anderson, 1997). The EDS is a ten-item self-report scale that was developed to measure discrimination experiences over a prescribed period of time (e.g., past year, past 6 months; Williams, Yu, Jackson, & Anderson, 1997). The original EDS focused on discrimination broadly, not specifically attributable to race/ethnicity. More recently, researchers have modified the scale to focus specifically on race/ethnicity with two approaches: 1) incorporating an attribution of race/ethnicity into the stem of the questions (Krieger, 2012), and 2) creating a branching design where respondents are first asked to indicate discrimination and followed by an attribution to a specific social category (e.g., gender, race/ethnicity, age, etc; Bastos, Harnois, Bernardo, Peres, & Paradies, 2017; Gomez & Trierweiler, 2001). Despite research showing that the branching approach generates a much higher response frequency than the race-attributed approach (Shariff-Marco et al., 2011, 2009), researchers interpret the results of two approaches as comparable. As such, there are methodological reasons to be weary of comparing prevalence estimates of RED across the two approaches.

The Psychometrics of Measuring Racial/Ethnic Discrimination

The EDS has been widely used to assess discrimination and its associations with various developmental outcomes from physical to mental health (Gee et al., 2007; Schulz et al., 2006; Williams & Mohammed, 2009), and across different racial/ethnic groups including African Americans, Asian-Americans, Latinx and Pacific Islanders (Chan, Tran, & Nguyen, 2012; Lewis et al., 2012; Shariff-Marco et al., 2011). Most analyses of the factor structure have concluded that the EDS comprises a single dimension of discrimination (Bernstein, Park, Shin, Cho, & Park, 2011; Clark, Coleman, & Novak, 2004; Lewis et al., 2012; Reeve et al., 2011; Shariff-Marco et al., 2011) with few exceptions (e.g., Guyll, Matthews, & Bromberger, 2001). Previous studies focused on the psychometric properties of the EDS reported high internal consistency (ranged from .81 to .88; Lewis et al., 2012; Reeve et al., 2011; Shariff-Marco et al., 2011) and validity (moderate to strong association with similar measures of RED; Harnois et al., 2019; Krieger et al., 2005).

Despite these psychometric analyses, there has been less item-level analysis of the EDS or other discrimination measures. Item-level analyses are particularly useful for researchers who wish to use abbreviated, or single-item measures in daily diary methods, or other intensive repeated measures. This is particular concern for scholars who are interested in everyday discrimination, its prevalence, and its health correlates. Item response theory (IRT) is one statistical procedure to investigate the psychometric properties of discrimination items. IRT uses probability functions to model the relationships between the categorical items and the latent trait (RED). The probability functions include parameters to evaluate the performance of items and explain the differences in person's responses to items. For example, a two-parameter logistic (2-PL) model, involves two parameters to model the relationships between the binary items and the latent trait: difficulty and loading. In the 2-PL model, the difficulty parameter is the value of the latent trait where the probability of an affirmative response is 50%. The loading parameter represents how well the items differentiate against different levels of the latent trait (de Ayala, 2009). Based on the item parameters, researchers can modify scales such as eliminating redundant items that

have similar item difficulty and loading with other items and identifying the estimating power of scales using test information functions. IRT framework also allows researchers to test whether the RED items are invariant across diverse populations (e.g., race/ethnicity). Measurement invariance means that persons from different populations have the same probabilities in the item responses, given the same level of the latent trait.

Application of IRT method to the Everyday Discrimination Scale

The following section reviews studies that have applied IRT methods to the Everyday Discrimination Scale (Berenbon, 2018; Lewis et al., 2012; Stucky et al., 2011), and discusses how the current study contributes to this body of work. It is worth mentioning that the studies reviewed here examine the EDS at the person level, with a particular focus on measurement invariance across certain groups (e.g., gender, race). Stucky et al. (2011) used IRT to examine the nine-item EDS in two samples (Study 1: 589 African American law school students, Study 2: 3,527 nationally representative community sample). Results supported a revised EDS with only five items. In addition, three items were non-invariant across gender: not as smart, dishonest, insulted/harassed. The revised EDS had satisfactory predictive validity to similar measures of overt discrimination. Lewis et al. (2012) tested the invariance of the 10-item EDS across race/ethnicity (African American, Chinese, Hispanic, Japanese, and White) among 3,295 middle-aged American women and found three non-invariant EDS items: poorer service, dishonest, treated with less courtesy. Most recently, Berenbon (2018) used IRT to study the nine-item EDS using a national sample ($N = 2,666$). The EDS items showed high loadings and high precision of measurement. One item (i.e., afraid of you) was non-invariant across gender. Taken together, these three studies provide suggestions that certain EDS items may function differently for men and women when administered at the person level. The current study builds upon this literature by testing invariance of the EDS scale administered at the daily and the person levels, across *both* gender and race/ethnicity.

Measuring Racial/Ethnic Discrimination at the Daily Level

Despite being assessed as a primarily person-level construct, RED is ultimately a daily-level phenomenon. This is especially true for "everyday" indicators of RED. In recent years psychologists have begun to assess RED using daily diary methods, intensive repeated measures, and novel methodological techniques to more closely link measurement to the lived experience of the construct. Using intensive repeated measures, respondents provide several measures of discrimination, typically at a daily frequency over a period of one to two weeks (Seaton, Neblett, Upton, Hammond, & Sellers, 2011; Torres & Ong, 2010). Daily RED assessment has several advantages. First, daily assessment can unpack how much of daily RED experiences are attributable to between-person (i.e., individual differences between participants) and within-person effects (i.e., daily variability due to context or person × context interaction). This enables research to investigate the relationship between RED and other outcomes (e.g., mental health) at both daily (within-person) and person (between-person) levels. Second, daily assessments minimize retrospective recall biases resulting in more accurate reports (Bolger, Davis, & Rafaeli, 2003; Iida, Shrout, Laurenceau, & Bolger, 2012). Third, although RED does not occur every day (Seaton & Iida, 2019;

Torres & Ong, 2010), daily assessments better approximate the prevalence of RED over time. To date, research employing daily diary designs report that RED is relatively low frequency, occurring approximately 1 to 2.44 times/week among ethnic minority adolescents and young adults. However, despite research suggesting that discrimination is not an everyday occurrence (Seaton & Douglass, 2014; Seaton & Iida, 2019; Torres & Ong, 2010; Yip, Cheon, et al., 2019; Yip, Wang, et al., 2019), it is a normative experience for most everyone (Seaton & Douglass, 2014; Seaton & Iida, 2019).

The existing evidence makes clear the utility of measuring RED at the daily level (Seaton & Douglass, 2014; Seaton & Iida, 2019), however, systematic psychometric analyses are limited. In the absence of psychometric analyses of daily measures, researchers have modified existing multi-item person-level measures (e.g., Daily Life Experiences; Seaton, Yip, & Sellers, 2009) or selected single items to assess daily RED (e.g. Hoggard, Byrd, & Sellers, 2015; Seaton & Iida, 2019). For example, the Everyday Discrimination Scale (EDS, Williams, et al., 1997) has been adopted as a daily RED measure while the psychometric properties of this adaption have not yet been investigated (Goosby, Cheadle, & Mitchell, 2018; Yip, Cheon, et al., 2019). Two key aims motivate this study: first, when administered daily, it is unclear how *each* EDS item contributes to the overall measure. The EDS is a compilation of multiple items, each measuring a discrete event (e.g., being threatened or harassed). When items are aggregated, it is unclear if one or two items are driving the overall prevalence of RED or if all the items are necessary to represent an individual's daily discrimination experiences. This analysis has practical applications as measures that are administered daily are optimized for brevity to minimize participant burden. Second, current studies have not partitioned the relative and independent contributions of the daily-level (within-person) and person-level (between-person) variability. The inability to separate within and between variability leaves open whether RED is an inherently person-level phenomena (i.e., a characteristic of the respondent) or a (person × environment) interaction, which raises concerns related to construct validity across the levels. The ability to distinguish sources of variability has important implications for interventions aimed to disrupt the negative effects of discrimination stress. To date, studies that investigate average scores across multiple days run the risk of conflating results at different levels, this introduces measurement error in making comparisons across groups. To address these limitations and to disentangle the measurement model for within- and between-person data, the multilevel extension of the 2PL IRT model is employed to account for the nested data structure.

Current Study

Using daily diary data collected from African American, Asian American, and Latinx adolescents across 14 days, this study used multilevel IRT to study the psychometric properties of two ethnic/racial discrimination measures: the Everyday Discrimination Scale (Williams, et al., 1997), and the Racial/Ethnic Discrimination Index (REDI, Wang & Yip, 2020). Both measures were administered at the daily and person levels to the same adolescent sample. Measurement invariance across racial/ethnic groups (African American, Asian American, and Latinx) and gender is also investigated. The results of these analyses form the basis of our recommendations for assessing daily experiences of RED.

Method

Participants

The data for this study came from a four-year longitudinal study among ethnic/racial minority adolescents (Wang & Yip, 2019). A total of 350 students from five high schools in a large, urban area participated and the study was approved by the relevant Institutional Review Boards. The schools were ethnically/racially diverse, as indicated by the Simpson's (1949) diversity index (median = .51, ranged from .29 - .62), which means the probability that two randomly selected students within the same school were from different ethnic/racial groups (Wang & Yip, 2020). On average, these schools had 31% African Americans (4% to 63%), 15% Asian Americans (3% to 57%), 46% Latinx (21% to 50%), and 6% Whites (2% to 16%). There were 108 males (31%; M_{age} = 14.31, SD = 0.61) and 242 females (69%; M_{age} = 14.26, SD = .61). The sample included 76 (22%) participants identifying as African American, 145 (41%) as Asian American, and 129 (37%) as Latinx. Among the Asian Americans: 69% identified as Chinese, 8% as Korean, and 23% as other. Among the Latinx, 43% identified as Puerto Rican, 14% as Dominican, 11% as Mexican, 7% as South American, 4% Central American, and 18% as other. Participants reported their parents' highest level of education as high school or below (36% for mothers, 36% for fathers), and as above high school (34% for mothers, 23% for fathers). Many participants did not know their parent's highest education levels (30% for mothers, 41% for fathers).

Procedures

All African American, Asian American, and Latinx ninth grade students in five schools in a large, urban area were invited to participate in a study focused on stress and development. Invitation letters were mailed to parents and only students with parental consent participated. Across the schools, the response rate varied from 6% to 31%, reflecting the challenging circumstances of the participating schools (i.e., 52% to 86% socioeconomically disadvantaged students). Many home addresses were not verifiable and invitation letters were returned as undeliverable. Data were collected in five cohorts (cohort 1: n = 90, cohort 2: n = 103, cohort 3: n = 25, cohort 4: n = 67, cohort 5: n = 96) across successive school years from 2015 to 2017. The cohorts did not differ on age, gender, race/ethnicity, or mother's and father's education level.

Participants met in groups of one to ten students after school dismissal, were assented to the study and completed an online demographic questionnaire. Participants were then given a data-enabled electronic tablet to access the five to seven-minute daily web-based daily diary survey, including Everyday Discrimination Scale (Williams et al., 1997) and Racial/Ethnic Discrimination Index (Wang & Yip, 2019). To measure person-level demographic information, a baseline assessment (pre-measure) was conducted before the beginning of daily diary survey. Participants were instructed to complete the daily diary survey before bedtime for 14 days. Research assistants monitored the compliance of the daily diary survey, and the average completion rate was 79% (= 11.1 days - out of maximum 14 days, SD = 9%, range = 65-97%). At the end of the 14 days, participants returned the tablet, completed another online survey and were compensated $20. All surveys were administered on Qualtrics.

Measures

Everyday Discrimination Scale (EDS).—The five-item EDS (Williams et al., 1997) is a short version of self-report measure that assesses participant's experiences of discrimination over a prescribed period of time. We explored a new five-item EDS developed based on the results of Stucky et al. (2011) and the characteristics of diverse racial/ethnic minority adolescents. In particular, we replaced three out of five items of Stucky et al. (2011) five-item version to account for the different sample characteristics. For example, "insulted" item was replaced with its locally dependent item "harassed," and "dishonest" item was replaced with "afraid" item as study suggested that Asian American students reported experiences of physical and verbal harassment by peers at school, while African American and Latinx students reported discrimination by adults, such as teachers, police, and shopkeepers (Rosenbloom & Way, 2004). Due to the comment about African American and Latinx adolescents' experiences with shopkeepers, we chose to include the "service" item. Furthermore, we removed the "better than you" item as Kim et al (2014) cautioned the use of this item for the Asian and Latinx populations (See Supplemental Materials, Table 2 for detailed comparison).

The EDS has been adapted for daily use with African American adolescents (Goosby et al., 2018). For the daily diary survey, instructions were modified to reflect the day's experiences (e.g. "Today, I received poorer service than others in restaurants or stores"). Branching logic assessed discrimination frequency and attribution to race/ethnicity with separate items. Specifically, the discrimination frequency dimension (EDS-DF; Table 1) asked participants whether they experienced discrete instances of discrimination (0 = no, 1 = yes). If participants indicated "yes", they were asked the degree to which the experience was attributed to race/ethnicity (EDS-AER): "how sure are you that this happened because of your race/ethnicity?" on a four-point scale (0 = not at all, 1 = not very sure, 2 = somewhat sure, 3 = very sure).

Racial/Ethnic Discrimination Index (REDI).—REDI is a six-item self-report measure developed to assess ethnic/racial discrimination (Table 1, Wang & Yip, 2019). For daily administration, the REDI includes a three-point scale (0 = did not happen/was not a problem today, 1 = somewhat of a problem today, and 2 = very much a problem today).

Analytic Plan

Analyses were performed on EDS-DF, EDS-AER and REDI. At the daily level, the missing data rate for EDS-DF, EDS-AER and REDI are 23%, 22.74% and 22.8%, respectively. Daily observations with no responses to all the items is the main contributor of the missingness, and were excluded from analyses by Mplus, resulting in a final analytic sample of 3799 observations for EDS scales and 3784 observations for REDI. In the final sample, the correlation between the missingness of EDS-DF, EDS-AER and REDI items and gender, nativity, race and parents' education (rs < .04) were examined and there were no significant correlations, suggesting that the missingness in daily items were missing completely at random. Maximum likelihood estimation accounted for the missing responses (Shafer & Graham, 2002). The low response frequency of categories (fewer than 1%) caused data sparseness and could make IRT model estimation unstable, therefore, response options were

collapsed for analysis. For EDS-AER items, response 1 (not very sure), 2 (somewhat sure) and 3 (very sure) were collapsed[1]. The outcomes of EDS-AER items were dichotomized into two categories with no (= 0) and yes (= 1 or 2 or 3). For all REDI items, response options 1 (somewhat a problem today) and 2 (very much a problem today) were collapsed. All items of REDI were collapsed into dichotomous variables.

Given the multilevel data structure of daily diary observations (level 1) nested within participants (level 2), the variability of the EDS-DF, EDS-AER and REDI items as well as the unweighted composite scores of EDS-DF, EDS-AER and REDI at both levels were assessed using the intraclass correlation (ICC). The ICC is the ratio of variance at level 2 to the total variance (Kaps & Lamberson, 2009)[2], with zero indicating that the variance is completely attributable to level 1 differences (i.e., daily differences) and one indicating that the variance is completely attributable to level 2 differences (i.e., individual differences). The average ICC of EDS-DF items was .61 (ranging from .45 to .70) and the ICC of their unweighted composite score was .48. The average ICC of EDS-AER items was .69 (ranging from .50 to .80) and the ICC of their unweighted composite score was .56. The average ICC of REDI items was .74 (ranging from .69 to .76) and the ICC of their unweighted composite score was .69. These ICCs indicate that more than half of the variance in daily discrimination experiences can be attributed to individual differences; however, contextual and daily variability is also evident. These values are consistent with the ICC of African American adolescents' daily RED (= .64) from Seaton and Iida (2019). Seaton and Iida conducted a daily diary assessment of 103 African American adolescents (13-18 years old) using the 18-item daily life experiences subscale of the Racism and Life Experiences Scale (Seaton, Yip, & Sellers, 2009). Since this study shares an inclusion of African American adolescents, a 14-day daily diary assessment, and an adolescent sample, the ICC of daily RED should be comparable. In summary, results support the use of multilevel IRT to capture the daily and individual differences.

Multilevel Item Response Theory (IRT) Analysis.—For the dichotomous EDS-DF, EDS-AER and REDI items (with the collapsed 2 category responses), the following two-level IRT models were fit for 1-PL, two-parameter logistic (2-PL) and 2-PL variant models. The probability function of an item in the two-level 2-PL model is written as:

$$P(q_{ijk} = 1 \mid \boldsymbol{\theta_{jk}}, \alpha_i, \lambda) = \frac{1}{1 + e^{-\alpha_i + \lambda_i^{day}\theta_{jk} + \lambda_i^{person}\theta_k}}$$

where $P(\cdot)$ is the probability function, q_{ijk} is the response (0 = not being discriminated or 1 = being discriminated) of the dichotomous item i on day j by person k. The vector $\boldsymbol{\theta_{jk}}$ contains the daily level latent trait θ_{jk} and the participant level latent trait θ_k. The latent

---

[1] The frequency of response "1" is relatively low in our sample. Collapsing response "1" and "0" yielded consistent results.

[2] The ICCs of the EDS discrimination frequency and REDI items were calculated using this formula: $\sigma_\mu^2/(\sigma_\mu^2 + \pi^2/3)$, where σ_μ^2 is the level-2 variance of the item estimated by the two-level logistic regression model (daily diary observations nested within participants) of each item without any predictor. The ICCs of the unweighted composite scores of EDS discrimination frequency and REDI were calculated using this formula: $\sigma^2/(\sigma^2 + \sigma_\mu^2 + \pi^2/3)$, where σ^2 and σ_μ^2 are the level-3 and level-2 variances estimated by the three-level logistic regression model (items nested within daily diary observations nested within participants).

traits are assumed to be normally distributed and are standardized for model identification. The item intercept a_i reflects the relationship between the item difficulty and loading parameter (difficulty $b_i = -a_i/\lambda_i$). The difficulty-loading form does not generalize well to the multilevel IRT model, the intercept-loading parameterization was adopted for all IRT models. The vector $\boldsymbol{\lambda}$ contains the item loading on the latent trait at the daily level λ_i^{day} and participant level λ_i^{person} (i.e., item differential capability to detect discrimination). Higher loading indicates stronger item differential capability to detect discrimination.

For multilevel model identification, either a loading at each level of analysis (daily and person) has to be constrained to 1 or constraining the variances of the random terms to 1. We constrained the loading to 1 in the multilevel IRT models for the purpose of comparing the variability of each level of analysis. If λ_i^{person} are supposed to have the same differentiate capacity across persons, then λ_i^{person} are constrained to 1 and the 2-PL model becomes the 2-PL(P) model. Similarly, the 2-PL(D) is another variant of 2-PL model which the λ_i^{day} are assumed to have same differentiate capacity across time. If all the item loadings λ_i^{day} and λ_i^{person} are set to 1, then the 2-PL model reduces to the 1-PL model. Before testing these IRT models, the item correlations of the EDS discrimination frequency and REDI were calculated. The likelihood ratio tests were used to comparing nested IRT models, and Akaike information criterion (AIC), and Bayesian information criterion (BIC) were reported as supporting evidence for model comparisons and were used for comparing non-nested models. For the likelihood ratio test, a non-significant result between the two models (e.g., 1-PL vs. 2-PL) supported the more restricted model (1-PL). Models with smaller AIC and BIC had better model fit.

Measurement Invariance across Gender and Race/ethnicity.—After selecting the best fitting multilevel IRT models for EDS-DF, EDS-AER and REDI, measurement invariance across gender (male and female) and race/ethnicity (African American, Asian American, Latinx) was tested. First, we used multi-group analysis (Liu et al., 2017; Millsap & Yun-Tein, 2004) to test the invariance of the selected model as a whole, which involves a stepwise sequence of models that constrained the item loadings and item intercepts across groups was fit. Consecutive invariance models were tested by the likelihood ratio test. Models were also compared using AIC and BIC. Second, we used multilevel logistic regression to test the differential item functioning (DIF) of each item. When grouping is a between cluster variable, DIF analyses is straightforward at this level as it does not vary within daily level (Ryu, 2013). Therefore, analysis was conducted by add person level direct effects to the multilevel-IRT model from the grouping variables to items to capture uniform DIF with respect to thresholds. DIF was tested by likelihood ratio test that comparing model with and without direct effect of grouping variable.

Results

The examination of response frequency was conducted before the IRT analysis. For the EDS, 155 (45%) participants reported experiencing discrimination (endorsed at least

one item over the 14-days of the study). Item 3 (not smart) was the most endorsed. Consistent with previous findings, endorsements for EDS items were relatively low occurring approximately less than once over the course of 14 days (M_{DF} = .30, SD_{DF} = .65; M_{AER} = .18, SD_{AER} = .53). There were no gender (DF: $t(348) = -.64$, $p = .52$; AER: $t(348) = -.04$, $p = .97$) or racial/ethnic differences (DF: $F(2, 347) = 1.57$, $p = .21$; AER: $F(2, 347) = .67$, $p = .51$). For the REDI, 30% of the sample endorsed at least one REDI item, and item 2 (felt stress because of race/ethnicity) was the most commonly endorsed item. Participants endorsed an average of 0.47 incidents ($SD = 1.37$) over the 2 weeks. There were no gender ($t(226) = -.80$, $p = .42$) or ethnic/racial differences ($F(2, 347) = .14$, $p = .87$).

Table 2 shows the item correlations for EDS-DF, EDS-AER and REDI at the daily level (upper diagonal) and person level (lower diagonal). At the person level, EDS-DF, EDS-AER and REDI items had high correlations among each other (rs > .87), suggesting that the items functioned similarly at this level for both scales. At the daily level, EDS-DF and EDS-AER had low to moderate correlations (rs ranging from .07 to .71), suggesting that these items functioned differently at this level. On the other hand, REDI items had high correlations among each other at this level (rs > .88), suggesting that REDI items functioned similarly at this level.

Multilevel IRT.

Two-level 1-PL and 2-PL models for the EDS-DF, EDS-AER and REDI were fit to the data (see Table 3). The likelihood ratio tests showed that the 2-PL demonstrated better fit than the 1-PL, and the AIC and BIC also generally favored the 2-PL model for EDS-DF, EDS-AER and REDI. In the 2-PL model of EDS-DF, EDS-AER and REDI, all the person-level item loadings were very similar to each other. Therefore, a variant of the 2-PL model where person-level item loadings were constrained to be equal across all items was considered. The AIC and BIC supported that the fit of this model was better than all other models for EDS-DF, EDS-AER and REDI and this 2-PL model was selected for interpretation and measurement invariance analysis.

The item intercepts and daily-level item loadings of EDS-DF, EDS-AER and REDI are presented in Table 3. The item intercepts and loadings of EDS-DF and EDS-AER were very similar. To better capture the RED, the following analyses focus on the results of EDS-AER. All EDS-AER and REDI items had high intercepts (EDS-AER: intercepts > 2; REDI: intercepts > 1.6), suggesting that only participants with high levels of the RED would likely to respond "being the target of discrimination" on these items. For EDS-AER, the daily-level item loadings ranged from .24 (item 4) to .84 (item 1) while all the person-level item loadings equaled 1. The higher the item loading value, the higher the probability of endorsing "being the target of discrimination" as the RED level increases. In other words, higher item loadings reflected that the item had a stronger ability to differentiate a participant who experience discrimination from another participant who did not. Results suggest that the EDS-AER items had stronger differentiation ability at the person-level than the daily-level; and not all EDS-AER items differentiated the daily observations equally well. For REDI, the daily-level item loadings ranged from .87 (item 4) to .96 (item 1) while

the person-level item loadings equaled 1, suggesting that all items had strong differentiation ability at both the daily- and person-levels.

To further describe the relationships between the model parameters and item responses, Figure 1 (panel A) shows the item characteristic curves of EDS-AER and REDI at the person level. The vertical axis is the response probability of the item and the horizontal axis is the z score of the latent trait at the person-level. Different lines in the plot represents different items. The daily-level latent trait z score was set to zero. Both the EDS-AER items and the REDI items functioned similarly across person-level RED. In other words, when used as a measure of person-level RED, EDS-AER and REDI should generate similar results.

Figure 1 (panel B) shows the item characteristic curves of EDS-AER and REDI at the daily level. It shows that some EDS-AER items had flatter curves due to low item loadings, some items had weak differential ability between participants with similar levels of RED. In other words, the items were not informative enough to differentiate participants with low or high levels of RED. For example, EDS item 4 ("People act as if they are afraid of me.") had the lowest loading (.39) and the flattest curve. A participant who had the person-level latent trait z score = 0, and a daily-level latent trait z score of 5.0 only had a 20% chance of reporting RED. On the other hand, all REDI items had steep curves (larger item loadings), indicating that the REDI items have higher loadings at the daily level. However, all of the REDI item characteristic curves at the daily level provided no item information about the latent trait z scores < 0, suggesting restricted utility to differentiate participants with low levels of RED.

Figure 2 shows the item information function (IIF) and test information function (TIF) of EDS-AER and REDI at the daily level. The person-level latent traits were set to zero. The item information curve is a function of the item parameters (intercepts, loadings) of each item separately, and the test information curve is the sum of item information functions of the whole scale. Item information > .1 is regarded as satisfactory (de Ayala, 2008). The item and test information curves can be used to show the precision (reliability) of the item and the test. Results showed that the EDS-AER items did not provide much information. Only item 1 ("Today, I was treated with less courtesy or respect than other people.") reached item information of .1. These results suggest that the EDS-AER items provided little precision to the scale. The highest daily-level information occurred at around the latent trait z score of 4, meaning that item 1 provided the highest precision at higher range of RED across latent trait. Therefore, these items have more utility for populations with higher RED levels. The test information curve had consistent results. EDS-AER provided information > .1 when the latent trait z score > 0.0, suggesting that the scale is appropriate for population with higher level of RED.

For the REDI, all the item characteristic curves were similar in shape, indicating that the items had very similar information. All REDI items provided the most information when the daily-level RED equaled 2 and provided limited information (< .1) when the daily-level latent trait z score < 0.0 or > 4.0. The REDI items had moderate precision within a narrow range of the daily-level latent trait. The item information curves also suggested that item 2 ("I felt stress because of my race/ethnicity.") provided the least information and item 3

("Others treated me poorly because of my race/ethnicity.") provided the most information. The highest daily-level information occurred at around the latent trait z score of 2, meaning that item 3 is more accurate at detecting higher RED levels. The test information curve had consistent results. REDI provided information $> .1$ when the latent trait z score > 0.0, suggesting the scale is appropriate for populations with higher level RED. Taken together, the results suggest that if a researcher is interested in reducing the number of items, item 2 can be dropped. If a researcher is interested in reducing REDI to a single item, item 3 should be used. Test information is the sum of all the item information in a scale, both the item precision and test length can affect the value. Even the REDI had higher test information than EDS-AER, neither REDI nor EDS-AER had adequate test information (> 4) across latent trait, suggesting that the measurement precision is not acceptable.

Measurement Invariance across Gender and Race/Ethnicity.

The likelihood ratio tests, AIC, and BIC of the factorial invariance test of EDS-AER and REDI across race/ethnicity (African Americans, Asian Americans, Latinx) and gender (males and females) are presented in Table 5. For EDS-AER, the likelihood ratio tests, AIC and BIC all supported that the item daily-level loadings and intercepts were invariant across race/ethnicity and gender. Therefore, the measurement invariance of item daily-level loadings and intercepts of EDS-AER across race/ethnicity and gender was supported.

For REDI, the likelihood ratio tests and AIC did not support invariance across race/ethnicity for the daily-level item loadings and intercepts. A careful examination of the baseline model showed that the daily-level loading and intercept for item 2 ("I felt stress because of my racial/ethnic.") were lower for African Americans compared to Asian Americans and Latinx, African Americans were more likely to endorse the item compared to Asian Americans and Latinx and the item had smaller loading for African Americans at the daily-level. The invariance model was revised by allowing the daily-level loading and intercept for item 2 to vary across race/ethnicity. The likelihood ratio test, AIC, and BIC supported the partial loading invariance model. BIC also supported the partial daily-level loading and intercept invariance model. Therefore, the measurement invariance of the daily-level loadings and intercepts (with the exception of item 2) across race/ethnicity was supported for the REDI measure and future work should consider administering the REDI excluding item 2. The likelihood ratio tests, and BIC supported that the item daily-level loadings and intercepts were invariant across gender. Examination of the item daily-level loadings and intercepts in the baseline model without these parameters constrained to be invariant across gender did not show any estimates to be different across gender. Therefore, the measurement invariance of the REDI was supported across gender.

To the best of our knowledge, there was no simulation study that has systematically examined the required sample size for multilevel factorial invariance test for categorical items (Liu et al., 2017). Thus, we conducted post-hoc analysis of the factorial invariance tests across gender and racial/ethnic groups through simulations (See Supplemental Materials, Table 3). For each invariance test, 500 datasets were simulated based on the parameters of loading and intercept constrained model for each gender and race/ethnic group. Power was estimated as the proportion of datasets that support the more constrained

invariance model for each comparison. Overall, results showed satisfactory statistical power (> .8) for the factorial invariance tests across gender and racial/ethnic groups using likelihood ratio test, AIC and BIC.

Besides the factorial invariance of EDS-AER and REDI scales across ethnic/racial and gender group, multilevel logistic regressions were conducted to test the differential item functioning (DIF) of each item (Table 6). Unlike the results of factorial invariance, EDS-AER items have shown significant DIF. Specifically, item 1 ("Today, I was treated with less courtesy or respect than other people"), item 3 ("People act as if I was not smart.") and item 5 ("Today, I was threatened or harassed.") showed significant DIF between Asian American versus African American and Latinx. Item 1 and item 3 also showed significant DIF between gender group. Further examination of the direct effects suggested that Asian American ($\beta = -1.64$) have less reported discrimination for item 1 and item 5 ($\beta = -2.03$) compared with African American and Latinx and more reported discrimination for item 3 ($\beta = 1.10$). Male have less reported discrimination for item 1 ($\beta = -1.24$), and more for item 3 ($\beta = 0.80$) compared with female. No significant DIF was observed between race and gender groups for REDI items.

Discussion

This study evaluated the psychometric properties of two daily and person-level discrimination measures, the modified daily Everyday Discrimination Scale (EDS) and the Racial/Ethnic Discrimination Index (REDI). The purpose of this study is to assess the psychometric properties of two measures of RED at both the person- and daily-level and to provide practical suggestions for future use. The study was conducted on a sample of adolescents who participated in a 14-day daily diary study. The results supported a single item solution for both scales as a daily measure.

The two measures were assessed and compared with several multilevel IRT models. The multilevel IRT models allow us to make comparative evaluations across time and subject level by looking at IRF and IIF plots conditioned on different levels of subjects/time. Both the EDS-DF and EDS-AER supported a 2-PL model with constrain on person-level item loading. Besides a lower response rate, there were no evident differences between EDS-DF and EDS-AER, for example, item 3 ("not smart") was most common item and item 4 ("afraid of me") was the least informative item for both measures. The parameter estimates for EDS-DF and EDS-AER were very close, suggesting that even without the attribution to ethnic/racial, discrimination frequency items remain good indicators of RED.

All items in EDS-AER and REDI had substantial item intercepts. High intercepts indicate the most informative location on the latent traits. The items are more informative for participants with high levels of RED. The person-level item loadings were all equal to 1 while the daily-level item loadings in EDS-AER ranged from .24 (item 4) to .84 (item 1). The daily-level item loadings for the REDI ranged from .87 (item 4) to .96 (item 1). The loadings are also illustrated in the item response function curves (Figure 1). Item loadings represent the ability to differentiate among participants on the basis of levels of RED. Therefore, the results of both scales suggest that the items had stronger differentiation ability

at the person- versus the daily-level. Comparing the two RED measures, the EDS-AER and REDI items did not differentiate the daily observations equally well. Item 4 ("afraid of me") of EDS-AER had low item loading at the daily level, suggesting that the item had weak differential ability between participants, while all REDI items showed high item loadings, indicating that REDI items had greater differential ability between participants. The results suggest that REDI is a more informative indicator of daily-level RED.

There are conventions for acceptable thresholds for item loadings at the daily level. Because the loadings reflect the degree to which an item, and the test as a whole, are measuring a latent trait, item loadings are interpreted in the context of the test which is being analyzed. Items with low loadings (e.g., EDS item 4) may be ambiguously worded or include a broader, less specified, range of experiences. EDS item 4 ("People act as if they were afraid of me.") refers to a wide range of behaviors with a higher level of ambiguity. As such, this item may have less differentiating information compared to more concrete items (e.g., "Today, I was threatened or harassed."). Therefore, it is reasonable to consider eliminating this item. Ideally, items with high levels of differential ability/loading and whose intercepts cover the interested latent construct continuum are retained. However, in practice, some items with large loadings cannot provide enough information for the interested latent construct continuum. Considering the fact that discrimination is inherently a daily-level construct, the lower range on the continuum of daily-level RED was of particular interest. It would be appropriate for scale construction to select items with smaller loadings that can add information to the lower range of RED, too.

One advantage of item response function curves (Figure 1) is that it provides a straightforward visualization of the function of each item on the continuum of latent trait. All REDI items functioned similarly, with high item intercepts and loadings. The REDI item response function curves of items overlapped with each other, suggesting that the items are interchangeable. All REDI items had their maximal information at the higher end of latent trait continuum, indicating that these items are more informative for participants with higher RED levels. These IRT results can be interpreted similar to a reliability, but the information can vary based on the location along the latent trait continuum. REDI item 2 ("I felt stress because of my race/ethnicity.") had the lowest information, while item 3 ("Others treated me poorly because of my race/ethnicity.") had the highest information. In practice, if only one REDI item is retained to measure RED, it should be the item with the highest information, in this case, item 3. Item 3 may be a more informative indicator of discrimination because it clearly implicates another person in an interaction, whereas item 2 does not make clear the source of the stress. Future studies should seek to understand how these items may be understood and interpreted differently by racial/ethnic minority respondents.

Because RED measures are often employed in large multi-group studies, it is important to establish measurement invariance. Often researchers want to compare prevalence rates across different racial/ethnic groups or compare experiences of the same group in different locations. Factorial invariance tests and multilevel logistic regression were conducted for EDS-AER and REDI to investigate whether group differences could be attributed to true differences in race/ethnicity or gender. The factorial invariance test results showed that daily-level loadings and intercept invariance was achieved for gender for both EDS-AER

and REDI. The same type of invariance was also supported for EDS-AER across race/ ethnicity. On the other hand, the daily-level loading and intercept invariance was not achieved for race/ethnicity for REDI. The DIF test, on the contrary, suggested that EDS-AER items failed to be invariant across gender and race group. Specifically, male have less discrimination reported for item 1("Today, I was treated with less courtesy or respect than other people"), and more for item 3 ("People act as if I was not smart.") compared with female. Previous studies testing gender as a covariate showed that male reporting more experiences with everyday discrimination (Seaton et al., 2008; Sellers, & Shelton, 2003), and Stucky et al. (2011) found that "not as smart" is non-invariant across. Our results confirm the findings with item 3, while the difference on item 1 suggesting that the difference between gender on feeling "less respect" might not reflecting the true gender difference on RED either.

The factorial invariance test suggested that the daily-level loadings and intercept for item 2 ("I felt stress because of my race/ethnicity") were lower for African Americans as compared to Asian Americans and Latinx, such that the item is more likely to be endorsed by African Americans than by Asian Americans and Latinx. In other words, the item is better able to distinguish participants with higher daily-level RED among African Americans than among Asian Americans and Latinx. A partial invariance model was established by removing the equal constraint for item 2, suggesting a partial measurement invariance of item daily-level loadings and intercepts (with the exception of item 2) for the REDI measure across race/ ethnicity. While DIF test suggested that the DIF indeed existed in three out of five items between Asian American versus African American and Latinx. Asian American reported less discrimination on item 1 ("less respect") and item 5 ("threatened or harassed") and more on item 3 ("not smart") compared with African American and Latinx. Previous studies have found that item "… acted as if they're better than you" have lower threshold for Hispanic and Asian groups compared to non-Hispanic white and Black groups (Kim, Chiriboga, & Jang, 2009). Most studies test the invariance of EDS focuses on compare African American with other ethnic/racial groups, for example, Shariff-Marco et al (2018) found that African Americans were less likely to endorse items like "…treated with less respect", "… acted as if they're better than you" and "… been threatened/harassed.", and Lewis et al. (2012) found that DIF existed on item "…treated with less courtesy" compared African American to Caucasian group and item "…receiving poorer service" compared African American with other groups (Hispanic, Chinese, Japanese and Caucasian). Our findings are intriguing that Asian American have differential responses in three out of five EDS-AER items. Giving Asian American RED are less studies in literature, and immigration-related factors might affect their RED, researchers using EDS studying RED should be careful about non-invariant items. Taken together, REDI would be recommended to study RED.

Implications and Conclusions

Two conclusions emerge from the current study. First, for both the Everyday Discrimination Scale and the Racial/Ethnic Discrimination Index, items functioned differently across person- and daily-level, suggesting that researchers interested in the construct of discrimination need to align measure selection with the level of interest/analysis. Previous research using daily indicators of RED shows that it is an infrequent experience with RED

occurring approximately 1 to 2.44 times/week among ethnic minority adolescents and young adults days over a two-week period, however, the percentages of adolescents reported at least one discriminatory experiences has been observed to be as high as 97% (Huynh & Fuligni, 2010; Seaton & Douglass, 2014; Torres & Ong, 2010; Wang & Yip, 2019; Yip, Cheon, et al., 2019). Despite its relative infrequency, these data suggest that daily RED measures complement scales measured at the person-level.

Second, the REDI is a more informative measure of daily racial/ethnic discrimination compared to the EDS. This recommendation is aligned with the original purpose of the measures. That is, the REDI was designed as a daily-level indicator whereas the EDS was developed as a person-level measure; however, in the absence of a validated daily measure, the EDS has been adapted for daily use. Previous studies indicate that the EDS-DF yields a higher percentage of respondents reporting discriminatory experiences than directly querying participants about race related discriminatory experiences (Shariff-Marco et al., 2011; Williams & Mohammed, 2009). However, due to the response burden of daily dairy approach, the response rates were much lower than other study formats as the mean score of EDS range from 0.09 to 0.32 using daily dairy approach (Seaton & Iida, 2019; Torres & Ong, 2010; Yip, Wang, et al., 2019) while the estimates were closer to a 2 with one-time assessments (e.g. Krieger et al., 2005; Lewis et al., 2012; Schulz et al., 2006). Relative to existing research (e.g. 0.12, Seaton & Iida, 2019; 0.32, Torres & Ong, 2010), the average counts of discrimination occurrences on the EDS-DF (.30), EDS-AER (.18) and REDI (.47) were lower in the current study, especially for the EDS-AER. Although the response rate for EDS-DF is higher, it does not offer a clear advantage over measures that were explicitly developed for daily use. In order to be comparable with REDI, the current study mainly focused on the item characteristics of EDS-AER. Given how infrequent daily RED responses were in the current and other samples (Seaton & Douglass, 2014; Seaton & Iida, 2019; Torres & Ong, 2010), researchers who are interested in RED may want to use the REDI which makes a racial/ethnic attribution, or EDS-AER which incorporates an attribution of racial/ethnic into the question stem.

For REDI, all items functioned very similarly according to the item loadings and intercepts. Taking a close look at the content of each question, the questions can be grouped into two aspects: stressful ("… feel stress …", "…feel unsafe…", "… feel uncomfortable …") and unfair ("… treated unfairly …", "… was teased …", "… treated me poorly …"). The item response functions of REDI items were largely overlapping, indicating that all questions are asking similar things; there is no difference between the two aspects mentioned before in terms of the responses (Table 1). It is highly possible that if someone is the target of RED on a given day, he/she would answer "yes" to all six items. These results suggest that the six-item scale can be reduced to one item. The item information curves show that item 2 ("I felt stress because of my race/ethnicity.") provided the lowest information, and item 3 ("Others treated me poorly because of my race/ethnicity") has the highest information. To reduce respondent burden, the recommendation is to reduce the REDI scale to a single-item scale retaining item 3.

Although the current study makes significant contributions to the measurement of daily discrimination experiences, several limitations need to be considered. First, response

categories of REDI were dichotomized for analysis due to low frequencies, and it is unclear if the difference between "was somewhat of a problem today" and "was very much a problem today" is meaningful, or if the distinction really lies between the presence and absences of RED on a given day. Second, although the results supported the 2-PL model with all factor loadings constrained to be the same at the person level for the EDS-DF, EDS-AER, and REDI, future research is needed to replicate this finding. Third, the sample size of our study may have limited statistical power to test for highly parameterized IRT models (e.g. 3PL and 4PL). Our sample size ($N = 350$) is above the average sample sizes of published daily diary assessments (Average $N = 233$; English et al., 2020; Hoggard, Byrd, & Sellers, 2015; Huynh & Fuligni, 2010; Seaton & Douglass, 2014; Seaton & Iida, 2019; Torres & Ong, 2010). Moreover, our number of level-1 observations are comparable with studies that implemented multilevel IRT analysis (Fox, 2004; Fox, 2005; Gorter et al., 2015). Our sample size allows us to test the 1-PL and 2-PL models for EDS-DF, EDS-AER, and REDI but not the more complicated 3-PL and 4-PL models. For example, the 4-PL models have addition lower and upper asymptote parameters of each item response function, which captures the behaviors that the participants who did not experience RED reporting as such out of obligation, and the behaviors that the participants who experienced RED not reporting as such. Asparouhov and Muthén (2020) suggested that sample size of $N = 5,000$ and 20,000 might be needed to estimate a 3-PL model and a 4-PL model "well", respectively. Adding priors for the model parameters to improve identifiability, yet we were unable to do so because the modified EDS and REDI are new or newly adopted for daily use and have not yet been studied as a measure for daily RED. Adding priors require switching the estimator from maximum likelihood to Bayesian. Measurement invariance test with Bayesian estimation involves the consideration of choosing the exact or approximate test (e.g., De Bondt & Van Petegem, 2015). This adds another layer of complexity. Fourth, we conducted both factorial invariance test and DIF test to study the measurement invariance for EDS-AER and REDI scales. Giving the inconsistent results, researcher should be cautious to compare RED across gender and ethnic/racial groups. To better understand whether we have enough statistical power to detect small but systematic non-invariant items, we also conducted a post-hoc power analysis for the factorial invariance tests based on the sample size and parameter estimates of each gender and race/ethnic group. Results showed that we have a satisfactory statistical power ($> .8$) for the factorial invariance tests across gender and racial/ethnic groups using likelihood ratio test, AIC and BIC. We did not conduct the post-hoc power analysis for the DIF tests because the simulation for all 33 tests was very time consuming. We believed that the comparison of our sample size with the average sample size of published daily diary assessments of ethnic/racial discrimination, as well as the post-hoc power analysis results of the factorial invariance tests provided some support that our sample size was sufficient for the analyses. We acknowledge that the lack of sample size guidelines and requirements of multilevel IRT analysis poses challenges of the generalizability of our findings. However, we believe that the benefits of our study outweigh the statistical deficiencies. In sum, giving previous studies showed ethnic/racial differences in RED measures (Berenbon, 2018; Lewis et al., 2012; Stucky et al., 2011), researchers should be cautious making racial/ethnic group comparisons. Last, although our study adds a large and diverse group of adolescents to a literature that has focused on adults, age and our

location in a large urban area may limit the generalizability of our results. These limitations point to fruitful areas for further research.

Despite these limitations this paper contributes to the extant literature and has implications for future research. Our findings suggesting that REDI is more appropriate for measuring RED at daily level and can be reduced to one item ("Others treated me poorly because of my race/ethnicity") to reduce respondent burden. Moreover, our study is the first one to look at the psychometric properties of comparing two daily RED measures. Few studies have examined item response functions considering within-person as well as between-person variability. Separating these levels of variability are important for studies that want to tease apart within- and between-person processes. For example, a within-person item function is important since daily discrimination measures are used in studies that provide implications for within- and between-person physical and mental health outcomes (Hoggard et al., 2015; Huynh, Guan, Almeida, McCreath, & Fuligni, 2016; Seaton & Iida, 2019; Torres & Ong, 2010). In studies such as these, using reliable and valid measures of daily discrimination would be critical to providing accurate information for devising effective health-related prevention and intervention strategies. Next, when measured at daily level, both the EDS-AER and REDI items suggest that incorporating a racial/ethnic attribution makes items more informative than questions that ask about unspecified discrimination, and there was not a difference in terms of response frequency in daily measures. These results suggest that the more informative REDI is a recommended for measuring daily discrimination. In the future, more studies with lager samples are needed to study the EDS at daily level with ethnic/racial attributions in question stems to exam the generalizability of our findings to other settings of daily study of RED. Finally, our study found that the REDI items functioned similarly and can be reduced to one item scale with acceptable psychometric properties. Being able to measure daily RED with one item will contribute to reducing participant burden in daily diary studies. In the future, it is important to administer the REDI scale with larger samples across longer periods of time.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Reference

Benner AD, Wang Y, Shen Y, Boyle AE, Polk R, & Cheng YP (2018). Racial/ethnic discrimination and well-being during adolescence: A meta-analytic review. American Psychologist, 73(7), 855–883.

Berenbon RF (2018). Using Rasch analysis to investigate the validity of the Everyday Discrimination Scale in a national sample. Journal of Health Psychology, 135910531880078.

Bernstein KS, Park SY, Shin J, Cho S, & Park Y (2011). Acculturation, discrimination and depressive symptoms among Korean immigrants in New York City. Community Mental Health Journal, 47(1), 24–34. [PubMed: 19888652]

Bolger N, Davis A, & Rafaeli E (2003). Diary Methods: Capturing Life as it is Lived. Annual Review of Psychology, 54(1), 579–616.

Carter RT, Johnson VE, Kirkinis K, Roberson K, Muchow C, & Galgay C (2019). A Meta-Analytic Review of Racial Discrimination: Relationships to Health and Culture. Race and Social Problems, 11(1), 15–32.

Chan KTK, Tran TV, & Nguyen TN (2012). Cross-Cultural Equivalence of a Measure of Perceived Discrimination Between Chinese-Americans and Vietnamese-Americans. Journal of Ethnic and Cultural Diversity in Social Work, 21(1), 20–36.

Clark R, Coleman AP, & Novak JD (2004). Brief report: Initial psychometric properties of the everyday discrimination scale in black adolescents. Journal of Adolescence, 27(3), 363–368. [PubMed: 15159094]

Collins PH, & Essed P (1992). Understanding Everyday Racism: An Interdisciplinary Theory. Contemporary Sociology.

de Ayala RJ (2009). In The theory and practice of item response theory. New York, NY, US: Guilford Press.

English D, Lambert SF, Tynes BM, Bowleg L, Zea MC, & Howard LC (2020). Daily multidimensional racial discrimination among Black U.S. American adolescents. Journal of Applied Developmental Psychology, 66. 101068. [PubMed: 33994610]

Fisher CB, Wallace SA, & Fenton RE (2000). Discrimination distress during adolescence. Journal of Youth and Adolescence, 29(6), 679–695.

Fox JP (2004). Applications of multilevel IRT modeling. School Effectiveness and School Improvement, 15(3-4), 261–280.

Fox JP (2005). Multilevel IRT using dichotomous and polytomous response data. British Journal of Mathematical and Statistical Psychology, 58(1), 145–172.

Gee GC, Spencer M, Chen J, Yip T, & Takeuchi DT (2007). The association between self-reported racial discrimination and 12-month DSM-IV mental disorders among Asian Americans nationwide. Social Science and Medicine, 64(10), 1984–1996. [PubMed: 17374553]

Goosby BJ, Cheadle JE, & Mitchell C (2018). Stress-Related Biosocial Mechanisms of Discrimination and African American Health Inequities. Annual Review of Sociology, 44(1), 319–340.

Gorter R, Fox JP, & Twisk JW (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. BMC Medical Research Methodology, 15(1), 1–12. [PubMed: 25555466]

Guyll M, Matthews KA, & Bromberger JT (2001). Discrimination and unfair treatment: Relationship to cardiovascular reactivity among African American and European American women. Health Psychology, 20(5), 315–325. [PubMed: 11570645]

Harnois CE, Bastos JL, Campbell ME, & Keith VM (2019). Measuring perceived mistreatment across diverse social groups: An evaluation of the Everyday Discrimination Scale. Social Science and Medicine, 232, 298–306. [PubMed: 31121440]

Harrell SP (2000). A multidimensional conceptualization of racism-related stress: Implications for the well-being of people of color. American Journal of Orthopsychiatry, 70(1), 42–57.

Hoggard LS, Byrd CM, & Sellers RM (2015). The lagged effects of racial discrimination on depressive symptomology and interactions with racial identity. Journal of Counseling Psychology, 62(2), 216–225. [PubMed: 25867694]

Huynh VW, & Fuligni AJ (2010). Discrimination Hurts: The Academic, Psychological, and Physical Well-Being of Adolescents. Journal of Research on Adolescence, 20(4), 916–941.

Huynh VW, Guan SSA, Almeida DM, McCreath H, & Fuligni AJ (2016). Everyday discrimination and diurnal cortisol during adolescence. Hormones and Behavior, 80, 76–81. [PubMed: 26853614]

Iida M, Shrout PE, Laurenceau J-P, & Bolger N (2012). Using diary methods in psychological research. In APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics. (pp. 277–305).

Kaps M, & Lamberson WR (2009). Discrete dependent variables. Biostatistics for Animal Science, 394–418.

Kessler RC, Mickelson KD, & Williams DR (1999). The Prevalence, Distribution, and Mental Health Correlates of Perceived Discrimination in the United States. Journal of Health and Social Behavior, 40(3), 208. [PubMed: 10513145]

Kim G, Sellbom M, & Ford KL (2014). Race/ethnicity and measurement equivalence of the Everyday Discrimination Scale. Psychological Assessment, 26(3), 892. [PubMed: 24708076]

Krieger N (2012). Methods for the scientific study of discrimination and health: An ecosocial approach. American Journal of Public Health, 102(5), 936–945. [PubMed: 22420803]

Krieger N, Smith K, Naishadham D, Hartman C, & Barbeau EM (2005). Experiences of discrimination: Validity and reliability of a self-report measure for population health research on racism and health. Social Science and Medicine. 61(7), 1576–1596. [PubMed: 16005789]

Lewis TT, Cogburn CD, & Williams DR (2015). Self-Reported Experiences of Discrimination and Health: Scientific Advances, Ongoing Controversies, and Emerging Issues. Annual Review of Clinical Psychology, 11(1), 407–440.

Lewis TT, Yang FM, Jacobs EA, & Fitchett G (2012). Racial/ethnic differences in responses to the everyday discrimination scale: A differential item functioning analysis. American Journal of Epidemiology, 175(5), 391–401. [PubMed: 22306556]

Liu Y, Millsap RE, West SG, Tein JY, Tanaka R, & Grimm KJ (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. Psychological Methods, 22(3), 486–506. [PubMed: 27213981]

Millsap RE, & Yun-Tein J (2004). Assessing factorial invariance in ordered-categorical measures. Multivariate Behavioral Research. 39(3), 479–515.

Mouzon DM, Taylor RJ, Woodward AT, & Chatters LM (2017). Everyday Racial Discrimination, Everyday Non-Racial Discrimination, and Physical Health Among African-Americans. Journal of Ethnic and Cultural Diversity in Social Work, 26(1–2), 68–80. [PubMed: 28286428]

Pager D, & Shepherd H (2008). The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. Annu. Rev. Sociol, 34, 181–209. [PubMed: 20689680]

Paradies Y (2006). A systematic review of empirical research on self-reported racism and health. International Journal of Epidemiology, 35(4), 888–901. [PubMed: 16585055]

Paradies Y, Ben J, Denson N, Elias A, Priest N, Pieterse A, … Gee G (2015). Racism as a determinant of health: A systematic review and meta-analysis. PLoS ONE, 10(9), e0138511. [PubMed: 26398658]

Reeve BB, Willis G, Shariff-Marco SN, Breen N, Williams DR, Gee GC, … Levin KY (2011). Comparing cognitive interviewing and psychometric methods to evaluate a racial/ethnic discrimination scale. Field Methods, 23(4), 397–419. [PubMed: 29887759]

Rosenbloom SR, & Way N (2004). Experiences of discrimination among African American, Asian American, and Latino adolescents in an urban high school. Youth & Society, 35(4), 420–451.

Ryu E (2014). Factorial invariance in multilevel confirmatory factor analysis. British Journal of Mathematical and Statistical Psychology, 67(1), 172–194.

Schmitt MT, Postmes T, Branscombe NR, & Garcia A (2014). The consequences of perceived discrimination for psychological well-being: A meta-analytic review. Psychological Bulletin, 140(4), 921–948. [PubMed: 24547896]

Schulz AJ, Gravlee CC, Williams DR, Israel BA, Mentz G, & Rowe Z (2006). Discrimination, symptoms of depression, and self-rated health among African American women in Detroit: Results from a longitudinal analysis. American Journal of Public Health, 96(7), 1265–1270. [PubMed: 16735638]

Seaton EK, Caldwell CH, Sellers RM, & Jackson JS (2008). The prevalence of perceived discrimination among African American and Caribbean black youth. Developmental Psychology, 44, 1288–1297. [PubMed: 18793063]

Seaton EK, & Douglass S (2014). School diversity and racial discrimination among African-American adolescents. Cultural Diversity and Ethnic Minority Psychology, 20(2), 156. [PubMed: 24773002]

Seaton EK, & Iida M (2019). Racial discrimination and racial identity: Daily moderation among Black youth. American Psychologist, 74(1), 117–127.

Seaton EK, Neblett EW, Upton RD, Hammond WP, & Sellers RM (2011). The Moderating Capacity of Racial Identity Between Perceived Discrimination and Psychological Well-Being Over Time Among African American Youth. Child Development, 82(6), 1850–1867. [PubMed: 21954919]

Seaton EK, Yip T, & Sellers RM (2009). A longitudinal examination of racial identity and racial discrimination among African American adolescents. Child Development, 80(2), 406–417. [PubMed: 19467000]

Sellers RM, & Shelton NJ (2003). The role of racial identity in perceived discrimination. Journal of Personality and Social Psychology, 84, 1079–1092. [PubMed: 12757150]

Shariff-Marco S, Breen N, Landrine H, Reeve BB, Krieger N, Gee GC, … Johnson TP (2011). Measuring everyday racial/ethnic discrimination in health surveys: How Best to Ask the Questions, in One or Two Stages, Across Multiple Racial/Ethnic Groups? Du Bois Review, 8(1), 159–177. [PubMed: 29354187]

Shariff-Marco S, Gee GC, Breen N, Willis G, Reeve BB, Grant D, … others. (2009). A mixed-methods approach to developing a self-reported racial/ethnic discrimination measure for use in multiethnic health surveys. Ethnicity & Disease, 19(4), 447. [PubMed: 20073147]

Stucky BD, Gottfredson NC, Panter AT, Daye CE, Allen WR, & Wightman LF (2011). An Item Factor Analysis and Item Response Theory-Based Revision of the Everyday Discrimination Scale. Cultural Diversity and Ethnic Minority Psychology, 17(2), 175–185. [PubMed: 21604841]

Sue DW (2010). Microaggressions in everyday life: Race, gender, and sexual orientation. John Wiley & Sons.

Torres L, & Ong AD (2010). A daily diary investigation of latino ethnic identity, discrimination, and depression. Cultural Diversity and Ethnic Minority Psychology, 16(4), 561–568. [PubMed: 21058819]

Tynes BM, Giang MT, Williams DR, & Thompson GN (2008). Online Racial Discrimination and Psychological Adjustment Among Adolescents. Journal of Adolescent Health, 43(6), 565–569.

Wang Y, & Yip T (2019). Sleep Facilitates Coping: Moderated Mediation of Daily Sleep, Ethnic/Racial Discrimination, Stress Responses, and Adolescent Well-Being. Child Development, cdev.13324.

Williams DR, Yu Y, Jackson JS, & Anderson NB (1997). Racial differences in physical and mental health. Journal of Health Psychology, 2(3), 335–351. [PubMed: 22013026]

Williams DR, & Mohammed SA (2009). Discrimination and racial disparities in health: Evidence and needed research. Journal of Behavioral Medicine, 32(1), 20–47. [PubMed: 19030981]

Williams DR, & Williams-Morris R (2000). Racism and mental health: The African American experience. Ethnicity and Health, 5(3–4), 243–268. [PubMed: 11105267]

Yip T, Cheon YM, Wang Y, Cham H, Tryon W, & El-Sheikh M (2019). Racial Disparities in Sleep: Associations With Discrimination Among Ethnic/Racial Minority Adolescents. Child Development, cdev.13234.

Yip T, Wang Y, Mootoo C, & Mirpuri S (2019). Moderating the association between discrimination and adjustment: A meta-analysis of ethnic/racial identity. Developmental Psychology, 55(6), 1724–1298.

Ziegert JC, & Hanges PJ (2005). Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. Journal of Applied Psychology, 90(3), 553.
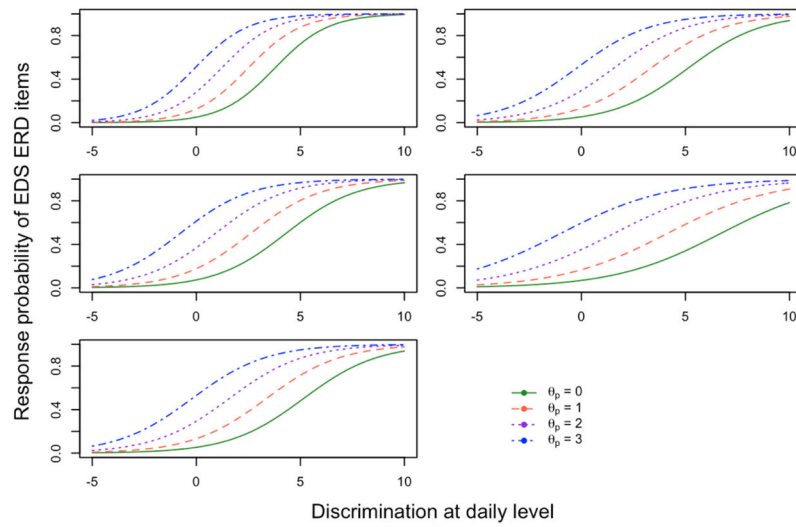
(A) Item characteristic curves conditional on participant-level RED
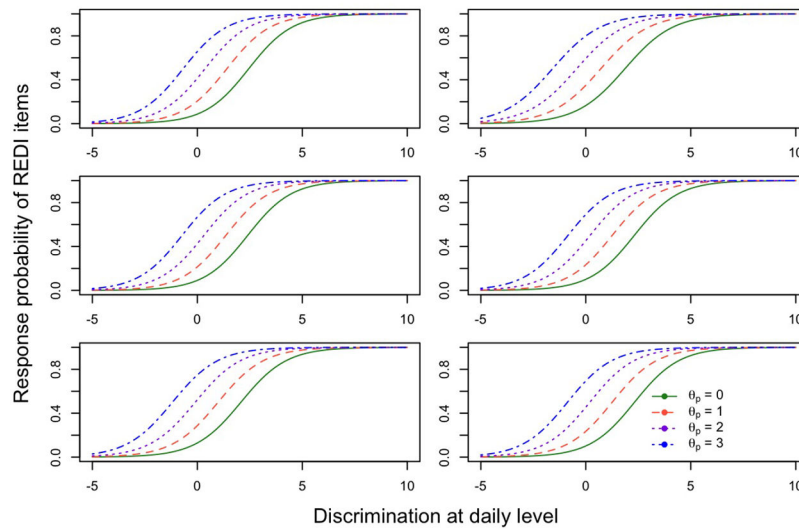
(B) Item characteristic curves conditional on daily-level RED

Figure 1.
Item characteristic curves of EDS-AER and REDI. In the top panel, the daily-level latent trait *z* score was set to 0. In the bottom panel, the participant-level latent trait *z* score was set to 0. Different lines represent different items.
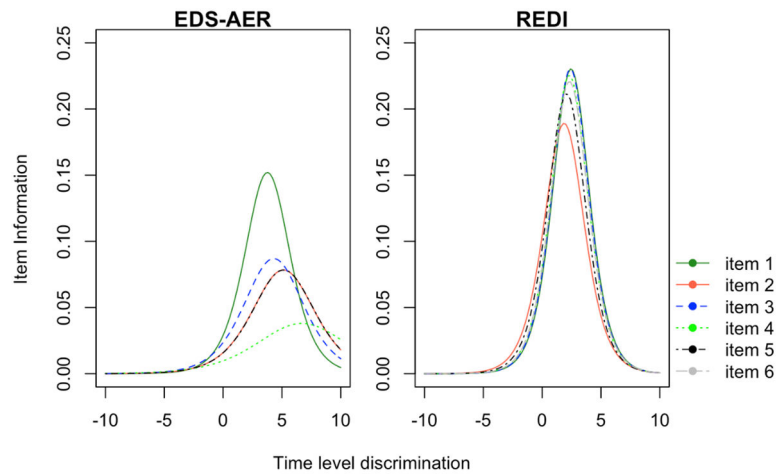
(A) Item Information Function
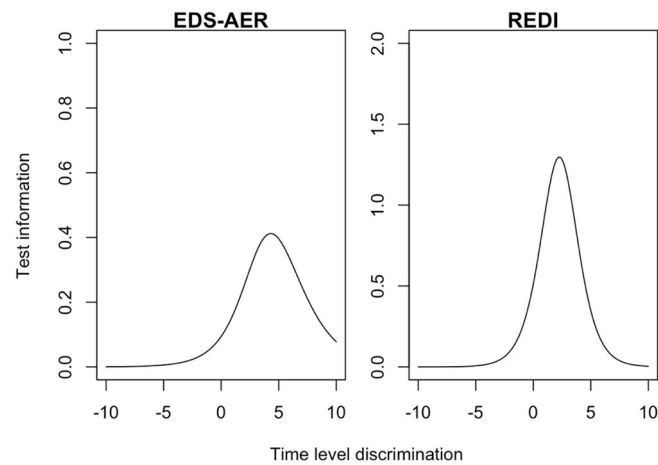
(B) Test Information Function

Figure 2.
Item information function and test information function of EDS-AER and REDI at the daily level. The participant-level latent traits were set to 0. In the item information function, different lines represent different items.

Table 1

Items of Everyday Discrimination Scale (EDS) and Racial Ethnic Discrimination Index (REDI)

Everyday Discrimination Scale (EDS)

Item	Discrimination Frequency (DF)		Attrition to Race/Ethnicity (AER)	
	Total %	Participant%	Total %	Participant%
1. Today, I was treated with less courtesy or respect than other people.	3.08	26.38	1.80	28.25
2. Today, I received poorer service than others in restaurants or stores.	1.26	9.86	1.06	16.56
3. People act as if I was not smart.	3.50	24.64	1.88	29.55
4. People act as if they were afraid of me.	1.74	14.20	1.10	17.21
5. Today, I was threatened or harassed.	1.16	9.86	.54	8.44

Racial/Ethnic Discrimination Index (REDI)

Item	Total %	Participant%
1. I was treated unfairly because of my race/ethnicity.	3.02	18.84
2. I felt stress because of my race/ethnicity.	4.93	24.06
3. Others treated me poorly because of my race/ethnicity.	3.04	17.39
4. I was teased because of my race/ethnicity.	3.21	17.68
5. I felt uncomfortable because of my race/ethnicity.	3.64	19.71
6. I felt unsafe because of my race/ethnicity.	2.80	16.23

Note. In EDS, when the participants responded "yes" to the DF item, they are asked whether race/ethnicity attributed (AER) to the discriminative experience: "how sure are you that this happened because of your race?". Participants responded the AER item using a four-point scale (1 = not at all, 2 = not very sure, 3 = somewhat sure, 4 = very sure). Total % indicates the percentage of reported the occurrence of the item. Participant % indicates the percentage of participants reported at least one occurrence of the item.

Table 2

Item Correlations of EDS and REDI at the Daily Level (Upper Diagonal) and Person Level (Lower Diagonal)

Item	1	2	3	4	5	6
EDS Discrimination Frequency (DF)						
1	1	.39	.37	.20	.40	
2	.80	1	.34	.37	.27	
3	.81	.62	1	.30	.40	
4	.79	.76	.84	1	.43	
5	.90	.63	.79	.72	1	
EDS Attrition to Race/Ethnicity (AER)						
1	1	.31	.37	.07	.12	
2	.90	1	.25	.10	.20	
3	.87	.93	1	.31	.40	
4	.93	.83	.89	1	.71	
5	.91	.82	.92	.86	1	
REDI						
1	1	.99	.99	.88	.92	.90
2	.90	1	.96	.91	.93	.94
3	.98	.89	1	.92	.96	.95
4	.99	.91	.99	1	.90	.93
5	.95	.94	.96	.97	1	.93
6	.98	.91	.98	.98	.96	1

Table 3

Model Fit of the Two-level 1-PL and 2-PL for the EDS and REDI

Model	Number of Parameters	Log-likelihood	AIC	BIC	Likelihood Ratio Test Comparison	χ^2	df	p value
					EDS Discrimination Frequency (EDS-DF)			
1-PL	7	−1866.15	3746.31	3790.00				
2-PL	15	−1851.09	3732.17	3825.81	1-PL	20.98	8	.007
2-PL(D)	11	−1855.90	3733.79	3802.46	2-PL	9.22	4	.056
2-PL(P)	11	−1852.08	3726.16	3794.83	2-PL	1.36	4	.851
					EDS Attribution to Ethnicity/Race (EDS-AER)			
1-PL	7	−1171.97	2357.93	2401.63				
2-PL	15	−1154.33	2338.66	2432.30	1-PL	28.71	8	< .001
2-PL(D)	11	−1158.49	2338.98	2407.65	2-PL	6.89	4	0.142
2-PL(P)	11	−1161.49	2344.98	2378.70	2-PL	12.00	4	0.017
					Racial/Ethnic Discrimination Index (REDI)			
1-PL	8	−1685.45	3386.90	3436.81				
2-PL	18	−1639.29	3314.58	3426.87	1-PL	38.20	10	< .001
2-PL(D)	13	−1672.27	3370.54	3451.65	2-PL	1467.91	5	< .001
2-PL(P)	13	−1650.83	3327.67	3408.77	2-PL	12.59	5	.028

Note. 2-PL(D) is a variant of the 2-PL model in which the daily-level item loadings are constrained to be equal across all items. 2-PL(P) is a variant of the 2-PL model in which the participant-level item loadings are constrained to be equal across all items. AIC is Akaike Information Criterion. BIC is Bayesian Information Criterion (BIC). χ^2 is the χ^2 difference. df is the difference of degrees of freedom.

Table 4

Parameter Estimates and Standard Errors (SE) of the 2-PL Model Constraining Equal Participant-level Item Loadings in EDS and REDI

Item	Intercept		Loading (Daily)		Loading (Participant)	
	Estimate	SE	Estimate	SE	Estimate	SE
EDS Discrimination Frequency (EDS-DF)						
Item 1	2.94	.29	.81	.05	1	---
Item 2	2.80	.25	.40	.14	1	---
Item 3	2.18	.16	.42	.08	1	---
Item 4	2.54	.18	.22	.13	1	---
Item 5	3.08	.32	.63	.10	1	---
EDS Attribution to Race/Ethnicity (EDS-AER)						
Item 1	3.22	.38	.84	.05	1	---
Item 2	2.71	.24	.37	.16	1	---
Item 3	2.44	.19	.46	.08	1	---
Item 4	2.66	.21	.24	.18	1	---
Item 5	3.07	.31	.38	.24	1	---
Racial/Ethnic Discrimination Index (REDI)						
Item 1	2.35	.23	.96	.01	1	---
Item 2	1.62	.12	.87	.04	1	---
Item 3	2.30	.24	.96	.01	1	---
Item 4	2.19	.21	.95	.01	1	---
Item 5	1.91	.19	.92	.03	1	---
Item 6	2.18	.22	.94	.02	1	---

Table 5

Measurement Invariance Test of EDS-AER and REDI across Ethnicity/Race and Gender

Invariance Model	Number of Parameters	Log-likelihood	AIC	BIC	Comparison	χ^2	df	p value
EDS Attribution to Ethnicity/Race (EDS-AER): Invariance across Ethnicity/Race								
2-PL(P)	35	−1509.16	3088.33	3306.81				
Loading	27	−1513.22	3080.44	3248.99	Baseline	8.33	8	0.40
Loading + Intercept	19	−1522.59	3083.18	3201.78	Loading	13.18	8	0.11
EDS Attribution to Ethnicity/Race (EDS-AER): Invariance across Gender								
2-PL(P)	23	−1367.54	2789.07	2932.65				
Loading	19	−1369.34	2776.67	2895.28	Baseline	6.97	4	0.14
Loading + Intercept	15	−1370.61	2771.23	2864.86	Loading	2.11	4	0.72
Racial/Ethnic Discrimination Index (REDI): Invariance across Ethnicity/Race								
2-PL(P)	41	−1958.85	3999.69	4255.47				
Loading	31	−1984.13	4030.25	4223.65	2-PL(P)	28.66	10	<.01
Loading + Intercept	21	−2004.89	4051.78	4182.78	Loading	27.99	10	<.01
Partial Loading	33	−1965.52	3997.03	4202.90	Baseline	6.95	8	.54
Partial Loading + Intercept	25	−1976.60	4003.20	4159.17	Partial Loading	72.54	8	<.01
Racial/Ethnic Discrimination Index (REDI): Invariance across Gender								
2-PL(P)	27	−1843.87	3741.74	3910.18				
Loading	22	−1851.16	3746.31	3883.56	2-PL(P)	7.65	5	.18
Loading + Intercept	17	−1861.26	3756.52	3862.58	Loading	9.44	5	.09

Note. 2-PL(P) is a variant of the 2-PL model in which the participant-level item loadings were constrained to be equal across all items. AIC is Akaike Information Criterion. BIC is Bayesian Information Criterion (BIC). χ^2 is the χ^2 difference. *df* is the difference of degrees of freedom. In the partial invariance model for REDI across ethnicity/race, the loading and intercept of item 2 were not constrained to be invariant.

Table 6

Differential Item Functioning Test of EDS-AER and REDI across Ethnicity/Race and Gender

| | Ethnicity/Race | | | | Gender | |
| | African American | | Asian American | | | |
Item	χ^2	p value	χ^2	p value	χ^2	p value
EDS Attribution to Ethnicity/Race (EDS-AER)						
Item 1	0.46	.50	7.34	.01	96.72	< .001
Item 2	0.25	.62	.01	.91	1.25	.26
Item 3	1.18	.28	4.20	.04	4.76	.03
Item 4	3.46	.06	1.49	.22	.78	.38
Item 5	.00*	1.00	13.75	< .001	.03	.87
Racial/Ethnic Discrimination Index (REDI)						
Item 1	1.21	1.00	1.02	1.00	.10	1.00
Item 2	4.79	.90	8.17	.61	.86	1.00
Item 3	.00*	1.00	.32	1.00	4.39	.93
Item 4	.00*	1.00	.00*	1.00	.04	1.00
Item 5	.01	1.00	.00*	1.00	8.63	.57
Item 6	13.53	.20	9.76	.46	.88	1.00

Note.

*
indicates that the χ^2 was slightly negative. We rounded them to .00 instead.