Race, Gender, and Teacher Equity Beliefs: Construct Validation of the Attributions of Mathematical Excellence Scale

Erik Jacobson[1], Dionne Cross Francis[2], Craig Willey[3], and Kerrie Wilkins-Yel[4]

[1]Indiana University, Bloomington

[2]University of North Carolina, Chapel Hill

[3]Indiana University-Purdue University, Indianapolis

[4]University of Massachusetts, Boston

Author Note

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Erik Jacobson, Dept. of Curriculum & Instruction, 201 N. Rose Ave., Bloomington, IN 47405. Email: erdajaco@indiana.edu

Abstract

Teachers' beliefs can have powerful consequences on instructional decisions and student learning. However, there is little research that focuses on how teachers' beliefs about the role of race and gender in mathematics teaching and learning influences educational equity within classrooms. This is partly due to the lack of studies focused on variation within classrooms, which in turn is hampered by the lack of instruments designed to measure mathematics-specific equity beliefs. In this study of 313 preservice and practicing elementary teachers, we report evidence of construct validity for the Attributions of Mathematical Excellence Scale (AMES). Factor analyses provide support for the four-factor structure including genetic, social, personal, and educational attributions. The findings suggest that the same system of attribution beliefs underlies both racial and gender prejudice among elementary mathematics teachers. The AMES has the potential to provide a useful outcome measure for equity-focused interventions both in teacher education and professional development.

Keywords. Teacher Beliefs, Mathematics, Racial Bias, Gender Bias, Attribution, Equity Beliefs

Teachers’ beliefs influence their instructional decisions (Pajares, 1992), including decisions that shape the mathematical learning opportunities of girls, Native American and Indigenous people, and Students of Color (SoC). However, the field lacks a validated, mathematics-specific instrument to measure teachers’ racialized and gendered beliefs about mathematics learning. To date, instrument development in this area has focused on learning more broadly. In the present study, we report our work developing the Attributions of Mathematical Excellence Scale (AMES), which we posit is an important step to advance research on the role of teachers’ equity beliefs in mathematics education. We consider evidence of substantive validity, structural validity, and external validity to build an initial argument for the construct validity (Flake, Pek, & Hehman, 2017) of the AMES as a measure of elementary teachers’ systems of racialized and gendered attribution beliefs about students’ struggles and success in mathematics. This instrument responds to calls for research on racial attitudes in mathematics education (Battey & Leyva, 2018) and comes at a time when public cries for racial justice and equity – and the valuing of Black, Latinx, and Asian American lives – has reached a new pinnacle. Such an instrument is critical for addressing the opportunity gap in school mathematics, a persistent challenge with wide implications for broadening participation in science, technology, engineering, and mathematics (STEM).

The validity evidence we report also has the potential to advance theory because the AMES design operationalizes several transformative theoretical claims about mathematics teachers’ equity beliefs. First, we claim that individuals hold attribution beliefs about mathematics learning that differ from their more general attribution beliefs. Second, we claim that a common belief system—attributions of mathematical excellence—underlies both racialized and gendered inequity in mathematics, although the effects are likely amplified for students with intersecting marginalized identities (Collins, 2000; Crenshaw, 1991; Leyva, 2017). Third, we claim that race-neutral and gender-neutral attribution beliefs are aligned with (rather than opposed to) racial stereotypes that make analogous attributions. For example, teachers who agree with ostensibly race-neutral statements (e.g., “Students who struggle to understand mathematics do not study enough.”) may be more likely than others to endorse analogous attributions of success to effort even if they echo racial stereotypes (e.g., “Black students struggle in mathematics because they are lazy,” is an analogous attribution which echoes the racist stereotype “Black people are lazy.”) Building on analyses of “color-blindness” in which individuals claim not to see race (e.g., Delgado & Stefancic, 2013) and related concepts of “color-evasiveness” (Annamma et al., 2017) and “race-evasiveness” (Chang-Bacon, 2021) in which individuals actively avoid discussing or acknowledging race, we conjecture that both attribution statements reflect the same underlying belief (i.e., the attribution of mathematical excellence to personal characteristics associated with race; c.f., *ideology,* see below). If this conjecture holds, equity in mathematics teaching and learning is likely shaped by teachers’ attribution beliefs.

Teachers’ Attribution Beliefs

Attribution beliefs are individuals’ thoughts about the causes of actions or behaviors. More broadly, attribution theory assumes that people try to determine why people do what they do by attributing behavior to causes (Fiske & Taylor, 1991; Graham, 2020). People make two kinds of attributions – internal or dispositional attributions, and external or situational attributions (Weiner, 1985). Internal attributions assign the cause of behavior to some internal characteristic of a person (e.g., personality, motives, race, gender). External attributions assign the cause of behavior to a situation or event outside of a person's control (e.g., social pressures, luck). Researchers have found that there is a tendency for individuals to explain their own behavior in ways that are favorable to them or their ingroup, referred to as attribution bias (Ross, 1977). The over-emphasizing of dispositional, or internal, explanations for people's behavior, even in cases with salient contextual factors, is referred to as correspondence bias or the over-attribution effect (Ross, 1977). In other words, people have a cognitive bias which assumes that what a person says or does is dependent on the "kind" of person they are instead of any situational or contextual factor. This preference for internal explanations appears to be particularly powerful in achievement domains like education (Reyna, 2008).

Existing studies suggest that teachers often fall prey to these forms of attributional biases (Bar-Tal & Guttmann, 1981; Bertrand & Marsh, 2015; Hall et al., 1989; Rolison & Medway, 1985). Teachers tend to attribute student failure to factors internal to the students (e.g., lack of effort), and external to themselves; however, students' successes are generally attributed to teachers (Guskey, 1982; Yehudah, 2002), for example, by virtue of instructional strategies (Gosling, 1994; Kulinna, 2007). How a teacher chooses to respond to a student's low achievement is determined by the teachers' attributions of the student's low performance (Reyna & Weiner, 2001) and teachers' attributions of students' success inform their expectations of student performance (Jussim et al., 2009). These expectations tend to be self-fulfilling (Rejeski & McCook, 1980; Reyna, 2000, 2008), because new information tends to be filtered through existing beliefs (Cooper & Burger, 1980; Fennema et al., 1990; Fives & Beuhl, 2012; Rejeski & McCook, 1980). Thus, once established teachers' attributions of students' achievement are unlikely to change without conscious awareness and deliberate effort.

Teachers' attributions are related to student variables, with gender and race/ethnicity being the most prominent (Espinoza et al., 2014). Although more recent research (Quinn, 2017) suggests that this is changing, teachers tend to perceive boys to be more mathematically capable than girls (Espinoza et al., 2014; Teidemann, 2002; Tindall & Hamil, 2004), thus attributing boys' success to ability and girls' to effort. Conversely, low performance among girls was attributed to a lack of ability and for boys to insufficient effort (Fennema et al., 1990; Espinoza et al., 2014). Researchers found that individuals' tendency to favor internal explanations is foregrounded in relation to race/ethnicity because cultural stereotypes serve as a fruitful source of attribution information (Reyna, 2008). Teachers make judgments about students' achievement and motivation based on race (Anderson-Clark et al., 2008). For example, White teachers provided more positive and less critical feedback to Black and Hispanic students than to White students (Harber et al., 2012). Although these studies do not explicitly capture teachers' attributions, they do show bias and differential actions towards students by virtue of their race.

We assume that most if not all K-8 mathematics teachers strive to provide all their students with the best instruction possible. However, research linking teacher expectations and student outcomes suggests that well-intentioned teachers are still influenced by biases that negatively impact Black and Brown students' mathematics success. Inspired by initial research on the role of colorblind racism beliefs in preservice teachers' emotion regulation during race-salient experiences (DeCuir-Gunby et al., 2020), we conjecture that teachers' attribution beliefs about the ultimate source of student differences in mathematics may explain how well they are able to follow through on their equity intentions.

Defining the Attributions of Mathematical Excellence Construct

The AMES measures teachers' beliefs about why students excel (or struggle) in mathematics. It includes four subscales describing attributions of mathematical excellence that are genetic (AME-G), social (AME-S), personal (AME-P), and educational (AME-E). From the perspective of social cognition research, these scales are designed to reflect specific *lay psychological theories* (e.g., Rangel & Keller, 2011) which help individuals make sense of why others act the way they do. *Psychological essentialism* (Medin, 1989) is the deterministic belief that individuals' behavior is explained by their underlying nature or essence. From the perspective of equity scholars within mathematics education research, these scales reflect specific *ideologies* (Battey & Leyva, 2016; Martin, 2012), which function to justify practices and policies in mathematics education. The AMES is designed to help researchers investigate teachers' attribution belief system about students' mathematical excellence by identifying the extent to which these beliefs are influenced by students' race and gender.

The genetic and the social AME scales (AME-G and AME-S, respectively) characterize mathematical excellence as a fixed trait. Our conceptualization of these two scales relies on a theoretical synthesis of research on social cognition and of scholarship on equity in mathematics education. Researchers have developed instruments to measure two different forms of psychological essentialism: *genetic determinism*, which is aligned with AME-G, and *social determinism*, which is aligned with AME-S. Genetic determinism refers to individuals' beliefs attributing personal characteristics (including academic ability and performance) to biology (Keller, 2005). As Jamieson and Radick (2017) posit, "Twenty-first century biology rejects genetic determinism, yet an exaggerated view of the power of genes in the making of body and minds remains [common]" (p. 1260). By contrast, belief in social determinism implies "that a person's essential features … are shaped permanently and profoundly by social factors (e.g., upbringing, socialization, and social background)" (Rangel & Keller, 2011, p. 1056). Unlike genetic determinism which focuses on an internal cause, belief in social determinism attributes the mathematical excellence of students to external social circumstances. Belief in social determinism to explain academic achievement might point to parents' level of educational attainment or parents' inability or unwillingness to help their children in school.

Although psychological essentialism explains attributions about general traits and behaviors, scholarship in mathematics education has identified ideologies that are specific to mathematical traits and behaviors. The *racial hierarchy in mathematics* (Martin, 2009) is an ideology that builds the belief that race is a genetic trait, and it informs AME- G. "[B]elief in innate mathematics ability serves as a colorblind way of unconsciously believing in the racial hierarchy of ability" (Battey & Leyva, 2016, p. 64). The ideology of *colorblindness* (Bonilla-Silva, 2003; Bonilla-Silva & Forman, 2000; Neville, et al., 2000) shifts discourse from internal genetic factors to external cultural and social proxies (e.g., parenting, values) and thereby makes discursive space for racist claims in putatively nonracial terms. Scholarship on this ideology informs AME-S. For both AME-G and AME-S, student struggle and success in mathematics is attributed to factors which are ultimately outside of teachers' influence, thus these attribution beliefs may undermine teachers' motivation to support SoC.

Whereas AME-G and AME-S attribute mathematical excellence to immutable causes, the personal and educational AMES subscales (AME-P and AME-E, respectively) reflect the view that mathematical excellence is malleable. These subscales differ in whether mathematical excellence is a result of internal or external factors. The AME-P scale is informed by a long history in mathematics education of teachers' attribution of girls' (but not boys') mathematics achievement to effort (e.g., Fennema et al., 1990; Tiedemann, 2000, 2002). It also builds on stereotypes that Black and Latinx students are lazy and do not try at school (Nasir & Shah, 2011; Oppland-Cordell, 2014). Some may be surprised by the apparent overlap between AME-P and growth mindset (e.g., Dweck, 1986; 2006; 2008), which more recent work indicates is an asset for students. However, we distinguish between students' views on the efficacy of their own effort and the way a teacher's focus on student effort can absolve teachers' duty of care by holding students wholly responsible for their learning. In this way, AME-P builds on the long-standing critique of the racist function of meritocracy in mathematics education (e.g., Battey & Franke, 2015; Martin, 2009). Meritocracy claims success is based on effort, implies that lack of effort explains lack of success, and therefore compounds—and provides justification to ignore—the historical, systemic, and institutional ways that opportunities and rewards are (and have been) distributed by race and gender instead of merit (Rubel, 2017).

The AME-E scale captures teachers' beliefs that mathematical excellence is a consequence of the schools, teachers, and educational opportunities a student has experienced. Our distinction between AME-E and the other AMES subscales draws on the contributions of Jackson et al. (2017) describing teachers' views of students' mathematical capabilities. Wilhelm and colleagues' (2017) take up this work and distinguish teachers' productive explanations ("ones that attribute student difficulty to instructional and/ or schooling opportunities", p. 349) from unproductive explanations ("ones that attribute student difficulty to inherent traits of the student, or their family or community," p. 349). The AME-E measures the productive beliefs that SoC and girls struggle in mathematics because of a lack of educational access and that mathematical excellence often involves extraordinary access to educational resources. These beliefs are productive in the sense that teachers recognize their own role as a teacher as a consequential aspect of mathematical excellence.

Validation Argument Overview, Research Questions, and Analytic Plan

In the current study, we investigated the validity of the AMES by drawing on a sample of both preservice and inservice teachers. Our work is guided by the construct validation framework discussed in Flake et al. (2017) and more recently applied to the validation of a novel instrument for mathematics teacher anxiety (Ganley et al., 2019). In this framework, evidence for substantive validity, structural validity, and external validity are integrated to make an argument for the construct validity of an instrument. We have framed our research questions both in relation to a specific kind of validity and in the context of open theoretical questions to clarify how our results advance knowledge even as they provide warrants for the use and further development of the AMES. This section also describes the statistical and psychometric analyses we used as warrants for the validation argument and to answer the research questions.

To establish initial evidence for substantive validity, we report our process of item design, item review and revision, and the innovation of using race/gender-neutral statements to measure race and gender bias. We do not report a research question for substantive validity because this aspect of our work is not an empirical study in the traditional sense. Instead, we draw on our synthesis of research in social cognition and equity in mathematics education (see above) to operationalize the four distinct factors within the AMES: genetic, social, personal, and educational attribution beliefs.

To assess structural validity at the item level, we used the survey response data and considered item-level descriptive statistics and relationships between items. Building on our experience piloting two previous versions of the AMES which suffered from skew and threshold effects, we applied two strategies to elicit a wider range of teachers' beliefs: writing negatively worded items and writing items using identity neutral language. Research Question 1 (RQ1) guided this phase of our study: *How are ratings of negatively worded (reverse coded) items related to teachers' ratings of corresponding positively worded items? To what extent do negatively worded items increase the range of rating responses across the AMES items?* The main hypothesis was that the negatively worded items would be correlated with corresponding positively worded items but have lower means, thereby increasing the range of AMES ratings. In addition, we asked Research Question 2 (RQ2): *How are ratings of identity neutral items related to teachers' ratings of corresponding identity specific items? To what extent do identity neutral items increase the range of rating responses across the AMES items?* The main hypothesis we investigated was that identity neutral items would be positively correlated with corresponding identity specific items but have lower means. To answer these questions, we compared the response patterns between negatively worded and positively worded items as well as the overall and within-factor item-total correlations.

To assess structural validity at the factor level, we used the survey response data and considered item-total correlations for each hypothesized factor, reliability estimates, and the results of confirmatory factor analysis (CFA). We found item dependencies between the identity specific and identity neutral versions of items that precluded modeling them as independent items with uncorrelated errors, a standard assumption of latent trait modeling. Ultimately, we combined the identity specific and identity neutral versions of items into testlets to account for between-item dependencies. Research Question 3 (RQ3) guided our work: *Are the AMES testlets better modeled as a single trait, as two factors (race versus gender prejudice), or as four factors corresponding to distinct attribution beliefs?* The main hypothesis for structural validity was that the hypothesized four-factor structure fit the data better than plausible alternatives. We report testlet statistics as well as the CFA results with the testlet data to answer this question.

To assess external validity, we investigated how scores on the AMES were correlated with other psychological constructs. In this study we were interested in whether social desirability played a role in teachers' responses, because if responses on the AMES were biased by social desirability, then AMES scores would have less utility for teacher education or research. We also examined social determinism and genetic determinism because these constructs heavily informed the development of the AMES. Research Question 4 (RQ4) guided this part of the study: *To what extent are scores on the AMES factors uncorrelated with social desirability and correlated with belief in social and genetic determinism in the ways theory predicts?* We hypothesized was that socially desirable responding would have a nonsignificant correlation with the four factors of the AMES, that social determinism would be correlated most with the AME-Social factor and least with the AME-Education and AME-Personal factors, and that genetic determinism would be correlated most with the AME-Genetic factor and least corelated with the AME-Education and AME-Personal factors. To answer this question, we used a CFA model with covariates to examine these correlations.

Method

Participants

The 313 participants included both practicing teachers (*n* = 223) and preservice teachers (*n* = 90) from [State BLINDED] in June 2020. All participants had previously participated in survey research for a larger project studying teachers' knowledge and beliefs (Author, 2019), and had indicated they were interested in follow-up research. The initial pool of teachers was a state-wide representative sample of public school teachers in Grades 2–5 which was stratified based on school urbanicity, school percentage of SoC, and percentage of students eligible for free-and-reduced price lunch. The sample of preservice teachers was constructed in two stages. First, volunteers were recruited from elementary teacher education programs in [State BLINDED], then participants were sampled from among volunteers in proportion to the size of each program.

Participants in the sample overwhelmingly identified as White (96%) and female (89%), following the regional demographics of elementary teachers in public schools (89% of elementary teachers in the Midwest identified as White; 88% as female; National Center for Educational Statistics, 2021). All but one teacher identified as having English as their first language. Further details about the background and characteristics of the participants are available in the supplemental materials.

Instruments

The social desirability scale (SDS-17; $\alpha = 0.75$; Stober, 2001) is an updated instrument for measuring desirable responding designed to update and replace the Marlowe-Crowne Scale (Crowne & Marlowe, 1960) as a reliable and valid measure of social desirability for adults. The belief in social determinism (BSD) and belief in genetic determinism (BGD) scales measure two components of psychological essentialism, the tendency of individuals to explain others' characteristics and behaviors by way of their underlying essence (Keller, 2005; Rangel & Keller, 2011). Both instruments have high reliability ($\alpha_{\text{BSD}} = 0.84$; $\alpha_{\text{BGD}} = 0.87$) and are each supported by several validation studies. More information about the instruments is available in the Instrumentation section of the Supplemental Materials.

Procedure

Participants were invited by email to take an online survey administered via Qualtrics, two follow up email reminders were sent, and data collection concluded after a two-week period that began with the first invitation email. The AMES items were included in a longer survey that also included questions about demographic and background characteristics as well as questions about mathematics teaching which are not germane to the present study. The whole survey took approximately 45 minutes to complete, and teachers were given a gift card as an incentive to participate. All participants provided informed consent before beginning the survey; all research instruments and procedures were approved by the Institutional Review Board at BLINDED before the study was conducted.

Results

Establishing Substantive Validity Through Operationalization

The AMES items are the product of three cycles of collaborative and cross-disciplinary item-writing based on a literature review coupled with field testing and item revision. We wrote and administered four pilot items tapping AME-G and AME-S in 2017 and administered them to 78 preservice teachers. Based on this pilot, we developed eight entirely new items tapping the same constructs and field tested them with 245 PSTs in 2018. The 64 AMES items used in the present study built on what we learned from these field tests and expand the instrument to include the AME-P and AME-E constructs, to include items with negative wording, and to include identity neutral items in addition to the identity specific items which were developed in previous cycles of item writing.

The AMES item design draws on the US General Social Survey (GSS) items that have been used for decades on nationally representative surveys of the US population (Quinn, 2017). The AMES items differ in two ways. The GSS questions require a yes or no response. Following recent work in sociology (Quinn, 2020; Valant & Newark, 2016), AMES items allow a range of responses which increases sensitivity to a broader range of beliefs. Second, the AMES items are mathematics specific.

The AMES items used in this study asked respondents to rate the truth (from *1: Completely true* to *7: Not at all true)* of statements that attribute stereotype-aligned indicators of mathematical excellence (e.g., "< White students / Boys > score higher on standardized math tests") to one of the four sources stereotypically associated with race or gender: genetic (e.g., "because of basic genetic differences", "... biological factors"), social (e.g., "because of cultural and religious expectations", "... upbringing"), personal (e.g., "because they put in more effort," "... spend more time studying"), and educational (e.g., "because they go to better schools," "... have more educational opportunities"). A subset of items for each indicator (1 genetic, 2 social, 4 personal, and 2 educational) were revised to contradict the relevant stereotype (i.e., negative wording). An example negatively worded genetic statement is, "In my view, genes do not determine which students excel in mathematics." An example negatively worded personal statement is, "The students who excel in mathematics rarely have to try very hard." These items were designed to be reverse scored.

For each identity-specific item, we wrote an identity-neutral version without specific race or gender identifiers (e.g., "Students struggle to learn ..." vs "Black students struggle to learn mathematics because they do not put in the required time and hard work.") This design encodes the theoretical claim that statements which make attributions of mathematical excellence that do not specify race or gender (i.e., identity neutral items) are different in degree but not in kind from statements that echo racial and gender stereotypes (in the case of AME-G, AME-S, and AME-P) or that acknowledge a racial and gendered opportunity gap (AME-E). The items are provided in Tables 1-4. We used item labels in which the first character indicates the construct (g: AME-G, s: AME-S, etc.), the numeral indicates a distinct attribution and mathematical excellence descriptor, and the second character indicates whether race ("r"), gender ("g"), or identity neutral wording ("n") is used; item labels appended with "x" are negatively worded. Thus, items g8nx and g8r share a genetic attribution and mathematical excellence descriptor, but g8nx is identity neutral and negatively worded ("In my view, genes do not determine which students excel in mathematics.”) whereas g8r is race specific (“In my view, genetic factors explain why Hispanic and Latino students struggle to learn mathematics.”).

Structural Validity

The new set of 64 AMES items were designed to increase the range of responses on AMES items because the first two field tests revealed skewed item distribution and possible restriction in range. We found that negatively worded items increased the range of responses to AMES items but were not highly correlated with the positively worded items. Because of this and other evidence that these items did not tap the intended constructs (see supplemental materials), we removed the negatively worded items from the subsequent stages of analysis. The identity neutral items also increased the range of responses to AMES items and were correlated with the identity specific items. However, the identity neutral and identity specific versions of items did not satisfy the assumption of local independence (see supplemental materials). To address this psychometric issue, we adopted a testlet approach (Wainer & Lewis, 1990; Wainer & Kiely, 1987) and scored each pair of items together as a single indicator. We used confirmatory factor analysis to evaluate how well the testlet scores could be modeled under the hypothesized factor structure.

In this section, we report on the structural validity evidence for items and then for testlets. We report both the results that address specific research questions about items (RQ1 and RQ2) and testlets (RQ3) as well as the results that contribute to the validity argument for AMES more generally in each category.

Items

We began our investigation of structural validity by evaluating the descriptive statistics for the 64 AMES items (Tables 1 - 4). Items that have skew exceeding an absolute value of 2

(Tabachnick & Fidell, 2013) or kurtosis exceeding an absolute value of 7 (Hair, et al., 2010) are considered problematic because they may violate the normality assumptions of CFA. Only two items (g3g; e8gx) exceeded the skew threshold, and only one item (e8gx) exceeded the kurtosis threshold. We flagged these items and next considered the item means.

The item means varied systematically by attribution. The range of item means and the grand mean tended to be lower for social and educational attributions and higher for items with genetic and personal attributions (see Table 5). To address RQ1, we compared the range of item means and grand means of the negatively worded items with the corresponding statistics for identity specific items. The negatively worded (reverse-scored) items had a substantially smaller grand mean and a lower range of item means, confirming our hypothesis that teachers would rate negatively worded attribution statements as more true than analogous positively worded items. The results in Table 5 also address RQ2 which concerns the distribution of responses to identity specific versus identity neutral items. We found that identity neutral items had smaller grand mean and a lower range of means than the identity specific items.

Next, we examined item-total correlations results. We calculated these statistics in three ways, corresponding to the hypothesized four-factor structure distinguishing genetic, social, educational, and personal attribution beliefs and two alternatives: a unidimensional structure and a three-factor structure comprising identity neutral-items, gender-specific items, and race-specific items. These three sets of item-total correlations are presented in the last three columns of Tables 1 to 4. Except for one item (g8n), the reverse-scored negatively worded items had low or negative item-total correlations, suggesting that these items were not effective at tapping the same constructs as the other items, regardless of the factor structure used (RQ1). As a result of the findings from item analysis, we excluded the negatively worded items from subsequent analyses (see Item Analysis in the supplemental materials).

Testlets

The AMES testlets were composed from pairs of dichotomized positively worded items that shared the same or very similar wording for the indicator of mathematical excellence (e.g., “high achievement scores in mathematics”) and the source of the attribution (e.g., “genetic factors”). One item in each testlet was identity neutral and the other was either race- or gender-specific. One challenge was scoring the testlets in a way that preserved the item meaning to maintain interpretability. For example, it did not make sense to add or average ratings on the two items, because these operations require an interval interpretation of item scores, but rating items are ordinal. Instead, we dichotomized the rating items at a meaningful cut point and used a simple rule to score each testlet on a 3-point ordinal scale: 1 if both items were rated 5 or below (partially to “completely true”), 2 if either item was rated partially or completely true, and 3 if neither was rated partially or completely true.

We chose a cut point of 5 to dichotomize items based on our examination of the empirical distribution of the identity specific item ratings. Many of these items evidenced a bimodal distribution with local minima near 5 (see Figure 1). Very few participants rated these statements completely true, but many rated these statements 5 or less, meaning *partially true*. Another group of participants tended to rate these statements as 6 or 7 meaning *not at all true* or nearly so. The cut score of 5 enabled us to distinguish between these groups and maintain meaningful scoring for the testlets. Significantly, we found that individuals who rated the identity neutral item true or partially true had higher odds to rate the identity specific item true or partially true (see RQ2). These results are presented in the fifth and sixth columns of Table 6. More analyses supporting the testlet scoring and factor structure are provided in the Testlet Analysis section of the Supplemental Materials.

Factor Analysis

We used CFA in Mplus 8.4 (Muthén, & Muthén, 2017) to examine the factor structure of the AMES testlets by treating them as categorical items with three levels and using weighted least squares estimation. We considered three different models: a unidimensional model with all items loading on the same single factor, a two-factor model with race-specific items loading on one factor and gender-specific items loading on the second factor, and a four-factor model with factors corresponding to the four kinds of attributions (genetic, social, educational, and personal). We follow Kline (2016) and report model chi-square, root mean square error of approximation (RMSEA), confirmatory fit index (CFI), Tucker-Lewis index (TLI), and the standardized root mean square residual (SRMR). The guidelines indicate there is good model fit when there is a nonsignificant chi-square test of model fit ($p > .05$), a high CFI (≥ .90), a high TLI (≥ .90), low RMSEA (< .08), and low SRMR (< 0.08). Table 7 presents these model fit statistics for Models 1-6.

The unidimensional model (Model 1 in Table 7 & 8; also see Figure 3) and two-factor model (Model 2 in Table 7 & 8; also see Figure 4) evidenced poor fit, with all fit indices falling below (or above) the recommended thresholds. The correlation between the race and gender factors in the second model was high, $r = 0.865$, $p = .000$, suggesting that these putative factors were not empirically distinct. The four-factor attribute structure exhibits much better fit, with every index indicating good fit except $\chi^2(1, N = 344) = 816.27, p = .000$) and SRMR = 0.086 > 0.08. In the four-factor model, the lowest standardized factor loadings were 0.42, 0.51, and 0.56 and all factor loadings were statistically significant at $p < .001$ (see Model 3, Table 7 & 8; also see Figure 5). The correlations among the factors were moderately high: between genetic and social, $r = .60$; between educational and personal, $r = .68$; between genetic and personal, $r = .62$; between genetic and educational, $r = .38$; between social and personal, $r = .71$; and between social and educational, $r = .70$. All correlations were statistically significant at $p < .001$. These suggested that the underlying latent constructs were clearly differentiated yet also strongly related, with the exception of the moderately low correlation between genetic and educational attribution beliefs.

To create a more parsimonious scale, we considered empirical item misfit and removed four items. More details are provided in the section titled "Removing Problematic Testlets" in the Supplemental Materials. The four removed items did not change the estimates of reliability for each factor appreciably, with Cronbach's alpha improving slightly for AME-Personal and worsening slightly for the other factors (AME-Genetic, α = .89 vs. .90; AME-Social, α = .76 vs. .79; AME-Educational, α = .80 vs. .82; AME-Personal, α = .82 vs. .80). We fit analogous versions of Model 1-3 with the 24 retained items, and found similar results (Models 4-6, see Table 7 & 8). The 24-item unidimensional model (Model 4) and two-factor model (Model 5) did not fit the data whereas the four-factor model (Model 6) fit the data very well, with all fit indices indicating good fit —including SRMR, which was slightly above the cutoff for the 28-item model. These results suggest that the 24-item version of the scale performs as well as, if not better, than the 28-item version, and we used this version of the model in subsequent analyses to address external validity.

External Validity

All analyses examining the relationship between scores on the AMES factors and the other instruments were conducted with Mplus, using a CFA model with covariates (Model 7, see

Figure 6) that extended Model 6, the parsimonious four-factor model. Raw scores for each external scale (SDS17, BGD, BSD) were included in the model, and we report the correlations between the latent constructs for each of the AMES factors and these raw scores.

AMES Factors and the Social Desirability Scale. To determine whether social desirability bias played a substantial role in teachers' responses on the AMES, we examined the correlation between the Social Desirability Scale (SDS-17) and each AMES factor. We found that the correlations between the SDS-17 and three factors were not statistically significant at the 0.05 level ($r_{Genetic}$ = -0.020, p = 0.73; $r_{Educational}$ = 0.121, p = 0.06; $r_{Personal}$ = 0.036, p = 0.560). We did find that SDS-17 and AME-Social were significantly correlated at the 0.05 level (r_{Social} = 0.150, p = 0.02). This very low correlation suggests that socially desirable responding explains about 2-3% of the variance in the AME-Social factor. Interestingly, when we examined item correlations with the SDS-17 scale, all correlations were between -.1 and .1, indicating the relationship may not be due to a single poorly performing item, but is more widespread across multiple items. This finding confirms that the relationship is weak, but also implies that further research is warranted to understand how social desirability might influence responses on the AME-Social items.

Relations among AMES factors and Belief in Genetic and Social Determinism. We were interested in examining how well the pattern of correlations among the AMES factors and Belief in Genetic Determinism (BGD) and Belief in Social Determinism (BSD) accorded with theoretical expectations. Because the AME-Genetic and AME-Social factors were designed to reflect mathematics, and learning-specific versions of these more general beliefs, we expected these AMES factors to be more highly correlated with BGD and BSD, respectively, than were other AMES factors. Because BGD and BSD are both deterministic beliefs, we expected the

AMES factors that frame mathematics excellence as a malleable trait to have the lowest correlations with BGD and BSD among the AMES factors. We found that BGD had the highest correlation with the AME-Genetic factor ($r = 0.524$, $p = 0.000$), then AME-Social ($r = 0.333$, $p = 0.000$), AME-Personal ($r = 0.245$, $p = 0.000$), and a non-significant correlation with AME-Educational ($r = 0.062$, $p = 0.311$). Similarly, BSD had the highest correlation with the AME-Social factor ($r = 0.392$, $p = 0.000$), then AME-Genetic ($r = 0.296$, $p = 0.000$), AME-Personal ($r = 0.265$, $p = 0.000$), and the lowest correlation with AME-Educational ($r = 0.183$, $p = 0.005$). Our results provide a pattern of correlations with magnitudes ordered in line with expectations based the social cognitive theory of psychological essentialism, which adds credibility to our interpretation of the AMES factors.

Discussion

Attribution beliefs help teachers make sense of their students' struggle and success in mathematics, but these attributions also shape how—and to whom—teachers respond. As is typical of beliefs in general, attribution beliefs about specific individuals tend to be stable (Green, 1971; Nespor, 1987), and this provides a window to both the problem and promise they pose for equity in mathematics education. On the one hand, theory suggests that once a teacher has attributed a student's mathematical success or struggle to a specific cause, there is little the student can do to change the attribution because it becomes a self-reinforcing filter (Wang & Hall, 2018). On the other hand, if teachers—through professional development, for example—become conscious of their attribution belief system, they may be able to reflect on the attributions they make and use this awareness to improve their teaching practice. As a first step in investigating these potential mechanisms of attribution beliefs for educational equity, it is necessary to have a well-validated measure of the construct for use with preservice and inservice teachers. In this study, our goal was to report initial validity evidence for a novel instrument, the Attributions of Mathematical Excellence Scale (AMES).

We addressed the substantive validity of the AMES instrument with a literature review to identify the construct and an iterative process to write and refine a wide range of items. The empirical factor analysis results evidenced structural validity by conforming to the theoretical structure we used to design the AMES: there were four moderately correlated factors related to genetic, social, educational, and personal attribution beliefs. Based on item analysis, we dropped the negatively worded items to preserve scale coherence, we combined the identity neutral and identity specific versions of the remaining items into testlets to satisfy local independence, and we ultimately identified a 24-testlet scale for future use after removing redundant items. We evaluated external validity by correlating AMES scores with a measure of social desirability which indicated that three of the factors are uncorrelated with social desirability and the social factor is weakly correlated with it. All four AMES factors were related in the hypothesized ways with belief in social and genetic determinism.

Substantive Validity

The development process for the AMES was designed with several characteristics to enhance the validity of the resulting instrument. First, to anchor the items in classroom practice and increase the potential utility of the resulting instrument, we began with interview-based descriptions of teachers' productive and unproductive beliefs about student struggle in mathematics (Jackson et al., 2017, Wilhelm et al., 2017). We expanded this construct to include a focus on mathematical excellence as well as struggle. Then, we drew on our novel synthesis of the research literature to identify four attributions of mathematical excellence that cross the internal versus external and the malleable versus non-malleable sources of attribution. At all stages, AMES items were developed through an iterative process of writing and revision that leveraged the varied expertise of our interdisciplinary team.

Structural Validity

The work we report on the structural validity of AMES advanced knowledge by generating a new hypothesis and supporting two theoretical claims. In response to skewed item ratings in initial pilots, we undertook two strategies to increase the range of responses and more completely capture the constructs. We answered Research Question 1 by evaluating the use of negative wording. This strategy was not successful and these items were ultimately eliminated from the instrument because of negative or low item-total correlations. These findings suggest that attribution beliefs form a loosely related system such that those who disagree with one attribution may agree not with its opposite but instead with an entirely different attribution.

We answered Research Question 2 by evaluating identity neutral items. First, the identity neutral items were successful in increasing the assessed range of the AMES constructs because the grand means of identity neutral items were lower than that of identity specific items for all four kinds of attribution statements. Second, item total correlation evidence and CFA results reveal that the identity neutral items loaded on the same constructs as the identity specific items. This finding provides robust evidence aligned with prior theoretical claims about the foundational meaning of teachers' "colorblind" or race-evasive statements (e.g., Battey & Leyva, 2016), and thereby makes an important empirical contribution to the field. Specifically, our survey data and psychometric methods suggest that stereotype-aligned, race-neutral attribution statements do not reflect different beliefs than race- and gender-based stereotypes, just milder (and more socially acceptable) versions of the same underlying belief.

The AMES was designed to include four distinct kinds of attribution beliefs including genetic, social, personal, and educational attributions. We answered Research Question 3 by evaluating this structure. We compared the hypothesized four-factor structure with two other plausible alternatives, a unidimensional model and a two-factor model distinguishing race-specific and gender-specific items. Item-total correlations and factor analysis fit indices strongly supported the four-factor model. The shortened instrument with the four-factor attribution structure also fit the data very well and provides additional evidence supporting the structural validity of the AMES. These results—and the comparison between the two- and four-factor models in particular—support another theoretical claim: attribution beliefs may be a common source of both gender and racial bias in mathematics education, something that has been largely overlooked because even among the rare studies that attend to both race and gender bias, researchers tend to frame each category of bias independently.

External Validity

We answered Research Question 4 by examining the relationship between AMES scores and several related constructs. We found that social desirability was not correlated with three of the AMES factors and only weakly correlated with the social factor. The findings of no (or low) correlations are remarkable in that many of the items reiterate racial and gender stereotypes and a substantial portion of the teachers participating endorsed them to some degree. We found that the pattern of correlations between AMES factors and belief in genetic determinism (BGD) and belief in social determinism (BSD) was consistent with social cognitive theory which defines these constructs as closely related yet distinct components of psychological essentialism. These findings increase our confidence that the AMES is measuring what it purports to measure.

Implications for Practice

The moderate size of the correlations also show that the AMES is measuring constructs which—although related to BGD and BSD—are clearly distinct, and this finding is in line with our theoretical claim that mathematics specific attribution beliefs are distinct from more general ones. Students perceive math as a difficult subject (Haag & Goetz, 2012) and are more anxious about it than other subjects (Pekrun et al, 2007); these differences may allow distinct attribution beliefs for mathematics to form. Teachers reported stronger beliefs in the role of innate ability for math than for German language arts (Heyder et al., 2020). Similarly, certain academic fields including mathematics are perceived by scholars in those fields to require more innate ability than others (e.g., Leslie, Cimpian,Meyer, & Freeland, 2015).

These observations lead to the question of what else beyond genetic and social determinism could contribute to mathematical attribution beliefs? Work in mathematics education (e.g., Battey & Leyva, 2016; Martin, 2009; 2012) suggests that there are discourses specific to mathematics that teachers maintain. The discursive practices within schools, districts, and teacher communities may play a larger role in shaping and maintaining attribution beliefs than does the variation between teachers' more general beliefs in social or genetic determinism, and such discourses might be influenced by professional development. Thus, the contexts in which teachers work likely reinforce these beliefs as deficit discourses are reaffirmed through teachers' interactions with colleagues and administrators (Horn, 2007).

Professional development focused on shifting teachers' attribution beliefs about mathematical excellence should include opportunities for teachers to understand how different forms of mathematical instruction support students in demonstrating different levels of competence (Jackson et al., 2017). Professional development and teacher education should explicitly acknowledge prevailing negative master narratives about SoC, and then support teachers in finding and retelling counter stories of mathematical competence (e.g., Stinson, 2008). Work should also explore how to adapt and adopt techniques used for disrupting other kinds of unproductive teacher beliefs. For example, Gill et al. (2020) found that preparing students to notice conflicts between their own beliefs and refutational texts produced more conceptual change in teachers' beliefs about mathematics teaching and learning than refutational texts alone.

Limitations and Future Directions

In this study, we reported preliminary evidence for the use of the AMES to measure mathematics attribution beliefs among preservice and inservice elementary teachers, but there are limitations in our work to date that suggest important directions for future research. First, the AMES items only reflect a small portion of the possible attribution statements that could be used in such items. Future work should include open-ended interviews with preservice and inservice teachers to contribute further evidence of substantive validity by illustrating how teachers reason about student struggle and success in mathematics and whether all of these ways of reasoning are adequately represented by the AMES items.

Second, further evidence should be collected to support the interpretation of AMES scores as a reflection teachers' racial and gender-related biases. For example, correlations between AMES scores and other instruments that measure teachers' racial or gender prejudice, including measures of implicit bias, would provide evidence about how well AMES taps race or gender bias. Third, the fit and reliability estimates for AMES should be confirmed with an independent sample. Fourth, the findings we report do not speak in any way to the level of attribution beliefs which are consequential for students. Future research that is sensitive to within classroom opportunity gaps either through test data or classroom observation would go a long way towards establishing how attribution beliefs are associated with educational equity.

Finally, this study was conducted with a large sample of preservice and inservice elementary teachers, but there are limitations of the sample. To the extent to which some groups of teachers were under- or overrepresented in the achieved sample, this might have introduced systematic bias in the implied distribution of mathematics attribution beliefs among elementary teachers. Available data suggests that the study sample is similar demographically to teachers in the same state, but the results we report in this study should be tested further with a nationally representative sample.

Conclusions

In this study, we presented the AMES and preliminary evidence of its construct validity by discussing evidence of substantive validity, structural validity, and external validity for its use with preservice and inservice elementary teachers. Our analyses showed the hypothesized structure fit the data much better than two alternatives, teasing apart four theoretical components related to distinct attribution beliefs. Scores on the instrument were correlated in expected ways with two constructs that informed the design of the measure, bolstering the theoretical grounding of the instrument design. By contrast, AME-G, AME-P, and AME-E scores were not correlated—and AME-S only weakly correlated—with social desirability, suggesting the design has avoided a major potential threat to validity.

Together, these findings support the use of AMES as a pre-assessment to inform the design of professional development that accounts for elementary teachers' beliefs about who excels in mathematics and why; to provide institutional feedback by tracking changes in these beliefs over time (for example in a teacher education program with a focus on equitable instruction); as a pre- and post-test for equity focused interventions that aim to shift teachers' attribution beliefs; and as a research tool to enable commensurate analysis of the relationships between teachers' beliefs, knowledge, and practice in a wide variety of educational contexts that exist in US schools. The AMES instrument may have wider applicability such as in other countries or with other populations of teachers (e.g., high school teachers), but we caution potential users to pilot the instrument before such use.

These research findings come at a time when the recent public reckoning with racial injustice in policing and the justice system has increased awareness of race- and gender-based inequities in other social and institutional systems of American society including education. We believe these conditions will continue to lead to increased interest in research efforts to understand inequity in education as well as new educational interventions to address it. Measurement is the cornerstone of all science, and in these endeavors, trustworthy instruments are of critical importance. Valid and reliable instruments for studying attribution beliefs at scale are necessary both to better understand inequity in classroom instruction and to understand how interventions influence—or are moderated by—teachers' attribution beliefs. We hope that researchers will use the AMES both to understand the role of attribution beliefs in classroom instruction and to evaluate interventions designed to address racial and gender equity in mathematics teaching and learning.

Acknowledgements

This work was funded in part by the National Science Foundation under Award #1561453 and Award #2200990. Any opinions, findings, and conclusions or recommendations expressed in this mate- rial are those of the authors and do not necessarily reflect the views of the NSF.

References

Anderson-Clark, T. N., Green, R. J., & Henley, T. B. (2008). The relationship between first names and teacher expectations for achievement motivation. *Journal of Language and Social Psychology*, *27*(1), 94–99.

Annamma, S. A., Jackson, D. D., & Morrison, D. (2017). Conceptualizing color-evasiveness: Using dis/ability critical race theory to expand a color-blind racial ideology in education and society. *Race Ethnicity and Education, 20*(2), 147-162.

Author, 2009, 2015;

Bar-Tal, D., and Guttmann, J. (1981). A comparison of teachers', pupils' and parents' attributions regarding pupils' academic achievements. *British. Journal of. Educational Psychology* 51, 301–311. doi: 10.1111/j.2044-8279.1981.tb02488.x

Battey, D., Bartell, T., Webel, C., & Lowry, A. (2021). Understanding the impact of racial attitudes on preservice teachers' perceptions of children's mathematical thinking. *Journal for Research in Mathematics Education, 52*(1), 62-93.

Battey, D., & Leyva, L. (2018). Making the implicit explicit: Building a case for implicit racial attitudes to inform mathematics education research. In T. Bartell (Ed.), Toward equity and social justice in mathematics education (pp. 21–41). New York: Springer.

Battey, D., & Leyva, L. A. (2016). A framework for understanding whiteness in mathematics education. *Journal of Urban Mathematics Education*, 9(2).

Bertrand, M., and Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. *American Educational Research Journal,* 52, 861–893.doi: 10.3102/0002831215599251

Bonilla-Silva, E., & Forman, T. A. (2000). “I am not a racist but...”: Mapping white college students’ racial ideology in the USA. *Discourse & Society*, 11(1), 50–85.

Bonilla-Silva, E. (2003). *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States.* Lanham, MD: Rowman and Littlefield.

Chang-Bacon, C. K. (2021). “We Sort of Dance Around the Race Thing”: Race-Evasiveness in Teacher Education. *Journal of Teacher Education*, doi: 00224871211023042.

Collins, P. H. (2000). Gender, black feminism, and black political economy. *The Annals of the American Academy of Political and Social Science*, 568(1), 41-53.

Cooper, H. M., and Burger, J. M. (1980). How teachers explain students’ academic performance: A categorization of free response academic attributions. *American Educational Research Journal*, 17, 95–109. doi: 10.3102/00028312017001095

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6), 1241-1279.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4), 349.

DeCuir-Gunby, J. T., Allen, E. M., & Boone, J. K. (2020). Examining pre-service teachers’ color-blind racial ideology, emotion regulation, and inflexibility with stigmatizing thoughts about race. *Contemporary Educational Psychology*, 60, 101836

Delgado, R., & Stefancic, J. (2013). *Critical race theory: The cutting edge*. Temple University Press.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*, 1087–1101.

Dweck, C. (1986). Motivational processes affecting learning. *American Psychologist*, 41, 1040–1048.

Dweck, C. S. (2006). Is math a gift? Beliefs that put females at risk. In S. J. Ceci & W. Williams (Eds.), *Why Aren't more women in science? Top researchers debate the evidence*. Washington, DC: American Psychological Association.

Dweck, C. S. (2008). *Mindsets and math/science achievement*. New York, NY: Carnegie Corp. of New York, Institute for Advanced Study, Commission on Mathematics and Science Education.

Espinoza, P., da Luz Fontes, A. B. A., & Arms-Chavez, C. J. (2014). Attributional Gender Bias: Teachers' ability and effort explanations for students' math performance. *Social Psychology of Education*, 17(1), 105-126.

Fennema, E., Peterson, P. L., Carpenter, T. P., & Lubinski, C. A. (1990). Teachers' attributions and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics*, 21(1), 55–69. doi:10.1007/BF00311015

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York: McGraw-Hill

Fives, H., & Buehl, M. M. (2012). Spring cleaning for the "messy" construct of teachers'beliefs: What are they? Which have been examined? What can they tell us? In K. R. Harris, S. Graham, T. Urdan, S. Graham, J. M. Royer, & M. Zeidner (Eds.), *APA Handbooks in Psychology. APA Educational Psychology Handbook, Vol. 2. Individual differences and cultural and contextual factors* (p. 471–499). American Psychological Association. https://doi.org/10.1037/13274-019

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370-378.

Ganley, C. M., Schoen, R. C., LaVenia, M., & Tazaz, A. M. (2019). The construct validation of the math anxiety scale for teachers. *Aera Open, 5*(1), doi: 10.1177/2332858419839702

Georgiou, S. N. (2008). Beliefs of experienced and novice teachers about achievement. *Educational Psychology, 28*, 119–131. doi: 10.1080/01443410701468716

Gill, M. G., Trevors, G., Greene, J. A., & Algina, J. (2020). Don't take it personally? The role of personal relevance in conceptual change. *The Journal of Experimental Education, 1*-22.

Gosling, P. (1994). The attribution of success and failure: the subject/object contrast. *European Journal of Psychology of Education*. 9, 69–83. doi: 10.1007/BF03172886

Graham, S. (2020). An attributional theory of motivation. *Contemporary Educational Psychology*. 61, https://doi.org/10.1016/j.cedpsych.2020.101861

Green, T. (1971), *The activities of teaching*, McGraw-Hill, New York, NY.

Guskey, T. R. (1982). Differences in teachers' perceptions of personal control of positive versus negative student learning outcomes. *Contemporary Educational Psychology,* 7, 70–80. doi: 10.1016/0361-476X(82)90

Hair, J. F., Anderson, R. E., Babin, B. J., & Black, W. C. (2010). Multivariate data analysis: A global perspective (Vol. 7). Upper Saddle River, NJ: Pearson.

Hall, B. W., Villeme, M. G., and Burley, W. (1989). Teachers' attributions for students' academic success and failure and the relationship to teaching level and teacher feedback practices. *Contemporary Educational Psychology*, 14, 133–144. doi: 10.1016/0361-476X(89)90031-3

Haag, L., & Goetz, T. (2012). Mathe ist schwierig und Deutsch aktuell. Vergleichende Studie zur Charakterisierung von Schulfächern aus Schülersicht [Math is difficult and German up to date: A study on the characterization of subject domains from students‘ perspective]. *Psychologie in Erziehung und Unterricht, 59*, 32–46.

Harber, K. D., Gorman, J. L., Gengaro, F. P., Butisingh, S., Tsang, W., & Ouellette, R. (2012). Students' race and teachers' social support affect the positive feedback bias in public schools. *Journal of Educational Psychology*, 104(4), 1149.

Heyder, A., Weidinger, A. F., Cimpian, A., & Steinmayr, R. (2020). Teachers' belief that math requires innate ability predicts lower intrinsic motivation among low-achieving students. *Learning and Instruction, 65*, 101-220. https://doi.org/10.1016/j.learninstruc.2019.101220

Horn, I.S. (2007). Fast kids, slow kids, lazy kids: Framing the mismatch problem in math teachers' conversations, J*ournal of the Learning Sciences*, 16(1), 37-79.

Hoy, A. W., Hoy, W. K., & Davis, H. A. (2009). Teachers' self-efficacy beliefs. In K. R. Wenzel & A. Wigfield (Eds.), *Educational psychology handbook series. Handbook of motivation at school* (p. 627–653). Routledge/Taylor & Francis Group.

Jackson, K., Gibbons, L., & Sharpe, C. (2017). Teachers' views of students' mathematical capabilities: Challenges and possibilities for ambitious reform. *Teachers College Record, 119*(7), 1-43.

Jameison, A., & Radick, D. (2017). Genetic determinism in the genetics curriculum. *Science & Education*, 26 (10), 1261–1290.

Jones, E. E. & Nisbett, R.E. (1971). *The actor and the observer divergent perceptions of the causes of behavior*. In E.E. Jones, D. E. Kanouse, H.H. Kelley, R.E. Nisbett, S. Valins & B. Weiner (Eds.) *Attribution; Perceiving the Causes of Behavior*, Morristown, General Learning Press, 79 – 94.

Jussim, L., Robustelli, S. L., and Cain, T. R. (2009). "Teacher expectations and self- fulfilling prophecies," in K. R. Wentzel and A. Wigfield (Eds.) *Handbook of Motivation at School*, New York, NY: Routledge, 349–380.

Keller, J. (2005). In genes we trust: The biological component of psychological essentialism and its relationship to mechanisms of motivated social cognition. *Journal of Personality and Social Psychology*, 88(4), 686–702. doi:10.1037/0022-3514.88.4.686

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press. doi:10.108 0/10705511.2012.687667

Kulinna, P. H. (2007). Teachers' attributions and strategies for student misbehavior. *Journal of Classroom Interaction,* 42, 21–30.

Leslie, S. -J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265. https://doi.org/10.1126/science.1261375.

Leyva, L. A. (2017). Unpacking the male superiority myth and masculinization of mathematics at the intersections: A review of research on gender in mathematics education. *Journal for Research in Mathematics Education*, 48(4), 397-433.

Martin. D.B. (2012). Learning mathematics while Black. *The Journal of Educational Foundations, 26*(1-2), 47-66.

Martin, D. B. (2009). Researching race in mathematics education. *Teachers College Record*, 111(2), 295–338

Matteucci, M. C., and Gosling, P. (2004). Italian and French teachers faced with pupil's academic failure: the "norm of effort. *European Journal of Psychology of Education.* 19, 147–166. doi: 10.1007/BF03173229

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44*(12), 1469.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus. Statistical analysis with latent variables. User's guide, 8th Edition*.

Nasir, N. S., & Shah, N. (2011). On defense: African American males making sense of racialized narratives in mathematics education. *Journal of African American Males in Education, 2*(1), 24–45.

National Center for Educational Statistics (2021). *Percentage distribution of public school teachers by sex, race/ethnicity, school level, and main teaching assignment: 2017–18.* Retrieved from https://nces.ed.gov/surveys/ntps/tables/ntps1718_21011202_t1n.asp

Nespor, J. (1987). The role of beliefs in the practice of teaching. *Journal of Curriculum Studies*, 19, 317-328.

Neville, H. A., Lilly, R. L., Duran, G., Lee, R. M., & Browne, L. (2000). Construction and initial validation of the color-blind racial attitudes scale (CoBRAS). *Journal of Counseling Psychology, 47*(1), 59.

Oppland-Cordell, S. B. (2014). Urban Latina/o undergraduate students' negotiations of identities and participation in an Emerging Scholars Calculus I workshop. *Journal of Urban Mathematics Education, 7*(1), 19–54.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, 62(3), 307. doi:10.2307/1170741

Pekrun, R., Frenzel, A.C., Goetz, T., & Perry, R. P. (2007). The Control-Value Theory of Achievement Emotions: An Integrative Approach to Emotions in Education In *P.A. Schutz & R. Pekrun* (Eds.), *Emotions in Education* (pp. 13-37). London: Academic Press

Pirrone, C. (2012). The influence of teachers' preexisting notions about students on scholastic achievement. *AASA Journal of Scholarship and Practice,* 9, 18–28.

Quinn, D. M. (2017). Racial Attitudes of PreK–12 and Postsecondary Educators: Descriptive Evidence From Nationally Representative Data. *Educational Researcher*, 0013189X17727270.

Quinn, D. M. (2020). Experimental Effects of "Achievement Gap" News Reporting on Viewers' Racial Stereotypes, Inequality Explanations, and Inequality Prioritization. *Educational Researcher*, 49(7), 482–492. https://doi.org/10.3102/0013189X20932469

Rangel, U., & Keller, J. (2011). Essentialism goes social: Belief in social determinism as a component of psychological essentialism. *Journal of Personality and Social Psychology*, 100(6), 1056–1078. doi:10.1037/a0022401

Rejeski, W. J., and McCook, W. (1980). Individual differences in professional teachers' attributions for children's performance outcomes. *Psychological Reports.* 46,1159–1163.

Reyna, C. (2000). Lazy, dumb, or industrious: when stereotypes convey attribution information in the classroom. *Educational Psychology Review*. 12, 85–110. doi:10.1023/A:1009037101170

Reyna, C. (2008). Ian is intelligent but Leshaun is lazy: antecedents and consequences of attributional stereotypes in the classroom. *European Journal of Psychology in Education*. 23, 439–458. doi: 10.1007/BF03172752

Reyna, C., and Weiner, B. (2001). Justice and utility in the classroom: an attributional analysis of the goals of teachers' punishment and intervention strategies. *Journal of Educational Psychology*. 93, 309–319. doi: 10.1037/0022-0663.93.2.309

Rolison, M. A., and Medway, F. J. (1985). Teachers' expectations and attributions for student achievement: effects of label performance pattern, and special education intervention. *American Educational Research Journal,* 22, 561–573. doi: 10.3102/00028312022004561

Ross, Lee 1977 "The intuitive psychologist and his shortcomings: Distortions in the attribution process.? In Leonard Berkowitz (ed.), *Advances in Experimental Social Psychology*. Volume J0. New York: Academic Press.

Rubel, L. H. (2017). Equity-directed instructional practices: Beyond the dominant perspective. *Journal of Urban Mathematics Education, 10*(2) 66-105.

Rubie-Davies, C.M., Flint, A., & McDonald, L. (2012). Teacher beliefs, teacher characteristics and school contextual factors: What are the relationships? *The British Journal of Educational Psychology, 82 Pt 2*, 270-288.

Solomon, D., Battistich, V., & Hom, A. (1996). Teacher beliefs and practices in schools serving communities that differ in socioeconomic level. *The Journal of Experimental Education, 64*(4), 327-347. http://www.jstor.org/stable/20152497

Stinson, D. W. (2008). Negotiating sociocultural discourses: The counter-storytelling of academically (and mathematically) successful African American male students. *American Educational Research Journal, 45*(4), 975-1010.

Stöber, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment*, 17(3), 222.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics (6th ed.).* Upper Saddle River, NJ: Pearson.

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology*, 99(2), 253-273. http://dx.doi.org/10.1037/0022-0663.99.2.253

Tiedemann, J. (2000). Gender-related belief of teachers in elementary school mathematics. *Educational Studies in Mathematics*, 41(2), 191–207.

Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school. *Journal of Educational Psychology*, 92(1), 144.

Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in mathematics*, 50(1), 49-62.

Tindall, T., & Hamil, B. (2004). Gender disparity in science education: The causes, consequences, and solutions. *Education, 125*, 282–295.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and teacher education, 17*(7), 783-805.

Valant, J., & Newark, D. A. (2016). The politics of achievement gaps: US public opinion on race-based and wealth-based differences in test scores. *Educational Researcher*, 45(6), 331–346.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1-14.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement, 24*(3), 185-201.

Wang, H., & Hall, N. C. (2018). A systematic review of teachers' causal attributions: Prevalence, correlates, and consequences. *Frontiers in Psychology*, 9, Article 2305. https://doi.org/10.3389/fpsyg.2018.02305

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, 92, 548–573.

Wilhelm, A. G., Munter, C., & Jackson, K. (2017). Examining relations between teachers' explanations of sources of students' difficulty in mathematics and students' opportunities to learn. *The Elementary School Journal, 117*(3), 345-370.

Yehudah, Y. B. (2002). Self-serving attributions in teachers' explanations of students' performance in a national oral essay competition. *Social. Behavior and. Personality: An international journal,* 30 (4), 411–416. doi: 10.2224/sbp.2002.30.4.411

Table 1

Item wording and descriptive statistics for the AME-Genetic subscale.

							Item-total Correlation		
		< 6					1-	3-	4-
Item	Label*	(%)	M	SD	skew	kurtosis	factor	factor	factor
Students' natural ability with mathematical reasoning is the primary factor that determines who studies advanced mathematics.	g1n	0.77	4.17	1.54	0.1	-0.95	0.16	0.20	0.37
Girls have less natural ability with mathematical reasoning than boys, so it makes sense that they are less likely to study advanced mathematics.	g1g	0.19	6.24	1.12	-1.71	2.88	0.26	0.25	0.47
I think that basic genetic differences determine to a large degree who becomes a professional mathematician.	g2n	0.52	5.24	1.48	-0.47	-0.71	0.41	0.44	0.66
I think that basic genetic differences explain why there are far more male than female mathematicians.	g2g	0.43	5.45	1.42	-0.6	-0.75	0.30	0.15	0.61
I believe that students who are less successful at pursuing mathematics-related career paths often lack genetic potential.	g3n	0.38	5.68	1.31	-0.82	-0.01	0.48	0.50	0.62
I believe that girls are less successful than boys at pursuing mathematics-related careers because of their genetic potential.	g3g	0.13	6.46	1.02	-2.16	4.3	0.26	0.22	0.59
Innate differences in ability largely account for those who excel in mathematics and those who do not.	g4n	0.73	4.44	1.42	0.11	-0.89	0.42	0.47	0.53
I think that innate gender differences account the large number of boys who excel in mathematics.	g4g	0.37	5.64	1.44	-0.81	-0.37	0.38	0.28	0.57
Inherent biological factors explain why some children demonstrate exceptional mathematical knowledge for their age.	g5n	0.75	4.21	1.63	0.09	-0.95	0.44	0.47	0.63
Inherent biological factors explain why Black children are less likely than White children to demonstrate high mathematical achievement.	g5r	0.21	6.21	1.17	-1.44	1.21	0.34	0.53	0.61
I believe that basic genetic differences often explain which students are identified as mathematically gifted.	g6n	0.63	4.82	1.51	-0.24	-0.9	0.45	0.49	0.68
I believe that genetic differences by race explain the large number of White children who are identified as mathematically gifted.	g6r	0.27	5.98	1.45	-1.34	0.78	0.31	0.36	0.51
Fundamental biological differences explain why some students have higher mathematical achievement than others.	g7n	0.70	4.56	1.55	0	-0.88	0.41	0.46	0.67
Fundamental biological differences explain why Asian students have higher mathematical achievement than White students.	g7r	0.28	5.96	1.32	-1.19	0.6	0.33	0.43	0.62
In my view, genes do not determine which students excel in mathematics.	g8nx	0.88	4.38	1.66	-0.06	-0.92	0.07	0.06	0.38
In my view, genetic factors explain why Hispanic and Latino students struggle to learn mathematics.	g8r	0.25	6.12	1.21	-1.25	0.58	0.38	0.53	0.65

* Labels are coded with “g” to indicate gender-specific, “r” for race-specific, and “n” for identity neutral wording, and with “x” to indicate negatively worded (i.e., counter-stereotype) items.

Table 2

Item wording and descriptive statistics for the AME-Social subscale.

Item	Label*	< 6 (%)	M	SD	skew	kurtosis	Item-total Correlation		
							1-factor	3-factor	4-factor
In my view, children's' interest in mathematics, science, and engineering is determined by how their parents raise them.	s1n	0.74	4.5	1.42	0.15	-0.84	0.42	0.44	0.52
In my view, boys' parents raise them to be interested in mathematics, science, and engineering.	s1g	0.77	3.97	1.66	0.31	-0.94	0.42	0.47	0.52
Cultural and religious expectations can profoundly influence which students work hard to master mathematics and which students give up.	s2n	0.77	4.1	1.68	0.08	-0.92	0.34	0.35	0.48
In my opinion, more boys than girls are identified as mathematically gifted because society expects boys to be more mathematical than girls.	s2g	0.77	3.74	1.85	0.35	-1.03	0.24	0.36	0.41
Differences in upbringing explain which students are selected for mathematically gifted programs.	s3n	0.73	4.34	1.57	0.09	-0.89	0.40	0.42	0.39
Differences in how girls and boys are raised explain why fewer girls than boys are selected for mathematically gifted programs.	s3g	0.74	4.14	1.65	0.18	-1.04	0.41	0.42	0.51
I think that differences in upbringing explain why some children are more likely than others to have an interest in mathematics.	s4n	0.92	3.38	1.34	0.56	-0.06	0.45	0.46	0.58
I think that differences in upbringing explain why Asian children are more likely than White children to have an interest in mathematics.	s4r	0.81	3.62	1.76	0.59	-0.7	0.50	0.49	0.50
In my opinion, some students are not interested in mathematics because of their cultural heritage.	s5n	0.66	4.7	1.52	-0.2	-0.74	0.40	0.36	0.53
In my opinion, Black students are not interested in mathematics because of their cultural heritage.	s5r	0.29	5.98	1.35	-1.16	0.28	0.43	0.55	0.36
Students often decide to pursue mathematics-related careers because of how they are raised.	s6n	0.90	3.47	1.35	0.44	-0.33	0.49	0.52	0.52
Hispanic or Latino children are less likely to pursue mathematics-related careers than White children because of how they are raised.	s6r	0.60	4.86	1.63	-0.18	-1.16	0.57	0.65	0.59
Students are more likely to succeed in mathematical professions if they have had a stable and supportive upbringing.	s7n	0.88	3.55	1.51	0.32	-0.6	0.44	0.48	0.43
Hispanic or Latino students are less likely than White students to succeed in mathematical professions because of their upbringing.	s7r	0.37	5.6	1.46	-0.87	-0.29	0.50	0.61	0.45
Socialization has minimal impact on children's interest in mathematics.	s8nx	0.68	3.14	1.32	0.26	-0.49	-0.07	-0.11	0.09
Gender socialization has minimal impact on girls' interest in mathematics.	s8gx	0.67	3.32	1.5	0.39	-0.52	-0.07	-0.05	0.07

* Labels are coded with “g” to indicate gender-specific, “r” for race-specific, and “n” for identity neutral wording, and with “x” to indicate negatively worded (i.e., counter-stereotype) items.

Table 3

Item wording and descriptive statistics for the AME-Educational subscale.

							Item-total Correlation		
Item	Label*	< 6 (%)	M	SD	skew	kurtosis	1-factor	3-factor	4-factor
Everyone who succeeds in mathematics has had at least one excellent teacher.	e1n	0.79	3.98	1.63	0.22	-0.79	0.32	0.3443	0.39
When boys do well in mathematics classes it is because they have had at least one excellent teacher in their life.	e1g	0.80	4.1	1.54	0.32	-0.68	0.47	0.375	0.62
Students who end up in advanced mathematics classes have had more instructional support and better learning opportunities than other students.	e2n	0.84	3.88	1.5	0.36	-0.71	0.46	0.4584	0.48
I think that boys do better than girls in advanced mathematics classes because they get more instructional support and better learning opportunities from their teachers.	e2g	0.50	5.17	1.66	-0.51	-0.84	0.33	0.434	0.41
I am convinced that students in math-intensive career paths have had better math teachers than those who end up in other careers.	e3n	0.80	4.28	1.45	0.04	-0.65	0.43	0.4754	0.48
I think that girls in math-intensive career paths have had better math teachers than girls who end up in other careers.	e3g	0.58	4.85	1.65	-0.3	-1.03	0.42	0.412	0.58
Students who attend better schools have higher mathematical achievement.	e4n	0.82	3.91	1.5	0.3	-0.68	0.47	0.484	0.46
White students have higher mathematical achievement than Black students because they go to better schools.	e4r	0.62	4.64	1.75	-0.12	-1.16	0.53	0.48	0.56
It is my opinion that when students struggle in mathematics it is because they have insufficient instructional support.	e5n	0.87	3.73	1.42	0.28	-0.57	0.26	0.2681	0.49
In my opinion, when Hispanic or Latino students struggle in mathematics, it is because they have insufficient instructional support.	e5r	0.74	4.05	1.79	0.26	-1.04	0.43	0.43	0.60
In my view, students who excel in mathematics usually have had more educational opportunities than students who do not excel in mathematics.	e6n	0.88	3.53	1.45	0.57	-0.43	0.53	0.519	0.54
I believe that Asian students who excel in mathematics have more educational opportunities than students from other groups who do not excel in mathematics.	e6r	0.66	4.48	1.76	0.03	-1.17	0.53	0.53	0.45
I think that poor instruction is the main reason that students do poorly in a mathematics class.	e7n	0.79	4.14	1.53	-0.02	-0.76	0.22	0.1962	0.48

I think that inadequate instruction is the main reason that Black students do poorly in a mathematics class.	e7r	0.66	4.38	1.86	-0.02	-1.16	0.41	0.37	0.61
No matter how good the instruction, some students will never achieve mathematical excellence.	e8nx	0.64	3.5	1.79	0.22	-1.09	-0.17	-0.1765	0.01
Even if teachers make an extra effort, few girls can achieve mathematical excellence.	e8gx	0.10	1.45	0.98	2.81	8.69	-0.24	-0.305	-0.07

* Labels are coded with “g” to indicate gender-specific, “r” for race-specific, and “n” for identity neutral wording, and with “x” to indicate negatively worded (i.e., counter-stereotype) items.

Table 4

Item wording and descriptive statistics for the AME-Personal subscale.

							Item-total Correlation		
Item	Label*	< 6 (%)	M	SD	skew	kurtosis	1-factor	3-factor	4-factor
Students who study more get higher scores on standardized mathematics tests.	p1n	0.76	4.32	1.51	-0.02	-0.72	0.34	0.3521	0.45
There are more boys than girls with high scores on standardized math tests because boys spend more time studying.	p1g	0.16	6.35	1	-1.68	2.5	0.33	0.296	0.27
Students struggle to learn mathematics if they do not put in the time and hard work that is required to succeed.	p2n	0.90	3.66	1.42	0.19	-0.54	0.26	0.3376	0.38
Black students struggle to learn mathematics because they do not put in the required time and hard work.	p2r	0.24	6.06	1.27	-1.28	0.72	0.48	0.61	0.39
When it comes to mathematics, students with grit and determination will succeed.	p3n	0.96	2.86	1.34	0.78	0.49	0.20	0.2642	0.31
When it comes to mathematics, the boys with grit and determination are those who succeed.	p3g	0.81	3.73	1.65	0.29	-0.92	0.41	0.329	0.41
Students who pursue a career that requires mathematics must put in more effort at school than their peers.	p4n	0.76	4.33	1.51	0.11	-0.8	0.49	0.5688	0.45
White students are more likely than Black students to pursue a career that requires mathematics because they put in more effort learning mathematics in school.	p4r	0.28	5.92	1.38	-1.26	0.73	0.48	0.63	0.35
The students who score highly on standardized math tests have spent more time studying than other students.	p5n	0.67	4.64	1.49	-0.15	-0.93	0.41	0.4162	0.55
White students score higher on standardized mathematics tests than Hispanic or Latino students because White students spend more time studying.	p5r	0.29	5.9	1.27	-0.97	-0.13	0.52	0.61	0.45
Students who end up studying advanced mathematics have put in more effort than those who do not.	p6n	0.79	3.96	1.57	0.29	-0.85	0.44	0.5107	0.53
Boys tend to get farther in studying advanced mathematics than girls because they put in more effort than girls do.	p6g	0.17	6.27	1.06	-1.58	2	0.44	0.385	0.36
The students who excel in mathematics rarely have to try very hard.	p7nx	0.94	2.85	1.51	0.58	-0.64	-0.23	-0.305	-0.16
How hard girls try in school has very little to do with their success in mathematics.	p7gx	0.92	2.69	1.61	0.91	0.03	-0.10	-0.158	0.04
Success in mathematics has very little to do with how hard students try in school.	p8nx	0.62	3.05	1.3	0.21	-0.79	0.00	-0.0067	0.22

Asian students will excel in mathematics whether or not they try very hard.	p8rx	0.19	1.74	1.03	1.29	0.67	-0.33	-0.5	-0.15

* Labels are coded with “g” to indicate gender-specific, “r” for race-specific, and “n” for identity neutral wording, and with “x” to indicate negatively worded (i.e., counter-stereotype) items.

Table 5

Grand mean and range of item means by wording type and source of attribution.

Source of Attribution	Identity specific, positively worded	Identity neutral	Negatively worded*
Genetic	6.01 [5.45, 6.46]	4.73 [4.17, 5.68]	4.38
Social	4.56 [3.62, 5.98]	4.01 [3.38, 4.70]	3.23 [3.12, 3.34]
Educational	4.52 [4.05, 5.15]	3.92 [3.53, 4.28]	2.48 [1.45, 3.50]
Personal	5.71 [3.73, 6.35]	3.96 [2.84, 4.64]	2.58 [1.74, 3.05]

**There was only a single negatively worded item with a Genetic attribution; two with Social and Educational attributions, and four with Personal attributions.*

Table 6

Testlet statistics by AME subscale.

Testlet Label	Items	Rating < 6		Fisher's Test		Testlet Distribution			Item-total Correlation		
AME-Genetic		Neutral	Specific	Odds-ratio	*p*	Neither	One	Both	1-factor	2-factor	4-factor
g1	g1n, g1g	0.77	0.19	2.1	0.038	0.21	0.63	0.16	0.39	0.39	0.50
g2	g2n, g2g	0.52	0.43	4.8	0.000	0.36	0.32	0.32	0.55	0.46	0.72
g3	g3n, g3g	0.38	0.13	9.2	0.000	0.59	0.30	0.11	0.57	0.51	0.66
g4	g4n, g4g	0.73	0.37	6.3	0.000	0.24	0.42	0.34	0.56	0.5	0.65
g5	g5n, g5r	0.75	0.21	14.2	0.000	0.24	0.55	0.21	0.62	0.65	0.74
g6	g6n, g6r	0.63	0.27	3.3	0.000	0.32	0.46	0.22	0.57	0.57	0.72
g7	g7n, g7r	0.70	0.28	11.0	0.000	0.29	0.45	0.27	0.59	0.61	0.75
g8	g8nx, g8r	0.88	0.25	1.9	0.115	0.65	0.33	0.02	0.31	0.47	0.47
AME-Social											
s1	s1n, s1g	0.74	0.77	2.7	0.000	0.10	0.30	0.6	0.41	0.42	0.52
s2	s2n, s2g	0.77	0.77	2.5	0.002	0.08	0.29	0.63	0.33	0.32	0.50
s3	s3n, s3g	0.73	0.74	2.1	0.005	0.10	0.33	0.58	0.49	0.5	0.51
s4	s4n, s4r	0.92	0.81	5.3	0.000	0.04	0.19	0.77	0.43	0.44	0.49
s5	s5n, s5r	0.66	0.29	3.5	0.000	0.29	0.47	0.24	0.56	0.58	0.51
s6	s6n, s6r	0.90	0.60	3.2	0.002	0.07	0.37	0.56	0.56	0.6	0.57
s7	s7n, s7r	0.88	0.37	3.7	0.002	0.11	0.54	0.35	0.56	0.61	0.52
AME-Educational											
e1	e1n, e1g	0.79	0.80	10.0	0.000	0.12	0.18	0.71	0.40	0.42	0.55
e2	e2n, e2g	0.84	0.50	2.1	0.012	0.11	0.45	0.44	0.45	0.42	0.54
e3	e3n, e3g	0.80	0.58	7.1	0.000	0.16	0.31	0.53	0.49	0.51	0.58

e4	e4n, e4r	0.82	0.62	4.6	0.000	0.12	0.31	0.57	0.53	0.49	0.54
e5	e5n, e5r	0.87	0.74	4.3	0.000	0.07	0.25	0.68	0.44	0.41	0.61
e6	e6n, e6r	0.88	0.66	3.9	0.000	0.07	0.31	0.62	0.59	0.58	0.58
e7	e7n, e7r	0.79	0.66	3.9	0.000	0.12	0.30	0.58	0.39	0.35	0.58
AME-Personal											
p1	p1n, p1g	0.76	0.16	2.2	0.041	0.22	0.64	0.14	0.42	0.45	0.52
p2	p2n, p2r	0.90	0.24	3.3	0.031	0.09	0.68	0.23	0.57	0.58	0.59
p3	p3n, p3g	0.96	0.81	2.2	0.172	0.01	0.20	0.79	0.27	0.31	0.23
p4	p4n, p4r	0.76	0.28	3.1	0.001	0.21	0.55	0.25	0.66	0.67	0.66
p5	p5n, p5r	0.67	0.29	2.3	0.002	0.27	0.50	0.23	0.60	0.56	0.68
p6	pn, p6g	0.79	0.17	17.3	0.000	0.20	0.63	0.17	0.54	0.52	0.60

Table 7

Fit statistics for the CFA Models of the AMES.

Model	Model Description	df	chi-sq	RMSEA	90% C.I.	CFI	TLI	SRMR
Model 1 - 28 items	Unidimensional	350	1811.315	0.115	[.110, .121	0.774	0.756	0.135
Model 2 - 28 items	Race & gender factors	349	1748.657	0.113	[.108, .0119]	0.783	0.765	0.133
Model 3 - 28 items	Four AME Factors	344	816.268	0.066	[.060, .072]	0.927	0.92	0.086
Model 4 - 24 items	Unidimensional (short)	251	1312.168	0.116	[.110,.122]	0.823	0.806	0.121
Model 5 - 24 items	Race & Gender factors (short)	251	1261.461	0.113	[.107, .120]	0.832	0.815	0.119
Model 6 - 24 items	Four AME factors (short)	246	508.97	0.058	[.051, .066]	0.956	0.951	0.069

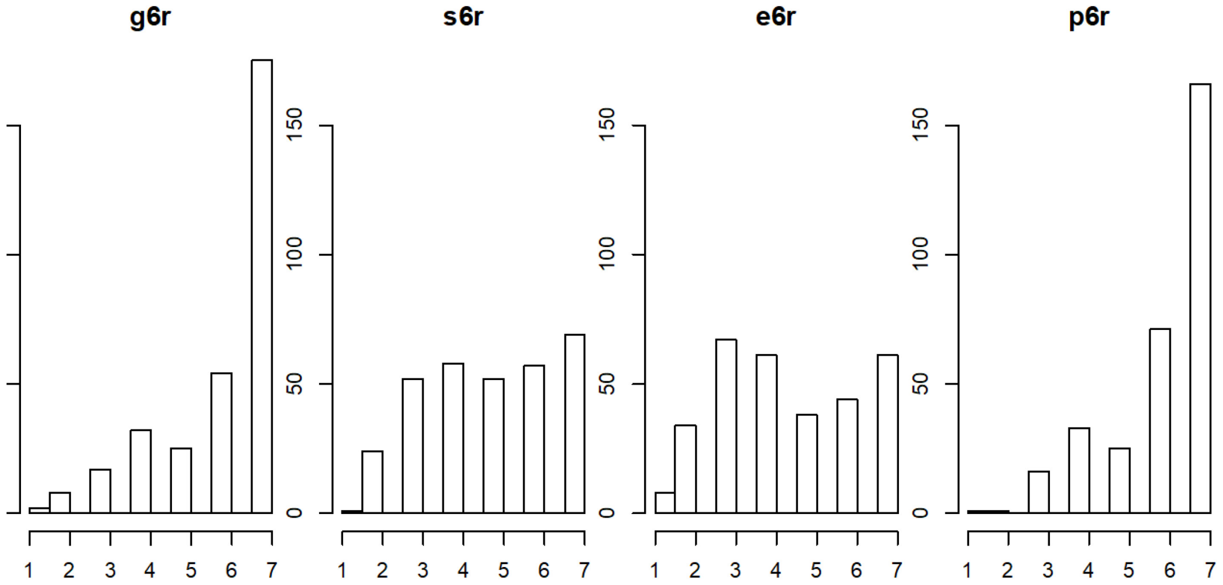
Table 8

Standardized Item Loadings for the CFA Models of the AMES.

Testlet Label		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
AME-Genetic	Items	1-factor	2-factor	4-factor	1-facor	2-factor	4-factor
g1	g1n, g1g	0.51	0.54	0.60	0.52	0.55	0.60
g2	g2n, g2g	0.74	0.78	0.83	0.75	0.79	0.83
g3	g3n, g3g	0.76	0.81	0.86	0.77	0.82	0.86
g4	g4n, g4g	0.70	0.74	0.79	0.71	0.75	0.79
g5	g5n, g5r	0.85	0.86	0.92	0.85	0.86	0.91
g6	g6n, g6r	0.77	0.78	0.84	0.77	0.79	0.84
g7	g7n, g7r	0.82	0.83	0.88	0.83	0.84	0.88
g8	g8nx, g8r	0.47	0.48	0.56	-	-	-
AME-Social							
s1	s1n, s1g	0.50	0.53	0.60	0.46	0.48	0.54
s2	s2n, s2g	0.41	0.43	0.51	-	-	-
s3	s3n, s3g	0.56	0.59	0.66	0.54	0.57	0.63
s4	s4n, s4r	0.60	0.61	0.69	0.59	0.60	0.68
s5	s5n, s5r	0.65	0.67	0.77	0.65	0.67	0.75
s6	s6n, s6r	0.70	0.72	0.81	0.70	0.71	0.81
s7	s7n, s7r	0.68	0.69	0.79	0.69	0.70	0.79
AME-Educational							
e1	e1n, e1g	0.51	0.55	0.67	0.50	0.53	0.66
e2	e2n, e2g	0.51	0.54	0.67	0.49	0.51	0.64
e3	e3n, e3g	0.57	0.60	0.72	0.55	0.58	0.69
e4	e4n, e4r	0.62	0.63	0.78	0.62	0.63	0.79
e5	e5n, e5r	0.59	0.60	0.76	0.53	0.53	0.68
e6	e6n, e6r	0.70	0.72	0.86	0.70	0.71	0.86
e7	e7n, e7r	0.50	0.51	0.67	-	-	-
AME-Personal							
p1	p1n, p1g	0.53	0.56	0.61	0.54	0.57	0.62
p2	p2n, p2r	0.72	0.73	0.81	0.72	0.73	0.81
p3	p3n, p3g	0.36	0.39	0.42	-	-	-
p4	p4n, p4r	0.80	0.81	0.91	0.80	0.81	0.91
p5	p5n, p5r	0.73	0.74	0.82	0.74	0.75	0.82
p6	pn, p6g	0.64	0.68	0.74	0.64	0.68	0.74

Figure 1

Items from all factors evidenced a bimodal distribution with local minima near 5.



Note: Items were rated from 1: Completely true to 7: Not at all true, with no intermediate labels).

Figure 2

Distribution of responses between AMES educational attributions e2g (positively worded) and e8g (negatively worded).

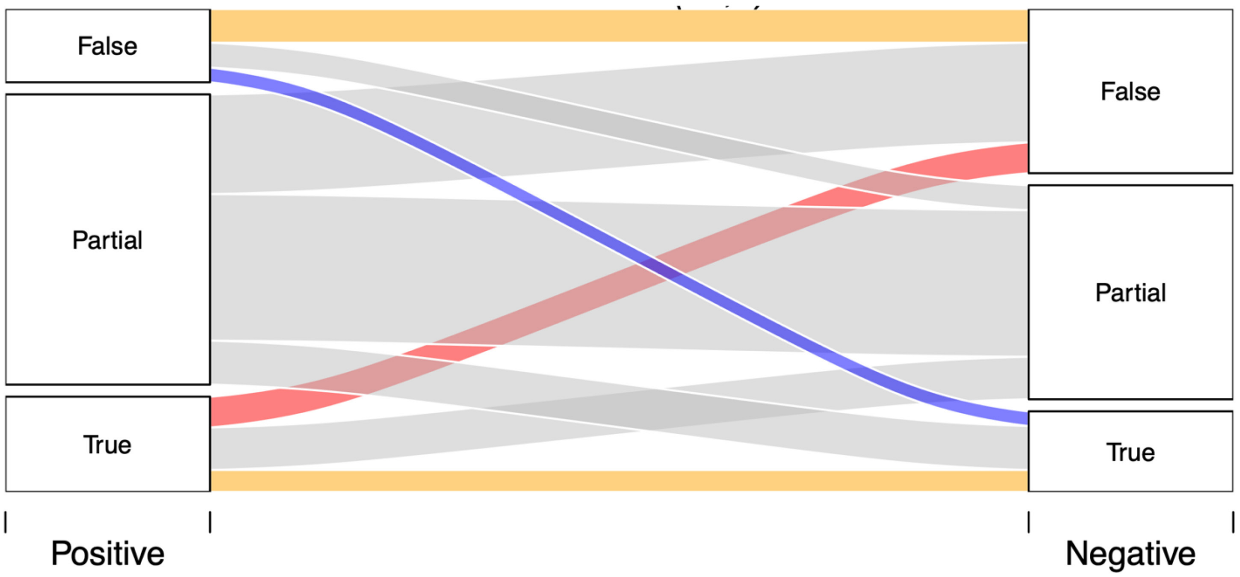


Figure 3

Path diagram for Model 1.

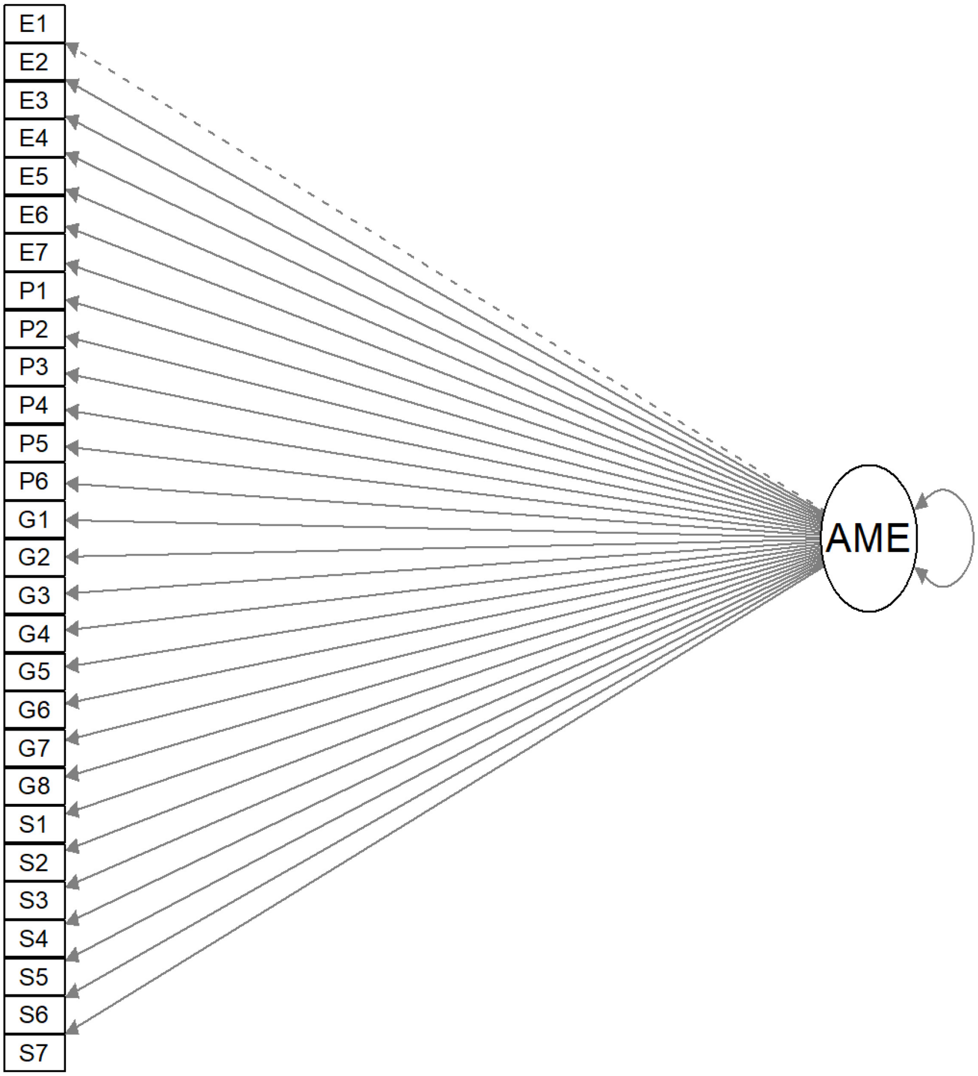


Figure 4

Path diagram for Model 2.

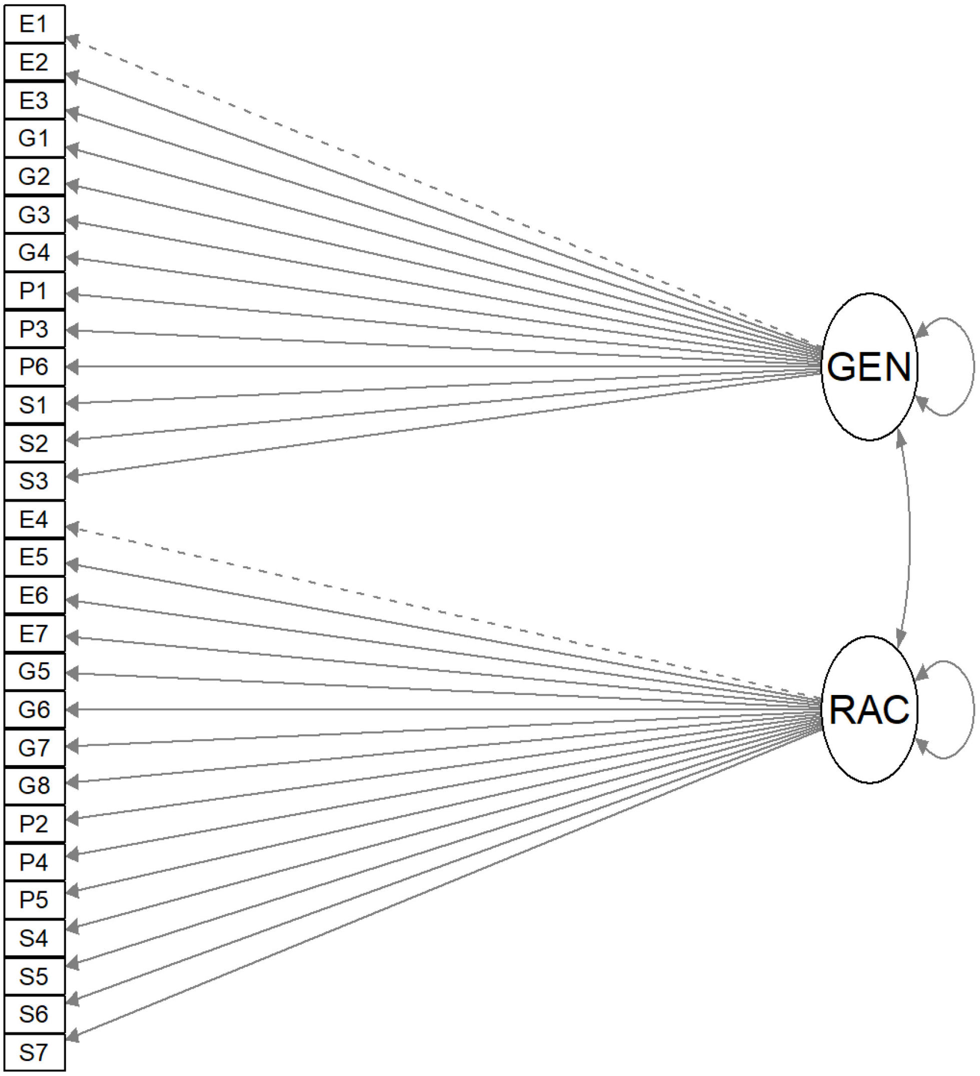


Figure 5

Path diagram for Model 3.

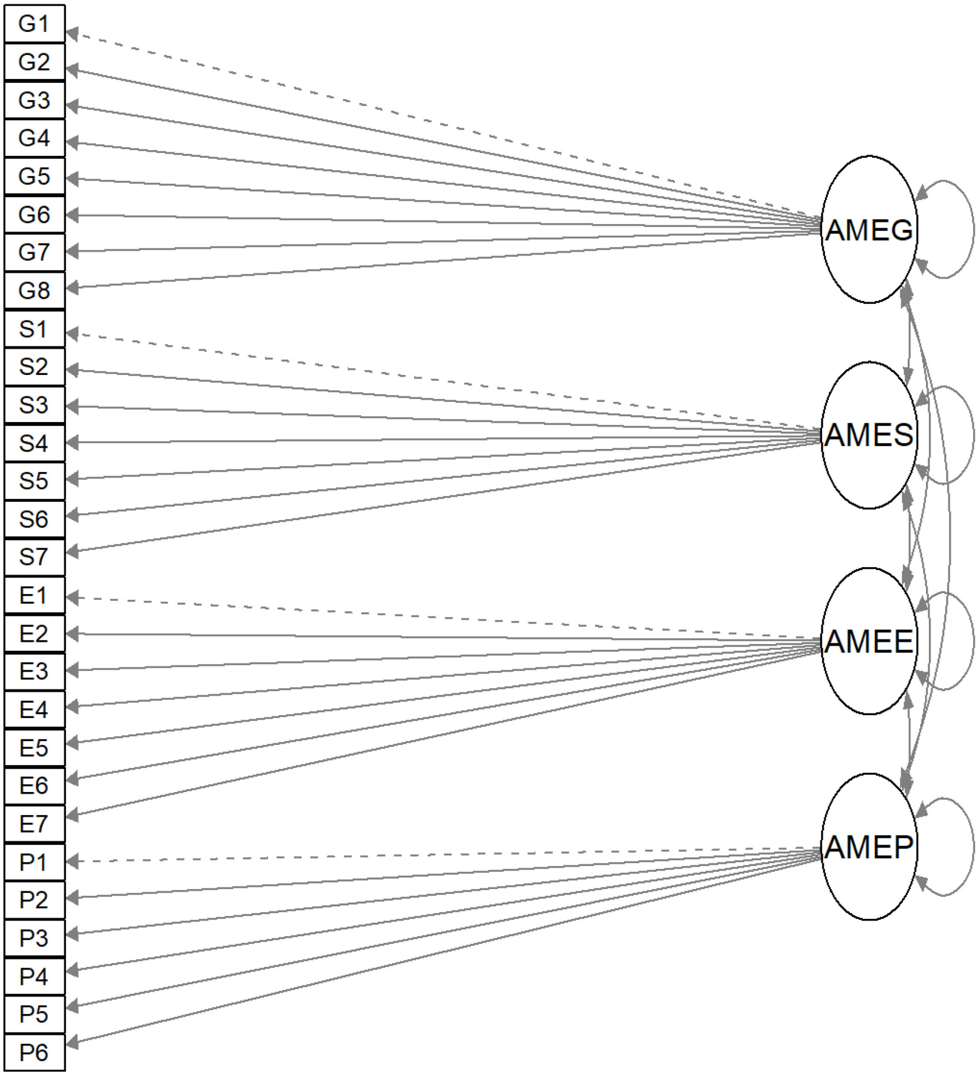


Figure 6

Path diagram for Model 7.

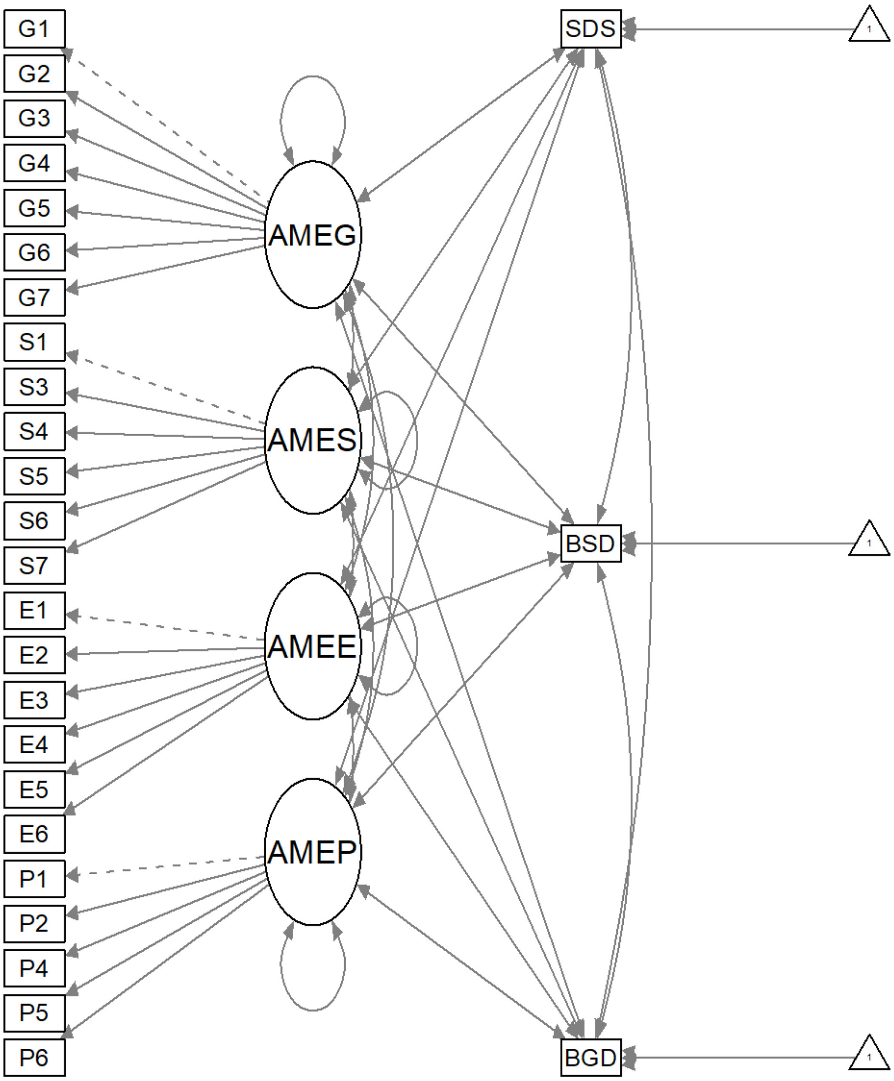
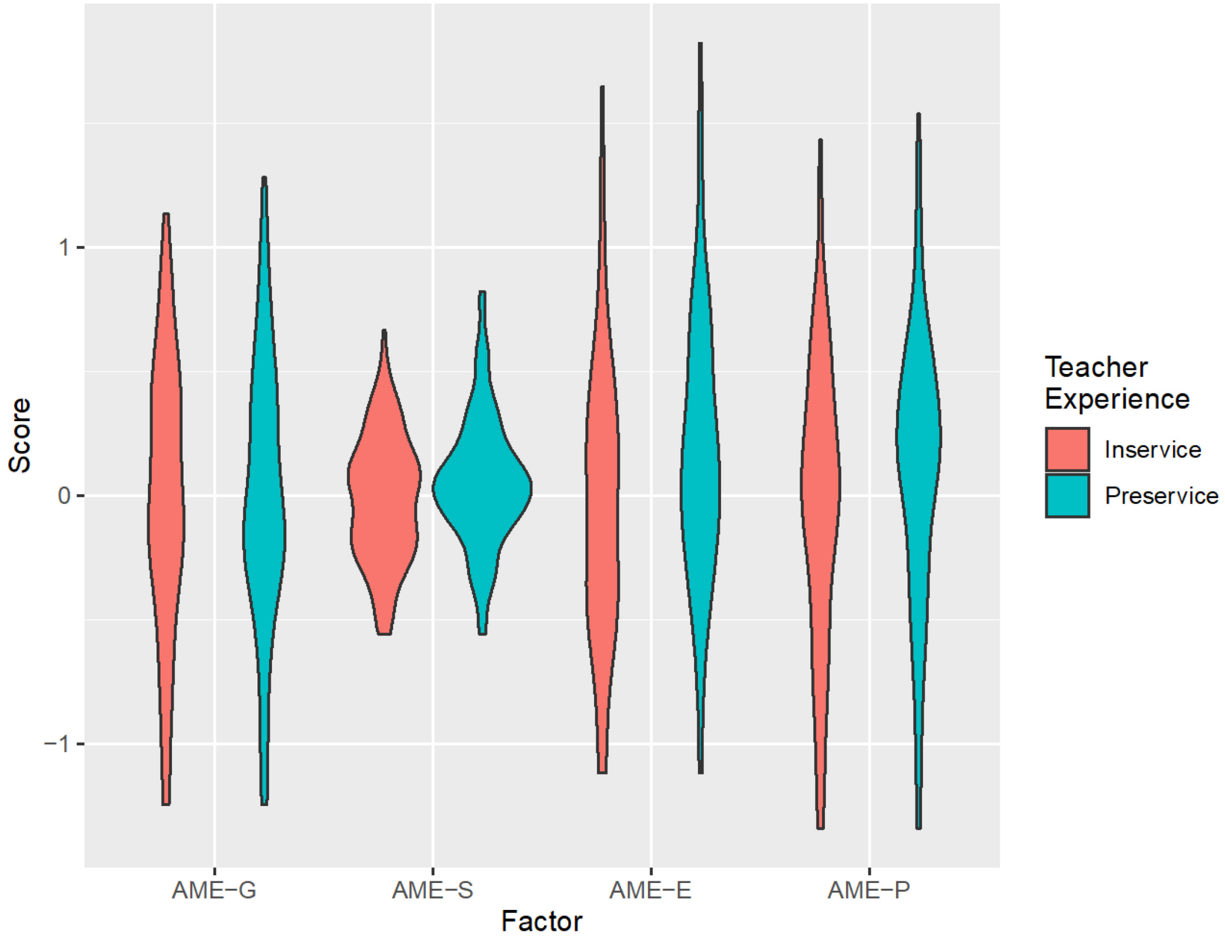


Figure 7

AMES factor score distribution of inservice versus Preservice teachers.



Supplemental Material

Participants

All the teachers surveyed had (or will have upon graduation) credentials to teach at the elementary (K-5) level but only 92% of the practicing teachers reported certification to teach all subjects. Among the practicing teachers, 95% had full credentials and 6 teachers (3%) had provisional certification. About 25% of the practicing teachers had between 0 and 5 years of teaching experience, about 50% were mid-career (6 – 20 years), and about 25% were late career teachers (over 20 years of experience). All but 6 of the practicing teachers (97%) were responsible for teaching math as a part of their curriculum.

In terms of preparation to teach mathematics, 37% reported having 1-2 postsecondary mathematics courses, 47% reported having 3-5 courses, 9% reported taking 6 or more mathematics courses, and 7% reported taking no mathematics courses during college. The distribution in mathematics education coursework was similar, with 67% reported having 1-2 mathematics methods courses, 22% reported having 3-5 method courses, 4% reported taking 6 or more mathematics methods courses, and 7% reported taking no mathematics methods courses during college. With respect to general education courses, about 79% of the participants took 6 or more general education classes which aligns with the fact that 91% of the practicing teachers hold a bachelor's degree in education. Only 1 participant had bachelor's degrees in Mathematics Education. About 10% held non-mathematics bachelor's degrees that varied including STEM, social sciences, and liberal Arts foci. With respect to professional development (PD), 14% of practicing teachers reported having zero hours of math-specific PD annually, 42% reported receiving less than 6 hours, 29% reported having 6-15 hours, and 16% reported 16 or more hours.

Instruments

Belief in Social Determinism (BSD) and Belief in Genetic Determinism (BGD) Scales. The BSD and BGD scales measure two components of psychological essentialism, the tendency of individuals to explain others characteristics and behaviors by way of their underlying essence (Keller, 2005; Rangel & Keller, 2011). Both instruments have high reliability (α BSD = 0.84; α BGD = 0.87) and are each supported by several validation studies. Five studies based on a common sample of 564 participants demonstrated strong psychometric properties including expected factor structure, high reliability, expected dimensionality, and expected correlations with other measures of essentialism and theoretically related motivational constructs (Rangel & Keller, 2011). Across these correlational studies and two additional experimental studies (N = 59, N = 64), findings consistently suggest that both BSD and BGD “imply negative consequences, such as stereotyping, prejudice, and discriminatory tendencies” (Rangel & Keller, 2011, p. 21).

Social Desirability Scale (SDS-17). The SDS-17 ($\alpha = 0.75$; Stober, 2001) is an updated instrument for measuring desirable responding designed to update and replace the Marlowe-Crowne Scale (Crowne & Marlowe, 1960) as a reliable and valid measure of social desirability for adults. The Marlowe-Crowne Scale was the standard instrument used to control for social desirability for four decades, but by that point some of the questions were outdated. The SDS-17 was found to be correlated with other measures of social desirability, including the Marlowe-Crowne Scale. Furthermore, the SDS-17 was sensitive to experimental conditions that were designed to provoke socially desirable responding, such as a job interview (Stober, 2001). By using the SDS-17 we attended to the possibility that survey responses may not reflect genuine beliefs as much as participants sense of what beliefs are socially desirable.

Item Analysis

The identity specific AMES items were created first, with four gender- and four race-specific items for each factor. We wrote an identify neutral version of each item by replacing the gender or race descriptors with a word or phrase which did not specify identify, e.g., “Black students” and “Girls” became “Students. From among these we modified a handful of items to use negative wording (i.e., contradicting the implicit stereotype). As described in the main paper, the item-total correlations for the negatively worded items was consistently low across all three of the factor structures we considered.

As a result of our analyses to answer RQ1 and RQ2, we identified problematic items (i.e., high skew, low item-total correlation, or low factor loadings), investigated to determine why they were behaving unexpectedly, and removed them from further steps of analysis when substantive reasons explained the poor item performance. The problematic item-total correlation results for the negatively worded AMES items led us to conduct an additional analysis to identify why these items were performing so poorly. We used alluvial diagrams to further investigate by examining the relationship between negatively worded items and similar items with positive wording. Figure 2 presents an illustrative case, with the distribution responses between the positively worded item e2g on the left and the negatively worded item e8g on the right. More of the participants who rated the gendered education attribution as false went on to rate a negatively worded gendered education attribution in the same way than those who reversed their ratings. In this way, the negatively worded AMES items (except for g8n) failed to capture opposite beliefs as intended, perhaps indicating that participants who disagreed with a specific attribution belief did not hold an opposite attribution belief as much as a different attribution belief.

We computed Cronbach's alpha for all the items together and for each of the four hypothesized factors. As a further check on the utility of the negatively worded items, we computed alpha both with and without the negatively worded items. Overall, alpha was .90 with the negatively worded items and .92 without. For the genetic attribution, alpha was .90. (We retained the only negatively worded item, g8g). Alpha was .83 for the social attribution items, and 0.85 without the negatively worded items. Alpha was .83 for the educational attribution items, 0.86 without the negatively worded items. Finally, alpha was .71 for the personal attribution items, 0.80 without the negatively worded items. We concluded that the items demonstrated adequate internal consistency overall and for the hypothesized four-factor structure. Moreover, removing the negatively worded items increased internal consistency, especially in the case of the personal attribution items. We retained item g8g but dropped all other the negatively worded items from subsequent analysis.

After excluding the negatively worded items, the five lowest item-total correlations under the four-factor structure were acceptable at 0.27, 0.31, 0.35, 0.36, 0.37, and 0.38. These correlation patterns answer RQ2 by showing that the identity neutral items tapped the same construct as the original race- and gender-specific items. Moreover, almost all items had the highest item-total correlations under the four-attribute structure. The item-total correlation results contributed to RQ3 by giving initial confidence in the hypothesized four-factor structure over the alternative structures that we evaluated.

Testlet Analysis

We evaluated the structural validity of the testlets in several ways. First, we compared the responses on each pair of dichotomized items composing a testlet to see if they were consistent with the theorized relationship that the identity neutral items were more likely to be rated as true or partially true than were analogous identity specific items (see RQ2). The percentage of true or partially true ratings for the identity neutral and identity specific items are presented in the first two columns of Table 6. Of the 28 item pairs, 24 pairs had the expected relationship and in the remaining four pairs the percentages were equal or very close (s1: 74% vs. 77%; s2: 77% vs. 77%; s3: 73% vs. 74%; e1: 79% vs. 80%). No item pairs included an identity specific item that was substantially more likely to be rated true or partially true than the corresponding identity neutral item.

We also used Fisher's test to probe the odds ratio within each pair of items to understand whether individuals who rated the identity neutral item true or partially true had higher odds to rate the identity specific item true or partially true (see RQ2). These results are presented in the fifth and sixth columns of Table 6. Of the 28 item pairs, the lowest odds ratios was 1.9 and the median was 3.6. All tests were significant at the .05 level except for testlet p3 (OR = 2.2; *p* = .172) and testlet g8 (OR = 1.9, *p* = .115). Among the remaining testlets, participants were two or more times as likely to rate the identity specific item as true or partially true if they rated the identity neutral item as true or partially true, as predicted by theory.

Taking each testlet as an item, we considered the item total correlations of the testlets under three possible factor structures: a single factor, two factors for race and gender specific items, and four factors for the different types of attributions: genetic, social, educational, and personal. These three sets of item-total correlations are presented in Table 6 and they extend the previously reported item-total correlations for the items. We found that all but one of the testlest (p3) had the highest item-total correlations for the four-attribute structure. Testlet p3 had an item total correlation of .27 in the unidimensional structure and an item-total correlation of .31 in the two-factor structure, but only 0.23 in the four-factor structure. We also found that the next four lowest item-total correlations under the four-factor structure were reasonably high at 0.47 (g8), 0.49 (s4), 0.50 (s2), and 0.50 (g1). These results gave us confidence that the testlets supported the hypothesized four-factor structure.

Finally, we computed a coefficient for reliability under the three conjectured structures: unidimensional, two factors, and four factors. If considered a unidimensional scale, the 28-items had a Cronbach's alpha of .90. When considered as a two-factor scale, the race-specific items had a Cronbach's alpha of .87, and the gender-related items had a Cronbach's alpha of .81. When the hypothesized four-factor attribute structure was used, all factors had high internal consistency (AME-Genetic, α = .90; AME-Social, α = .79; AME-Educational, α = .82; AME-Personal, α = .80). Notably, the reliability of the four hypothesized scales when based on testlets was not appreciably smaller than when using items (reported above) even though the number of testlets was half the number of items.

Removing Problematic Testlets

We attempted to reduce the number of items on the AMES to increase scale quality and parsimony. We examined empirical evidence of item misfit by considering factor loadings and looking at item statistics including the testlet odds ratio and item-total correlation. We also investigated within-factor item redundancy by examining the modification indices (MIs) to identify item pairs which would improve model fit by allowing correlated errors. Statistically significant correlated errors of items measuring the same latent construct is a violation of the local independence assumption of latent variable models and can be used as an indicator of potential redundancy (e.g., Ganley, et al. 2019). Before removing items, we carefully considered the item wording in light of theory and any implications for construct representation in the modified scale.

In the AME-Genetic factor, item g8 (Race - genetic factors, excel in / struggle to learn mathematics) had low odds ratio of 1.9 and a non-significant Fisher test (*p* = 0.115). Furthermore, the item total correlation of g8 with the rest of the AME-Genetic items was below 0.5 (0.47). Considering wording, this was the only testlet which included negative phrasing. Given that all the other negatively phrased items functioned so poorly they did not warrant testlets, we judged that excluding this item would make the factors have more parallel meaning by not including any negatively phrased items. Although this limits the operationalization of the constructs, it does so in a consistent way across all four of the AME constructs. We removed item g8.

In the AME-Social factor, s1 (Gender - how parents raise them, interest in mathematics) and s2 (Gender - cultural and religious / societal expectations, work hard [at] math / gifted) were empirically identified as being potentially redundant (MI = 43.57). Although s1 comprised two rating items that were precisely parallel, the wording for s2 items varied somewhat, with the identity neutral statement attributing hard work in mathematics to cultural and religious expectations and the gender-specific item attributing gifted status in mathematics to societal expectations. This item was also problematic because the identity neutral item used a marker of mathematical excellence ("work hard to master mathematics") which was not similar to any other markers of excellence but rather was similar to the attribution sources for AME-Personal items. Looking back at our item development records, it seems that this item was retained from an earlier version of the instrument and was not appropriately edited to reflect the addition of the AME-Personal scale. We removed item s2.

In the AME-Educational factor, the items e5 (Race - insufficient instructional opportunities, struggle in mathematics) and e7 (Race - poor / inadequate instruction, do poorly in math class) were empirically identified as being potentially redundant (MI = 41.08). The items do have considerable overlap in meaning. The biggest difference we could discern is that e5 refers to “Hispanic or Latino” students whereas e7 refers to “Black students.” Another item (e4) also refers to Black students, but e5 is the only item that refers to Latino students, so we removed item e7.

In the AME-Personal factor, item p3 had a low odds ratio of 2.2 and a non-significant Fisher test (*p* = 0.172). Furthermore, the item total correlation of p3 with the rest of the AME-Personal items was low at .23. Considering the wording (Gender - have grit and determination, succeed in mathematics), this was the only testlet which included the word “grit” a phrase which has received a lot of attention in the press and in schools in the last 5 years (Duckworth et al., 2007). This context may explain why the rating items were almost universally accepted: 99% of participants agreed with one or both, leaving little room for the testlet to distinguish between participants. This item poor performance was likely due to this restriction in range. We removed item p3.

AMES Factor Scores by Teacher Service

The sample for this study included both inservice and preservice teachers, and a natural question is whether the instrument functioned in similar ways across these different teacher populations. We did not have enough preservice teachers to evaluate measurement invariance across these groups. Therefore, as a preliminary step pending future research, we examined the distribution of factor scores for preservice and inservice teachers (see Figure 7). Based on this descriptive analysis, it appeared to us that the distribution of each factor was very similar between the preservice and inservice teachers in our sample. Future research with larger samples from each population is required to fully investigate this important question.