Reliability generalization of the multigroup ethnic identity measure- Revised (MEIM-R)

Timothy B. Smith
Brigham Young University, tbs@byu.edu

Hayley Marie Herrington
Brigham Young University - Provo

Erin Feinauer

Derek Griner
Brigham Young University

Original Publication Citation
Herrington, H. M., Smith, T. B., Feinauer, E., & Griner, D. (2016). Reliability generalization of the Multigroup Ethnic Identity Measure-Revised (MEIM-R). Journal of Counseling Psychology, 63(5), 586-593.http://dx.doi.org/10.1037/cou0000148

Herrington, H. M., Smith, T. B., Feinauer, E., & Griner, D. (2016). Reliability generalization of the Multigroup Ethnic Identity Measure-Revised (MEIM-R). *Journal of Counseling Psychology, 63*(5), 586-593.http://dx.doi.org/10.1037/cou0000148

Reliability Generalization of
the Multigroup Ethnic Identity Measure–Revised (MEIM–R)

Abstract

Individuals' strength of ethnic identity has been linked with multiple positive indicators, including academic achievement and overall psychological well-being. The measure researchers use most often to assess ethnic identity, the Multigroup Ethnic Identity Measure (MEIM), underwent substantial revision in 2007. To inform scholars investigating ethnic identity, a reliability generalization analysis was performed on data from the revised version (MEIM–R) and compared with data from the original MEIM. Random-effects weighted models evaluated internal consistency coefficients (Cronbach's alpha). Reliability coefficients for the MEIM–R averaged $\alpha = .88$ across 37 samples, a statistically significant increase over the average of $\alpha = .84$ for the MEIM across 75 studies. Reliability coefficients for the MEIM–R did not differ across study and participant characteristics such as sample gender and ethnic composition. However, consistently lower reliability coefficients averaging $\alpha = .81$ were found among participants with low levels of education, suggesting that greater attention to data reliability is warranted when evaluating the ethnic identity of individuals such as middle-school students. Future research will be needed to ascertain if data with other measures of aspects of personal identity (e.g., racial identity, gender identity) also differ as a function of participant level of education and associated cognitive/maturation processes.

Keywords: reliability, psychometrics, ethnic identity, personal identity

Reliability Generalization of

The Multigroup Ethnic Identity Measure–Revised (MEIM–R)

Individuals' self-perceptions and identity in relation to other people exert profound influence on many facets of behavior and psychological functioning (e.g., Erikson, 1968; Marcia, 1966; Tajfel & Turner, 1986). Ethnic identity, one's sense of self in ethnic terms (Phinney, 2000), is associated with positive behavioral, psychological, and academic outcomes (Phinney, 1990; Smith & Silva, 2011). A strong sense of ethnic identity correlates with self-esteem, self-efficacy, and psychological well-being, particularly among youth and people of color (Kiang, Gonzales-Backen, Yip, Witkow & Fuligni, 2006; Rivas-Drake et al., 2014; Schwartz, Zamboanga, & Jarvis, 2007; Umaña-Taylor & Updegraff, 2007). Similarly, ethnic identity is positively associated with academic adjustment, achievement, motivation, and aspirations among students of color (e.g., Altschul, Oyserman & Bybee, 2006; Fuligni, Witkow, & Garcia, 2005).

Over the past three decades social scientists and educational researchers have dedicated substantial attention to measuring and evaluating ethnic identity, as shown in thousands of research manuscripts representing millions of hours of scholarly effort (Smith & Trimble, 2016). Intensive scholarship, such as that reported in a special section of *The Journal of Counseling Psychology* (Ponterotto & Mallinckrodt, 2007), continues to expand and refine our understanding and measurement of this important form of self-perception and affiliation (Umaña-Taylor, 2015).

Reliable and valid measurement of ethnic identity, along with its associated conceptualization and operationalization, deserves careful consideration (Cokley, 2007; Helms, 2007; Trimble, 2007). Although many measures of ethnic identity have appeared in the literature (e.g., Umaña-Taylor, Yazedjian, & Bámaca-Gómez, 2004), the vast majority of research on the topic (e.g., Smith & Trimble, 2016) has been conducted using the Multigroup Ethnic Identity

Measure (MEIM; Phinney, 1992). The MEIM, designed for use across all ethnic groups, originally consisted of 14 items assessing attachment or belonging, achieved identity, and ethnic behaviors. Subsequent scholarship (Roberts et al., 1999) recommended removing the two items assessing ethnic behaviors, with additional studies providing psychometric evidence for the resulting two subscales of Exploration and Commitment (e.g., Spencer, Icard, Harachi, Catalano, & Oxford, 2000; Yancey, Aneshelsel, & Driscoll, 2003). Although many researchers using the MEIM adhered to the recommended version, some removed or modified various MEIM items to fit their own purposes. Given concerns in the field regarding the use and factor structure of the MEIM, substantial effort was invested in revising this instrument (Phinney & Ong, 2007).

The Multigroup Ethnic Identity Measure–Revised

The Multigroup Ethnic Identity Measure–Revised (MEIM–R) consists of six items, half the number on the MEIM. The MEIM–R represents a two-factor model of ethnic identity, exploration and commitment, with each subscale having three items. To develop the MEIM–R Phinney and Ong (2007) conducted analyses of multiple datasets, conceptual analyses, and focus groups to refine MEIM wording and item selection. A confirmatory factor analysis fitting the intended two-factor structure (AGFI = .96, CFI = .98, RMSEA = .04) provided evidence for scale validity, with a subsequent study by the scale's authors (Phinney & Ganeva, 2010) also supporting the two-factor structure. Although lower than the reliability coefficients typically reported for the MIEM (e.g., Roberts et al., 1999), the reliability coefficients reported by Phinney and Ong for the MEIM–R were nevertheless in the acceptable range: .81 for the total scale, .76 for the Exploration subscale, and .78 for the Commitment subscale.

The MEIM–R has now been available for many years, and evaluations of its psychometric properties by independent scholars have begun to appear in the literature. Yoon

(2011) found support for the two-factor structure of the MEIM–R with a sample of 289 counseling students. Chakawa, Butler, & Shapiro (2015) found similar support for the two-factor structure with a sample of 105 African American and 91 European American adults from Alabama. The two-factor model was also supported with large samples of East Asian adolescents in Canada (Homma, Zumbo, Saewyc, & Wong, 2014) and pregnant women in California (Brown et al., 2014). Nevertheless, scholars have frequently cautioned that the conceptualization of ethnic identity may differ across people of different backgrounds (e.g., Ong, Fuller-Rowell, & Phinney, 2010), and mean differences in strength of ethnic identity are commonly found across ethnic groups (e.g., Brown et al., 2014; Yoon, 2011). So any evaluation of the psychometric properties of the MEIM–R must consider differences across participant characteristics, particularly ethnic background.

Although two studies (Brown et al., 2014; Yoon, 2011) found measurement invariance across ethnic groups and another study (Chakawa et al., 2015) found differential item functioning by ethnicity on only one of six items, the issue of measurement invariance needs to be addressed across multiple studies with people from many backgrounds if the MEIM–R is to be appropriately advocated for use across all groups of people. For instance, strength of ethnic identity and its association with well-being differs across age (Smith & Trimble, 2016), ostensibly due to developmental processes (Quintana, 2007). Thus education and maturation should be considered when attempting to measure ethnic identity (Phinney & Ong, 2007).

Nevertheless, very limited scholarship has investigated differences across multiple participant and study characteristics, even less focusing on how those differences impact the psychometric properties of ethnic identity measures. Although a few studies of measurement validity exist for the MEIM–R, many studies involving the MEIM–R report reliability coefficients, data which could be aggregated and compared across such characteristics as participant age and ethnicity. Thus detailed analyses of measurement reliability could be conducted as an initial step toward understanding how individuals of different backgrounds respond to questions regarding ethnic identity, most often administered via the MEIM–R. With such information, scholars could more precisely measure populations, select procedures, and interpret research findings.

Measurement Reliability

Measurement reliability refers to the consistency of test scores (Crocker & Algina, 1986), with differences between the true score and the observed score being *measurement error*. Reliable test scores have minimal measurement error. Conversely, unreliable results are comparable to completely random results and can be said to measure nothing. Measurement reliability denotes the ability to measure something rather than nothing (Vacha-Haase & Thompson, 2011). Verifying reliability is an important step in ascertaining validity, which is the degree to which the test actually measures the intended construct. Because the MEIM–R, like the original MEIM, was intended for use across all populations (Phinney & Ong, 2007), it would be highly beneficial to evaluate score reliability across different samples.

An important but sometimes overlooked aspect of reliability is that it is a property of the observed data, not of the measure itself: Results can be reliable or unreliable, but the measure itself remains the same. The same measure can generate reliable scores in one instance and unreliable scores in another, depending on the setting and sample. Thus different study characteristics (e.g., design, format, environment) and participant characteristics (e.g., age, gender, ethnicity) can affect variations in reliability (Vacha-Haase, 1998). For this reason, reliability should be examined for each sample of data. As mentioned previously, for the

MEIM–R ascertaining differences across participants of different ages and ethnic groups is of particular concern (Brown et al., 2014; Phinney & Ong, 2007; see also Swartz et al., 2014).

The most widely used measure of reliability is internal consistency, the magnitude of inter-item correlations. For example, identical responses to MEIM–R items about "strong attachment towards" one's ethnic group and "strong sense of belonging" to one's ethnic group (wording from Phinney & Ong, 2007, p. 276) would indicate internal consistency, and the associated scores would be said to be reliable. Consistency of responses indicates that items can appropriately be combined for purposes of scoring. Thus all six MEIM–R items should correlate highly with each other, with the three items of the Exploration subscale and the three items of the Commitment subscale also showing high correlation.

Reliability Generalization

Acknowledging the variability of reliability coefficients across different datasets, researchers have explored ways to analyze variables that impact data reliability. Vacha-Haase (1998) introduced a method for systematically evaluating data reliability, termed *reliability generalization*. Reliability generalization builds on Schmidt and Hunter's (1977) *validity generalization*, a meta-analytic approach to analyzing validity coefficients, with the primary difference being the analysis of reliability coefficients in reliability generalization. The method of reliability generalization seeks to characterize "(a) the typical reliability of scores for a given test across studies, (b) the amount of variability in reliability coefficients for given measures, and (c) the sources of variability in reliability coefficients across studies" (Vacha-Haase, 1998, p. 6). This approach allows an evaluation of a broad sample of data (multiple studies) with a particular instrument to see under what conditions (populations, settings, etc.) the reliability coefficients differ.

Thus reliability generalization studies inform the work of scholars using an instrument in different settings and with different populations (Henson, 2001). For instance, a reliability generalization study of the Beck Depression Inventory found low data reliability when completed by individuals experiencing substance abuse (Yin & Fan, 2000). Accounting for such population differences is essential because data reliability impacts the results and the interpretation of the results. Specifically, systematic sources of unreliability reduce both statistical power and effect size magnitude (Hellman, Fuqua, & Worley, 2006). When a researcher does not account for or mitigate systematic sources of measurement error, interpretation of the observed effect size and its confidence interval can be problematic and even erroneous.

Because the MEIM–R was intended to measure ethnic identity across multiple conditions, including participant ethnicity, immigration status, and age, it would be valuable to ascertain whether or not such variables influence the size of reliability coefficients obtained with the MEIM–R across studies conducted under various conditions with diverse populations. For instance, ethnic groups differ in terms of strength of ethnic identity (Avery et al., 2007; Yap et al., 2014). However, if reliability coefficients differ by ethnicity or by strength of ethnic identity (with response consistency enhanced by repeated exposure and introspection among those with higher levels of ethnic identity), comparisons of means across groups need to account for differences in data reliability. Similarly, MEIM–R scores and their reliability may differ by immigrant status (due to language and acculturation issues), gender (due to culturally influenced gender roles or experiences of sexism), or age (due to multiple maturation processes, including introspection and autonomy). Previous reliability generalization studies with other instruments have found participant age to influence data reliability, with lower reliability coefficients for data provided by adolescents than by adults (Barnes, Harp, & Jung, 2002; Hellman et al., 2006; Nilsson, Schmidt, & Meek, 2002). Given that measures of ethnic identity are often administered to adolescents (Swartz et al., 2014; Smith & Trimble, 2016), it is particularly important to evaluate participant age, along with other study conditions such as the number of other measures administered in the study (with long questionnaires possibly resulting in participant fatigue/disinterest).

A reliability generalization analysis is advisable when information regarding reliability coefficients is available across dozens of studies but information regarding validity coefficients is limited making a validity generalization analysis unfeasible. Although some information about the validity of the MEIM–R exists (e.g., Chakawa et al., 2015), the participant samples have been delimited and data are insufficient to conduct a validity generalization study. However, many reliability coefficients are available from a broad range of studies conducted across multiple settings with different types of participants. Examining trends in reliability coefficients across multiple studies should enable scholars to make better decisions regarding research samples and procedures when planning to evaluate ethnic identity using the MEIM–R.

Our study, therefore, evaluated data from the MEIM–R using reliability generalization methods. We hypothesized that reliability coefficients of the data from the MEIM–R would be adequate (alpha coefficients larger than 0.80) and not significantly different from data from the MEIM, given the coefficients reported by the scale's developers (Phinney & Ong, 2007). We also hypothesized that the reliability coefficients would be similar across participant composition (age, gender, and race) and across study settings (schools, clinics, etc.).

Method

Literature Search and Study Selection Criteria

Articles that cited the original MEIM–R article by Phinney and Ong (2007) were found by using Google Scholar and EBSCO Publishing (PsycINFO and ERIC databases). Research not conducted in English in the United States or Canada was excluded due to differences that might occur across language translations or cultural interpretations of MEIM–R items. Research administering the MEIM–R involved studies with both cross-sectional and longitudinal designs, so to avoid the potential methodological confound of repeated measurement over varying and sometimes unspecified lengths of time, data extraction in longitudinal studies was limited to their first time administration. For sake of consistency, Cronbach's (1951) alpha coefficients were extracted as the measure of internal consistency reliability. Being the average of all possible split-half coefficients for a particular dataset, alpha is assumed to estimate the expected correlation between tests of the same construct (Nunnally, 1978). Cronbach's alpha has been reported more often in research literature than any other indicator of measurement reliability (Sijtsma, 2009), with over 25,000 citations.

Coding Procedures

Information reported in the obtained studies was coded to enable statistical analyses. The categorical variables coded were (a) publication status (published or unpublished), (b) other variables measured in the same administration as the MEIM–R, (c) immigrant status of study participants, and (d) clinical or adjudicated status of the participants. Coded variables with continuous level data were (a) mean age and years of education of the participants, (b) number of participants in the sample, (c) race/ethnicity of participants, (d) percentage of female participants, (e) number of other measures administered at the same time as the MEIM–R, and (f) Cronbach's alpha coefficient for the MEIM–R. When a study reported reliability coefficients separately for non-overlapping independent groups of participants, such as for three distinct ethnic groups, those separate reliability coefficients were extracted because they could be statistically compared: for example, across participant ethnicity.

Trained graduate and undergraduate students with prior experience coding meta-analyses independently coded each article. Overall, the inter-rater agreement for data abstraction was adequately high for categorical variables (with Cohen's *kappa* averaging .93) and for continuous variables (with intraclass correlations for single measures averaging .97). The few discrepancies between coders were resolved following additional review of the manuscript and, when needed, consultation with a research team member who had previously conducted multiple relevant research projects.

Data Analysis

In this reliability generalization study, alpha coefficients were analyzed using random-effects weighted models, anticipating that study and participant characteristics moderated the values reported in studies. To avoid possible errors associated with the distribution of coefficients, Fisher's *z* transformation was performed prior to analyses (Corey, Dunlap, & Burke, 1998), with transformation of the data back to the metric of a reliability coefficient following analyses for purposes of interpretation.

The main analysis ascertained the averaged reliability coefficient across all studies coded, although findings were expected to vary across studies. To determine sources of coefficient homogeneity, we evaluated possible moderation due to study and participant characteristics through two types of analyses. We calculated random effects weighted correlations between the continuous level data coded and the alpha coefficient, for instance as the average participant age might be associated with the reliability coefficients obtained within studies (perhaps with data obtained from younger participants being less reliable on average). We also calculated random effects weighted analyses of variance for the categorical level data coded, for instance as data from student populations might be less reliable than data from general community members. These analyses would thus inform researchers intending to use the MEIM–R about possible study and participant characteristics that they should consider when planning their research.

Data Comparison Between the MEIM–R and the MEIM

A separate analysis compared the reliability coefficients obtained for the MEIM–R to those obtained for the MEIM using data reported in 75 research manuscripts that had been included in a published meta-analysis (Smith & Silva, 2011) identifying MEIM data in 30 published and 45 unpublished studies on the association between ethnic identity and well-being (including mental health, self-esteem, and positive mood states). The 75 studies had a median of 133 participants (15,105 participants altogether) who were on average 21 years of age, with an average of 63% being female.

Results

Study Characteristics

A literature search found that 120 of 391 articles citing the MEIM–R (Phinney & Ong, 2007) contained quantitative data. Of those 120 articles, 35 reported Cronbach's alpha coefficients on the total MEIM–R, including data from 37 unique samples. Two studies each reported total MEIM–R coefficients for two separate samples of participants. Reliability coefficients for the three-item Exploration subscale were reported in 21 studies involving 24 distinct samples, and reliability coefficients for the three-item Commitment subscale were reported in 22 studies involving 26 distinct samples.

The studies included unpublished dissertations (60%) and peer-reviewed publications (40%). The median number of participants across all samples was 200 (mean = 287), with a minimum of 67 and maximum of 1,884; the total number of participants was 10,611. Participants had been obtained from normal community samples or from students enrolled in schools, middle-school through university. No clinical or adjudicated populations had been used in the studies. Although 19 did not report information regarding acculturation level and immigrant status, 17 explicitly included immigrant populations and 1 omitted immigrants. Across all studies, participants were 64% female, 26% White/European American, 25% Hispanic/Latino(a), 23% Asian American, 18% Black/African American, 0.5% Native American, and 0.06% Pacific Islander; about 8% were reported as “other” people of color or unspecified “ethnic minorities.” The average age of participants was 27.0 years (SD = 10.0), and the average education level was 14.0 years (SD = 3.2). The mean level of ethnic identity measured on a 1-5 scale was 3.65 (SD = .42).

Every study measured variables in addition to ethnic identity; 73% of studies measured more than two other variables, averaging four other variables. The most common variables measured were related to academic achievement and psychological well-being.

Reliability Generalization Analyses

The overall random effects weighted average reliability coefficient for the MEIM–R was .88, a respectable value for a six-item measure. As can be seen in Table 1, the averaged reliability coefficients for MEIM–R were statistically significantly higher than those of the original MEIM ($\alpha = .84$, $p < .05$) evaluated using coefficients obtained from 75 studies included in a meta-analytic review (Smith & Silva, 2011). Moreover, the six-item MEIM–R yielded more internally consistent data than the three-item Exploration and Commitment subscales of the MIEM–R, which had average random effects weighted coefficients of .84 and .86, respectively (see 95% confidence intervals in Table 1).

Because reliability coefficients were characterized by substantial heterogeneity across studies (Q = 211.6, $p < .0001$; $I^2 = 83.0$), subsequent analyses were conducted to ascertain the degree to which differences in reliability coefficients were attributable to participant and study characteristics (see Table 2). Of particular interest, no differences were found in average coefficients across samples with participants from different ethnic groups. Also the percentages of participants from each ethnic group were not significantly correlated with reliability coefficients within studies; thus the MEIM–R yielded equivalent reliability coefficients across participant ethnicity. Similarly, no differences were found in reliability coefficients when considering (a) how many other variables were measured in the studies (survey length), (b) the proportion of female participants in the sample, and (c) whether or not the participants included immigrants to North America.

However, reliability coefficients varied ($p < .05$) as a function of three participant variables: average age, education level, and student status. Specifically, higher averaged reliability coefficients were obtained when (a) the samples were adults (the older the participants, the more reliable the data); (b) the samples were college educated (the more years of education, the more reliable the data); and (c) the samples consisted of community members rather than currently enrolled students.

As interaction among the three variables was likely (participant age being related to both years of education and student status), it was essential to evaluate the relative contribution of each variable in a random effects weighted meta-regression model. As seen in Table 3, the model reached statistical significance ($p > .0001$) and explained 37.3% of the variance in reliability coefficients across 34 samples. The only one of the three variables to remain statistically significant in the model was participant education ($\beta = .57$, $p < .0001$). The univariate findings with the two variables of participant age and student status were better accounted for by participant overall level of education.

Overall, reliability coefficients differed systematically as a function of participant education. Higher reliability coefficients can be expected when the MEIM–R is administered to populations with higher levels of education. Specifically, reliability coefficients reported for samples with at least some college education averaged $\alpha = .90$, whereas samples with participants in grades 7 to 9 averaged $\alpha = .81$. Nevertheless, the average reliability coefficients obtained on the MEIM–R with populations having lower levels of education remained above the acceptable range. No reliability coefficient less than .73 was observed on the total MEIM–R across 37 distinct samples, although one study with middle-school students reported coefficients of about .50 for the 3-item Exploration and Commitment subscales.

Possible Reporting/Publication Bias

We had originally located 120 studies with MEIM–R data, but most of them had not contained information about reliability coefficients. If the studies lacking this information had systematically differed from those that did report reliability coefficients (e.g., low reliability coefficients being unreported), then the findings of our reliability generalization study would be biased. Thus we needed to estimate the degree to which author reporting bias might have occurred. A second possible source of bias involved the publication status of the research studies. Unpublished research is more difficult to locate, but must be considered because its findings may differ systematically from those from published data. We therefore evaluated the possibility that the omission of studies not reporting data or not located in our literature search influenced the findings reported previously.

First, we compared the reliability coefficients from published studies to those reported in unpublished studies located in our search (Table 2). Unpublished studies yielded slightly larger coefficients than those reported in published studies, the opposite trend of what would have occurred due to publication bias. A second method, "trim and fill" analysis (Duval & Tweedie, 2000), identified only five possibly missing values in the data distribution, meaning a minimal likelihood of either author-reporting bias or publication bias. Using that same method, the corrected average $\alpha = .875$ did not differ significantly from the observed value of .88. Thus the absence of the five hypothetically missing studies did not impact the overall findings. Finally, Orwin's fail-safe *N* was calculated to be 282 using a conservative estimate of $\alpha = .70$ for hypothetical "missing" studies. Given that the lowest reliability coefficient observed among located studies was .73, we considered it unlikely that 282 studies with coefficients of about .70 would exist but not have been located and/or not have reported the coefficients. Thus author-reporting bias and publication bias did not appear to be threats to the findings of this study.

Discussion

The purpose of this study was to assess the reliability of the MEIM–R, an already widely used instrument that will likely be used in hundreds of future studies. We had hypothesized that the MEIM–R would yield reliable data across different participant characteristics and across research settings. Random-effects weighted models with alpha coefficients were used to ascertain reliability generalization.

Summary of Findings

For data obtained with the six-item MEIM–R, the overall average reliability coefficient of .88 was respectable, with slightly lower but still satisfactory average coefficients observed for the two three-item MEIM–R subscales (Table 1). The design of the MEIM–R had been based on extensive analyses of MEIM data, with the intention of improving the psychometric properties of the original version (Phinney & Ong, 2007). The resulting MEIM-R item wording closely reflects the intended two-factor structure and is more homogeneous than the wording of the original MEIM items. Our findings confirmed that the data yielded by the MEIM–R have been characterized by significantly greater internal consistency reliability than those from the original MEIM. The MEIM–R improves on the original MEIM both in administration brevity and data reliability.

The MEIM–R was designed to measure ethnic identity across ethnicities, ages, settings, and so forth. As such contextual variables could influence individuals' responses to MEIM–R items, we considered it important to investigate the available data for any systematic differences in reliability. The analyses indicated that the MEIM–R yielded equivalently reliable data across several study and participant characteristics. Especially notable was that data from participants of different ethnic groups yielded about the same averaged reliability coefficients. Variables such as the inclusion of immigrant participants, number of measures administered simultaneously (survey length), and study publication status did not systematically influence the findings.

Reliability coefficients tended to differ as a function of participant age, level of education, and current student status. But due to the interaction of these variables (e.g., education level increases with age), we analyzed them simultaneously. In a meta-regression model, only participants' average education level remained statistically significant (Table 3). The higher the level of participants' education, the more reliable the data obtained in studies. In particular, MEIM–R data was least reliable when the participants had less than high school education. Individuals with less than high school education may be less likely to have deliberated their own ethnic identity and thus to respond to questions about ethnic identity in a consistent manner. Developmental processes are clearly relevant to ethnic identity (Quintana, 2007). Overall, given that reliability generalization studies with other psychological measures have shown that adolescents can be inconsistent in their responses (Barnes et al., 2002; Hellman et al., 2006; Nilsson, Schmidt, & Meek, 2002), scholars conducting research with young and/or less educated individuals should take steps to facilitate response consistency when possible, verify data reliability after survey administration, and interpret data accounting for confidence intervals and reliability coefficients. Future research will be needed to ascertain if data with other measures of aspects of personal identity (e.g., racial identity, gender identity) also differ as a function of participant level of education and associated cognitive/maturation processes.

Limitations and Recommendations

A reliability generalization analysis reflects the studies available, and in this instance 60% of the studies were unpublished dissertations and theses. Unpublished studies are often of lower methodological quality than published studies that have undergone peer review; the data from unpublished studies could be less reliable than those of published studies. However, the reliability coefficients obtained from published studies were similar to those of unpublished studies. Research participants completing the MEIM–R would have been exposed to the same questions irrespective of subsequent publication status, and their responses were consistent regardless of possible differences indirectly associated with subsequent publication.

A limitation of the available literature concerned the inconsistent reporting of reliability coefficients. Although 120 articles with MEIM–R data were identified, reliability coefficients were reported in only 37% of them. Thus our analyses could have overestimated the average reliability coefficients of the MEIM–R if unreported coefficients were systematically lower than reported ones. However, statistical examination of that possibility did not detect problematic patterns in the data, so we concluded that reporting bias was unlikely to have systematically impacted reported findings. However, the weakness of inadequate reporting of reliability coefficients in the majority of studies is a problem that should be corrected by dissertation and thesis committee members and peer reviewers.

Furthermore, the low percentage of studies reporting reliability coefficients restricted the statistical power of our analyses. Post hoc estimates of statistical power varied according to the number of studies in the analysis (and also due to the observed effect size), with all statistically significant values in Tables 2 and 3 having estimates above 0.90 but with other findings being characterized by much lower estimates of power. Had additional studies been available, the statistical analyses would have been much less susceptible to Type II error.

Another limitation of the literature concerns the selection and use of the MEIM–R. Even among recent reports, we observed that fewer studies have used the MEIM–R than the highly popular original MEIM. Evidence has verified the superiority of the revised version to the original, and we considered it past time for researchers to switch to the revised version, published in 2007. In addition, we found that some researchers administered only one of the two MEIM–R subscales, although the brevity of the MEIM–R, with only six items, would seem to make administration simple enough that researchers would include both subscales.

Another consideration relevant to the extant literature concerns sample size. We located only one study with a large number of participants (N = 1,884) in the available literature. Several of the studies had fewer than 100 participants, limited numbers for which results are likely to be impacted by sampling error. In general, a higher number of participants increases the likelihood that the results can be generalized to other populations.

A final limitation observed in the literature is that most of the participants were recruited using convenience samples of normal community members and students (which limitation has been widely observed, e.g., Schwartz et al., 2014). It would seem worthwhile to evaluate ethnic identity (and the psychometric properties of the MEIM–R) with different groups of people, such as clinical and at-risk populations, who are surprisingly underrepresented in multicultural psychology research (Dunn, Montoya, & Smith, 2006; Smith & Trimble, 2016). As it now stands, the reliability of the MEIM–R has been supported for normal community members only.

Evaluating the reliability generalization of the MEIM–R is a necessary but insufficient step for ongoing examinations of the scale's validity. Future studies of the psychometric properties of the MEIM–R should consider contextual factors, including participant and study characteristics (see Yap et al., 2014), with particular attention to the education level of participants, found to be a factor influential in how participants responded to the MEIM–R items.

We found the MEIM–R to be a reliable, practical measure of ethnic identity that will likely be widely used in future research. Nevertheless, scholars seeking to evaluate the popular concept of ethnic identity must still account for important conceptual issues not addressed in psychometric evaluations such as this study (e.g., Cokley, 2007; Smith & Trimble, 2016; Trimble, 2007).

References

Altschul, I., Oyserman, D., & Bybee, D. (2006). Racial-ethnic identity in mid-adolescence: Content and change as predictors of academic achievement. *Child Development*, *77*, 1155-1169.

Avery, D. R., Tonidandel, S., Thomas, K. M., Johnson, C. D., & Mack, D. A. (2007). Assessing the Multigroup Ethnic Identity Measure for measurement equivalence across racial and ethnic groups. *Educational and Psychological Measurement*, *67*(5), 877-888.

Barnes, L. L., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement*, *62*(4), 603-618.

Brown, S. D., Unger Hu, K. A., Mevi, A., Hedderson, M., Jun, S., Quesenberry, C. P., & Ferrara, A. (2014). The Multigroup Ethnic Identity Measure–Revised: Measurement invariance across racial and ethnic groups. *Journal of Counseling Psychology*, *61*(1), 154-161.

Chakawa, A., Butler, R. C., & Shapiro, S. K. (2015). Examining the psychometric validity of the Multigroup Ethnic Identity Measure–Revised (MEIM–R) in a community sample of African American and European American Adults. *Cultural Diversity & Ethnic Minority Psychology*,

Cokley, K (2007). Critical issues in the measurement of ethnic and racial identity: A referendum on the state of the field. *Journal of Counseling Psychology*, *54*, 224-234.

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging correlations: Expected values and bias in combined Pearson *r*s and Fisher's *z* transformations. *The Journal of General Psychology*, *125*(3), 245-261.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL:

Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 97-334.

Dunn, T. W., Smith, T. B., & Montoya, J. A. (2006). Multicultural competency instrumentation: A review and analysis of reliability generalization. *Journal of Counseling & Development*, *84*(4), 471-482.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455-463.

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

Fuligni, A. J., Witkow, M., & Garcia, C. (2005). Ethnic identity and the academic adjustment of adolescents from Mexican, Chinese, and European backgrounds. *Developmental Psychology*, *41*(5), 799-811.

Hellman, C. M., Fuqua, D. R., & Worley, J. (2006). A reliability generalization study on the Survey of Perceived Organizational Support the effects of mean age and number of items on score reliability. *Educational and Psychological Measurement*, *66*(4), 631-642.

Helms, J. E. (2007). Some better practices for measuring racial and ethnic identity constructs. *Journal of Counseling Psychology*, *54*(3), 235-246.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177-189.

Homma, Y., Zumbo, B. D., Saewyc, E. M., & Wong, S. T. (2014). Psychometric Evaluation of the Six-Item Version of the Multigroup Ethnic Identity Measure with East Asian Adolescents in Canada. *Identity*, *14*(1), 1-18.

Kiang, L., Yip, T., Gonzales, M., Witkow, M., & fuligni, A. (2006). Ethnic identity and daily psychological well-being of adolescents from Mexican and Chinese backgrounds. *Child Development*, *77*, 1338-1350.

Marcia, J. E. (1966). Development and validation of ego-identity status. *Journal of Personality and Social Psychology*, *3*(5), 551-558.

Nilsson, J. E., Schmidt, C. K., & Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-efficacy Scale. *Educational and Psychological Measurement*, *62*(4), 647-658.

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Ong, A. D., Fuller-Rowell, T. E., & Phinney, J. S. (2010). Measurement of ethnic identity: Recurrent and emergent issues. *Identity: An International Journal of Theory and Research*, *10*(1), 39-49.

Phinney, J. S. (1990). Ethnic identity in adolescents and adults: Review of research. *Psychological Bulletin*, *108*(3), 499-514.

Phinney, J. S. (1992). The multigroup ethnic identity measure: A new scale for use with diverse groups. *Journal of Adolescent Research*, *7*(2), 156-176.

Phinney, J. S. (2000). Ethnic and racial identity: Ethnic identity. In A. E. Kazdin (Ed.), *Encyclopedia of Psychology, Vol. 3* (pp. 254-259). Washington, DC: APA.

Phinney, J. S., & Ganeva, Z. (2010). The structure of ethnic identity of young adolescents of Bulgarian and Romany origin. *Bulgarian Journal of Psychology*, (1-4), 65-79.

Phinney, J. S., & Ong, A. (2007). Conceptualization and measurement of ethnic identity: Current status and future directions. *Journal of Counseling Psychology, 54,* 271–28.

Ponterotto, J. G., & Mallinckrodt, B. (2007). Introduction to the special section on racial and ethnic identity in counseling psychology: Conceptual and methodological challenges and proposed solutions. *Journal of Counseling Psychology*, *54*, 219-223.

Quintana, S. M. (2007). Racial and ethnic identity: Developmental perspectives and research. *Journal of Counseling Psychology*, *54*(3), 259-270.

Rivas-Drake, D., Seaton, E. K., Markstrom, C., Quintana, S., Syed, M., Lee, R. M., ... & Yip, T. (2014). Ethnic and racial identity in adolescence: Implications for psychosocial, academic, and health outcomes. *Child Development, 85*(1), 40-57.

Roberts, R. E., Phinney, J. S., Masse, L. C., Chen, Y. R., Roberts, C. R., & Romero, A. (1999). The structure of ethnic identity of young adolescents from diverse ethnocultural groups. *Journal of Early Adolescence, 19*, 301–322.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.

Schwartz, S. J., Zamboanga, B. L., & Jarvis, L. H. (2007). Ethnic identity and acculturation in Hispanic early adolescents: Mediated relationships to academic grades, prosocial behaviors, and externalizing symptoms. *Cultural Diversity and Ethnic Minority Psychology*, *13*(4), 364-373.

Smith, T. B., & Silva, L. (2011). Ethnic identity and personal well-being of people of color: A meta-analysis. *Journal of Counseling Psychology*, *58*(1), 42-60.

Smith, T. B., & Trimble, J. E. (2016). *Foundations of multicultural psychology: Research to inform effective practice.* Washington, DC: American Psychological Association.

Spencer, M., Icard, L., Harachi, T., Catalano, R., & Oxford, M. (2000). Ethnic identity among monoracial and multiracial early adolescents. *Journal of Early Adolescence*, *20*, 365-387.

Tajfel, H., & Turner, J. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. Austin (Ed.), *Psychology of intergroup relations* (pp. 7-24). Chicago: Nelson-Hall.

Trimble, J. E. (2007). Prolegomena for the connotation of construct use in the measurement of ethnic and racial identity. *Journal of Counseling Psychology, 54*(3), 247-258.

Umaña-Taylor, A. J. (2015). Ethnic identity research: How far have we come? In C. Santos and A. Umaña-Taylor (Eds.) *Studying ethnic identity: Methodological and conceptual approaches across disciplines.* (pp. 11-26). Washington, DC: APA.

Umaña-Taylor, A. J., & Updegraff, K. A. (2007). Latino adolescents' mental health: Exploring the interrelations among discrimination, ethnic identity, cultural orientation, self-esteem, and depressive symptoms. *Journal of Adolescence, 30*(4), 549-567.

Umaña-Taylor, A. J., Yazedjian, A., & Bámaca-Gómez, M. (2004). Developing the ethnic identity scale using Eriksonian and social identity perspectives. *Identity: An International Journal of Theory and Research*, *4*, 9-38.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*(1), 6-20.

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, *44*(3), 159-168.

Yancey, A., Aneshensel, C., & Driscoll, A. (2003). The assessment of ethnic identity in a diverse urban youth population. *Journal of Black Psychology*, *27*, 190-208.

Yap, S. C., Donnellan, M. B., Schwartz, S. J., Kim, S. Y., Castillo, L. G., Zamboanga, B. L., ... & Vazsonyi, A. T. (2014). Investigating the structure and measurement invariance of the Multigroup Ethnic Identity Measure in a multiethnic sample of college students. *Journal of Counseling Psychology*, *61*(3), 437-446.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*(2), 201-223.

Yoon, E. (2011). Measuring ethnic identity in the Ethnic Identity Scale and the Multigroup Ethnic Identity Measure-Revised. *Cultural Diversity and Ethnic Minority Psychology*, *17*(2), 144-155.

Table 1

Random Effects Weighted Internal Consistency Coefficients (Cronbach's α) of the MEIM–R and Original MEIM

Scale/Subscale	*α*	95% CI	*k*
MEIM–R	.88	.87, .89	37
Exploration	.84	.82, .87	24
Commitment	.86	.84, .88	26
MEIM[1]	.84	.83, .85	75

Note: [1] Sample of studies using items from the original MEIM. *k* = number of coefficients.

Table 2

Random Effects Weighted Moderation Analyses with Study and Participant Characteristics

Variable	*k*	*α*	*r*	*p*
Categorical Variables				
Participant Type				.007
Current Students	22	.87		
Community Members	15	.90		
Immigrants in the Sample				.84
Immigrants included in the data	17	.88		
No immigrants/not reported	20	.88		
Published manuscript				.08
Yes	16	.87		
No	21	.89		
Continuous Variables				
Avg. Age	37		.43	.002
Years of Education	34		.60	< .001
% Female	37		.22	.13
% African Americans	18		.35	.13
% Asian Americans	18		.01	.95
% Hispanic/Latino(a) Americans	17		-.10	.69
% White/European Americans	16		-.01	.97
% “Other” Americans	18		-.18	.42
Mean Level of Ethnic Identity	28		.18	.32
Number of Measures Administered	37		.18	.24
Total Number of Participants	37		.13	.42

Note. *k* = Number of samples included in the analysis.

Table 3

Meta-Regression of Participant Characteristics on Reliability Coefficients

Variable	*B*	*SE*	*p*	β	R^2	*k*
Model					.373	34
Education Level	0.030	.009	< .001	.57		
Avg. Age	-0.002	.004	.72	-.09		
Current Student	-0.066	.074	.37	-.20		