Evolution: Education and Outreach

RESEARCH

Open Access

Measuring evolution acceptance using the GAENE: influences of gender, race, degree-plan, and instruction

Gena C. Sbeglia[1*] and Ross H. Nehm[2]

Abstract

Background: The evolution education research community has defined the construct of "evolution acceptance" in different ways and measured it using different instruments. One of these instruments—the GAENE—has not been analyzed across different student populations, demographic groups, degree plans, and instructional treatments. Such comparisons are crucial for examining whether the inferences drawn from instrument measures are valid, reliable, and generalizable. In this study, we attempt to replicate findings produced in the original validation study and explore aspects of the instrument not previously examined.

Methods: We use Rasch analysis to study a large sample (n > 700) of undergraduates enrolled in standard introductory biology classes in the Northeastern USA. Participants completed the GAENE pre- and post-course for two semesters, and the MATE pre- and post-course for one semester. We assessed dimensionality, reliability, item fit, and rating scale functioning. We used regression analyses and generalized eta squared to evaluate the contribution of demographic and background variables to pre-course measures and pre-post course acceptance gains.

Results: Our analyses of GAENE dimensionality and item properties were generally in line with prior work, including the finding that particular items displayed psychometric problems. Surprisingly, GAENE measures did not differ between biology majors and non-majors. Evolution instruction produced significant but small pre-post improvements in GAENE measures. GAENE measures were significantly associated with MATE measures (0.68–0.80). White and male participants had the highest evolution acceptance measures using both the MATE and the GAENE; race had a much stronger contribution to MATE measures as compared to GAENE measures. Race and gender acceptance differences were found to be as large as the differences produced in response to evolution instruction.

Conclusions: Overall measures of acceptance change will be similar, but not identical, using the MATE and the GAENE. We make several recommendations for the modification or removal of some GAENE items, as well as future research directions for the measurement of evolution acceptance.

Keywords: Evolution acceptance, Rasch, GAENE, MATE, Validation, Race, Gender, Degree plan

Introduction

The evolution education research community has attempted to define the construct of "evolution acceptance" and to empirically measure it using three instruments: the Measure of Acceptance of the Theory of Evolution (MATE) (Rutledge and Warden 1999), the Inventory of Student Evolution Acceptance (I-SEA) (Nadelson and Southerland 2012), and the Generalized Acceptance of EvolutioN Evaluation (GAENE) (Smith et al. 2016). Although all three instruments have been used to measure evolution acceptance in *separate* studies using *different* participant samples across a variety of educational levels and geographic regions, remarkably few studies have (1) replicated validity claims (psychometrically or conceptually), (2) compared how the measures derived from different instruments function in the

*Correspondence: gena.sbeglia@stonybrook.edu
[1] Department of Ecology and Evolution, Stony Brook University, Stony Brook, USA
Full list of author information is available at the end of the article

same populations, or (3) examined how gender, race, and academic background impact acceptance measures. A better understanding of evolution acceptance measures is important for aligning the findings of different studies and ensuring that validity inferences for instrument measures generalize to a broad range of educational contexts and participant samples (AERA, APA, and NCME 2014). For example, some research has found that magnitudes of evolution acceptance differ across demographic groups (e.g., underrepresented minorities [URM] vs. white males; Metzger et al. 2018; Pew 2015). Many aspects of the measurement of evolution acceptance remain in need of empirical and conceptual attention.

The MATE has been the mostly widely used instrument to measure evolutionary acceptance, but it has notable weakness, including: limited validity testing; conflation of evolutionary acceptance, knowledge, and religiosity; signatures of multidimensionality; and items that lack clear alignment to evolutionary scales and contexts (Romine et al. 2017; Smith et al. 2016; Sbeglia and Nehm in press). In a recent study, Romine et al. (2017) addressed some of these concerns, most notably conducting validity testing using Rasch analysis and reconceptualizing the instrument as two dimensional.

The I-SEA instrument was developed to address some of the limitations of the MATE. Specifically, the I-SEA measures only acceptance—not belief, religiosity, or knowledge (Nadelson and Southerland 2012). In addition, it assesses acceptance in specific aspects of evolution using three item sets: microevolution, macroevolution, and human evolution (Nadelson and Southerland 2012). However, like the MATE, the I-SEA has weaknesses including: limited validity testing and replication (Smith et al. 2016); the inclusion of both microevolution and macroevolution items in the human evolution item set; and signatures of multidimensionality in the human evolution item set (Sbeglia and Nehm in press).

In an attempt to address criticisms of both the MATE and the I-SEA, Smith et al. (2016) developed the GAENE. The GAENE contains 13 items intended to measure generalized evolution acceptance as a unidimensional construct. The GAENE items ask respondents about their acceptance of patterns of change (1 item), their acceptance of evolution as true and/or explanatory (6 items), their willingness to argue in favor of evolution in public (2 items), and the importance of understanding or appreciating evolution (4 items). Furthermore, although the GAENE was designed to test *generalized* evolution acceptance, some items invoke a specific organismal context (e.g., item 5: plants, animals, humans; item 8: bacteria; item 12: humans), some invoke specific evolutionary scales (e.g., item 8: microevolution; item 12 and 14: speciation/macroevolution), some invoke both (e.g., item 8: microevolution in bacteria; item 12: macroevolution of humans), and other items are abstract (e.g., they do not specify a scale or a context).

The GAENE has been the subject of validity testing using Rasch methods in a sample of high school and undergraduate students across the United States (n > 650). However, the GAENE has not yet been psychometrically analyzed in contiguous populations across geographic regions, across semesters of the same class, across gender and racial groups, and among participants with different degree plans. It also has not been analyzed in a pre- to post-course study design. These comparisons are important aspects of validity testing because they provide evidence that the inferences drawn from the instrument can be appropriately generalized across groups.

Robust measurement of magnitudes of evolution acceptance may be relevant to observed patterns of differential persistence in Science, Technology, Engineering, and Mathematics (STEM) degree programs (PCAST 2012). In particular, race and gender have received considerable attention as likely contributors to STEM persistence (e.g., Gender: Lauer et al. 2013; Wright et al. 2016; Race: Ma and Liu 2015; Nehm and Schonfeld 2008). The contributions of race and gender to evolution acceptance—which is a central feature of the life sciences—remain understudied. Well-validated tools capable of measuring evolution acceptance across a diversity of respondents is an essential first step towards generating robust inferences that can inform evidence-based interventions.

Research questions

In this study, we use Rasch analysis to examine the psychometric properties of the GAENE instrument. We first attempt to replicate findings produced in the original validation study of the GAENE. We go on to explore aspects of the instrument that were not previously examined. Specifically, we ask: (RQ1) Do Rasch analyses of pre- to post-course GAENE measures from a large sample (n > 700) of undergraduates align with prior validation work? (RQ2) Are GAENE measures sensitive to evolution instruction? (RQ3) Does the GAENE measure comparable levels of evolution acceptance between genders, among races, and across intended degree programs (e.g., biology majors and non-majors)? And (RQ4) To what extent do GAENE measures align with the most widely-used evolution acceptance instrument (i.e., the MATE)?

Materials

Course

The course examined in this study is a large (n > 250), 3-credit, undergraduate introductory biology class at a research-intensive (R1) public university in the

Table 1 Overall consent rates, demographic breakdown, and final sample sizes (after incomplete and problematic responses were removed)

Semester	% Non-consenting	% Consenting	% Consenting female	% Consenting URM	GAENE	MATE
Fall 2016	37.60	62.40	53.14	19.50	295	282
Spring 2017	29.89	70.11	56.13	27.67	475	–

Northeastern United States. This course is taken early in the academic careers of both biology majors and non-majors. It is a stand-alone course without a lab section. The prerequisites for this course include high school biology and freshman-level mathematics. The course content is aligned with the five core concepts of biological literacy described in the American Association for the Advancement of Science's *Vision and Change* policy document (Brewer and Smith 2011). Central themes in the course include microevolutionary processes (e.g., mutation, natural selection, genetic drift) and macroevolutionary patterns (e.g., phylogenetics, fossil records, biodiversity). A unit on the nature and practice of science is taught at the beginning of the course, which focuses on observations, facts, laws, models, inferences, theories, and experiments. The course is taught by an overlapping team of three instructors (Ph.D.s in evolutionary biology). The course does not address or discuss acceptance of evolution or religiosity at any point during the semester. Therefore, the course represents a standard approach to biology instruction that is common in undergraduate biology education in the United States.

Participants

Participants in two semesters (Fall 2016 and Spring 2017) were invited to complete the GAENE instrument pre- and post-course. In one of the semesters in which participants completed the GAENE (Fall 2016), we also invited participants to complete the MATE at the beginning and end of the course (Table 1). An average of 76% of participants (n = 823; 55% female and 23% underrepresented minority [URM]) consented to both the pre- and the post-course survey across the two semesters (Table 1). URM students included those who identified as Black/African American, American Indian/Alaska Native, Hispanic of any race, or Native Hawaiian/Other Pacific Island. In addition, we gathered demographic and background variables on the sample of consenting participants (e.g., gender, race, age, English Learner [EL] status, previous biology courses taken, intended degree program).

Instrument

The GAENE 2.1 (Smith et al. 2016) is composed of 13 Likert-scale items (numbered 2–14). Although the authors recommend a 5-option response format in GAENE 2.1, we used the four-option response format [i.e., strongly disagree (SD), disagree (D), agree (A), and strongly agree (SA)] of GAENE 1.0. The rating scale was scored from 1 to 4 and required respondents to choose between agreement and disagreement. The four-option response format was described in GAENE 1.0 but the authors chose to add an "undecided" option in later versions of the instrument after "…participants expressed a need for an option between acceptance and rejection" (Smith et al. 2016, p. 10). However, because the authors found little distinguishing power between levels of disagreement in the GAENE 2.1, and because GAENE 2.1 items were easier for students to agree with than the GAENE 1.0 items (Smith et al. 2016), we retained the rating scale of GAENE 1.0 (i.e., excluded the "undecided" option). All items are of the same valence. The four response options have three boundaries between them (known as "thresholds"): SD-D, D-A, and A-SA (see Andrich et al. 1997; Nijsten et al. 2006; Wilson 2005 for more information on Rasch thresholds).

Of the 823 (318 Fall 2016, 505 Spring 2017) participants that consented to the pre- and post-course survey in the two semesters that the GAENE was administered, some were excluded from the analysis if they (1) answered none of the GAENE items on either the pre- or the post-course survey (n = 40), (2) received a perfect measures on the pre- *and* post-course survey (n = 10), or (3) took the class previously (n = 3). The final data set for the GAENE analyses consisted of 770 participants (57% female, 22% URM).

The Measure of Acceptance of the Theory of Evolution (MATE) is composed of 20 Likert-scale items with a five-option response format (i.e., strongly disagree [SD], disagree [D], neutral [N], agree [A], and strongly agree [SA]) (Rutledge and Warden 1999). Of these items, 10 have been psychometrically shown to group into a "facts" dimension (i.e., these items measure the "facts and supporting data for evolution") and the remaining 10 items group into a "credibility" dimension (i.e., these items measure the "acceptance of the credibility of evolutionary science and rejection of non-scientific ideas") (Romine et al. 2017, p. 20). The MATE has negatively-worded items interspersed among positively-worded items. A positive answer is considered the normative response for the positively-worded items, and a negative answer is considered the normative response for the negatively-worded items. The five-option rating scale was scored from 1 to 5 and negatively-worded items were reverse coded.

Of the 318 participants that consented to the pre- and post-course survey in the semester in which the MATE was administered, some were excluded if (1) they answered none of the MATE items on either the pre- or the post-course survey (n = 14), (2) they received a perfect score on the pre- *and* post-course survey (n = 15), (3) they took the class previously (n = 3), or (4) had illogical answer patterns (n = 4). Students were classified as having illogical answer patterns if they agreed or disagreed with all instrument items (i.e., the same responses despite reverse coding across items). However, we were conservative in the removal of students based on these patterns because the MATE includes items that target knowledge, acceptance, and belief; different answers for different types of items may not be inherently illogical. The final data set for the MATE analyses consisted of 282 participants (57% female, 19% URM).

Methods

To address RQ1 (*Do Rasch analyses of GAENE measures from a large sample (n > 700) of undergraduates align with prior validation work?*), we examined several instrument properties: (1) dimensionality, (2) item and person reliability, (3) item fit, (4) rating scale functioning, and (5) person-item alignment (Wright maps).

Participants' raw response scores were converted into interval scale measures using a polytomous partial credit Rasch model in the R package Test Analysis Modules (TAM) v. 2.10-24 (Robitzsch et al. 2018). Before running the Rasch model, we modified the rating scale coding to begin at zero (e.g., 1–4 rating scale converted to 0–3). We ran a separate Rasch model for the pre- and post-survey by constraining items in the pre-survey Rasch model and then anchoring pre-survey Rasch item measures to the post-survey Rasch model (Wright 2003; see Sbeglia and Nehm in press for additional detail on these approaches). Rasch-transformed data are represented in "logits" and contain information about the difficulty of each item (known as "item difficulty") and the ability of each person (known as "person ability"), which share a common scale (Boone et al. 2014). Person ability is calculated using a weighted maximum likelihood estimation (WLE) of the item difficulty parameters. We used TAM to generate: Model fit statistics, item difficulties, person abilities, separation reliabilities, Wright maps, mean overall Rasch person measures as a function of the answer option selected for each item, Rasch-Andrich thresholds, and the frequency of participants selecting each answer option for each item. Collectively, these statistics can be used to evaluate the relative difficulty of the items and the extent to which they are productive for the measurement of the trait. Specifically, items that are productive for the measurement of the trait are those that behave as expected and that reliably separate respondents by their abilities. Each of these statistics are explained in detail below.

Dimensionality. We conducted a principal component analysis (PCA) of Rasch residuals to examine response pattern dimensionality. If the group of item response patterns being analyzed is one-dimensional, then the residuals should lack structure (e.g., an eigenvalue for the first contrast < 2). If the group of item response patterns being analyzed is multidimensional, then shared patterns will be apparent in the residuals, indicating that the group of items being analyzed share an attribute that was not accounted for in the one-dimensional Rasch model. In this case, the eigenvalue of the first contrast would be greater than 2. This approach is a standard method for evaluating the dimensionality of an instrument (Bond and Fox 2001). Smith et al. (2016) conducted an equivalent analysis in which they performed a PCA of Rasch *measures* (not a PCA of Rasch *residuals* as is frequently done) and analyzed the eigenvalue of the *second* contrast (which would be equivalent to the eigenvalue of the first contrast in a PCA of the Rasch residuals). If multidimensionality is suspected, the goodness of fit of the multidimensional Rasch model can be compared to the unidimensional Rasch model using a likelihood ratio test.

Item and person reliability. We used two methods to calculate reliability. The Expected A Posteriori/Plausible Value reliability (EAP/PV) index estimates if the order of item difficulties could be replicated in a different population with similar abilities. We also generated the WLE person separation index, which estimates if the order of person abilities could be replicated with a different set of items of similar difficulty (Bond and Fox 2001). Reliability values of greater than 0.70 are considered acceptable (Grigg and Manderson 2016; Yang et al. 2017).

Item fit. We calculated the fit of the items to the model by analyzing the weighted mean squares fit statistics for each item (WMNSQ; equivalent to infit MNSQ). Acceptable WMNSQ scores typically range from 0.7 to 1.3 logits, but a less conservative range of 0.5–1.5 logits is also used (Wright and Linacre 1994). High WMNSQ scores indicate that the data underfit the model and that items are poorly measuring the respondents for whom they are targeted.

Rating scale functioning. We assessed item-specific rating-scale functioning by evaluating the effectiveness of each item at separating respondents of different abilities. Failure to separate respondents could indicate unpredictability of the item response patterns. We used two related approaches to evaluate rating scale functioning. First, the mean overall Rasch person measures were examined as a function of the answer option selected for each item (Boone et al. 2014; Sbeglia and Nehm in press). If an item is functioning properly, there should be a correspondence between the participants' answer choices on a given item and their overall Rasch person measure, such that respondents that select the normative answer option for a particular item would have the highest Rasch person measures (Boone et al. 2014). A poor correspondence indicates that the item does not predictably discriminate person abilities.

The second approach to evaluate rating scale functioning involved the examination of Rasch-Andrich thresholds. These thresholds (also called step parameters or Andrich deltas) represent the locations on the Rasch category probability curve (see figure 2 from Smith et al. 2016, p. 17 for an example) where the curve for a given answer option crosses the curve for the subsequent answer option (Linacre 1999). If the thresholds are close together, or not in a sequential order (e.g., SD-D < D-U > U-A), then the items are unlikely to be discriminating person abilities in a predictable manner (Smith et al. 2016). This phenomenon is called rating scale disorder (or threshold disorder). Rating scale disorder occurs when participants that are predicted to receive a particular measure on an item (based on their other responses) instead receive a measure above or below this predicted value (Andrich 2013). Therefore, rating scale disorder is an anomaly that requires further examination and explanation (Andrich 2013). There are many possible explanations for rating scale disorder. Some of these explanations attempt to account for problems with the items, and some do not. For example, the generation of construct-irrelevant variation by an item could produce rating scale disorder and warrant the modification or removal of the problematic item (Andrich 2013). Unpredictable response patterns, and resulting rating scale disorder, may also be caused by participant guessing. This finding may not necessarily indicate that the items themselves are problematic. Rating scale disorder may also be associated with answer options that are selected by a small number of participants. For example, a low response frequency for some item options could amplify the impact of anomalous responses or guessing, resulting in rating scale disorder. The item and the rating scale would likely be retained in such cases. If the rating scale functions as expected for all but a few participants, the researcher may choose to not modify the item because it might be sufficiently productive for the measurement of the trait. For these reasons, rating scale disorder may not necessitate modification or removal of items (Adams et al. 2012; Boone et al. 2014), but it does indicate that the categories are not working as expected and that the nature and magnitude of the anomaly should be evaluated (Andrich 2013). Very little work has explored rating scale disorder using Rasch-Andrich thresholds for evolution instruments. Smith et al. (2016) used these Rasch-Andrich threshold patterns to evaluate the rating scale of the GAENE but in the format of Rasch category probability curves, not Rasch-Andrich thresholds per se. In summary, rating scale functioning and item fit were collectively used as metrics to assess the overall functioning and appropriateness of each item.

Wright maps. Wright maps plot item difficulties against person abilities and can be used to determine if the difficulties of the GAENE items were aligned with the abilities of the respondents. To generate Wright maps, we calculated the Thurstonian thresholds and item difficulties for each item (item difficulty = mean of the Thurstonian thresholds, see Sbeglia and Nehm in press for a further explanation of Thurstonian thresholds). Respondents at the top of the Wright map (with high logit measures) are estimated to have high abilities (i.e., high evolution acceptance), whereas those at the bottom of the map (with low logit measure) are estimated to have low abilities (i.e., low evolutionary acceptance). Conversely, items at the top of the map with high logit measures are more difficult (i.e., more challenging for participants to agree with) and items at the bottom of the map with low logit measures are less difficult (i.e., easier for participants to agree with). When respondents on a Wright map appear aligned with a specific Thurstonian threshold, there is an equal probability that the respondent selected an answer option that is above or below that threshold.

To address RQ2 (*How variable are GAENE measures across semesters, and are they sensitive to evolution instruction?*), we conducted a linear mixed-effects model with pre- and post-course GAENE measures as the outcome variable. We generated post-course Rasch person measures by anchoring the pre-course item difficulties and step parameters to the post-course Rasch model. We set instruction (pre/post) and semester as fixed effects, demographic and background variables as covariates (coding scheme for covariates described in "RQ3"), and person identifier as a random effect to control for repeated measures of the pre- to post-course design. We included interaction effects between instruction and several other variables (i.e., semester, race, gender, degree program, previous biology courses) to allow us to assess if there were differences from pre- to the post-course by semester. Because the regression model includes categorical variables, we report the unstandardized betas (*b*). Respondents that were missing any of the demographic or background variables were removed from the analysis.

Because we used a total of three regression models in this study (as described below), we used a critical p-value of 0.016 for all regression analyses.

To address RQ3 (*Does the GAENE measure comparable levels of evolution acceptance between genders, among races, and across intended degree programs?*) we utilized several general linear models. The model described in RQ2 (model 1) can address this research question, but for ease of interpretation, we ran two additional regression models and used a Bonferroni-corrected critical p-value of 0.016 (to account for the multiple tests). This approach resulted in complete correspondence of results between model 1 and the subsequent models (models 2 and 3) described below. In model 2, we conducted a linear regression model with *pre-course* GAENE measures as the outcome variable, demographic and background variables as fixed effects, and semester as a covariate. Demographic and background variables included: (1) Race (coded as "White", "Asian", "URM" [underrepresented minority: Black/African American, American Indian/Alaska Native, Hispanic of any race, Native Hawaiian/Other Pacific Island, Other]), (2) Gender (coded as "Male" or "Female"), (3) Intended degree plan (coded as "bio" [biology major], "non-bio STEM" [STEM major—Science, Technology, Engineering, Math–but not biology], "non-STEM" [not a biology or other STEM major]), and (4) Previous biology courses (coded as "none", "Advanced Placement biology only", "one introductory bio course", or "two introductory bio courses"). This model allowed us to analyze the influence of key demographic and background variables on pre-course measures. In model 3, we conducted a general linear model with *post-course* GAENE measures as the outcome variable, demographic and background variables as fixed effects, and semester and pre-course GAENE measures as covariates. This approach facilitated exploration of how key demographic and background variables influenced pre- to post-course gains. Respondents that were missing any of the demographic or background variables were removed from the analysis.

In the above models, we examined the magnitude of the unique impact (i.e., effect size) of each significant variable. We also examined the unique impact of the interaction between these significant variables. We measured this effect size using generalized eta squared ($\eta^2 G$) via the R package Analysis of Factorial Experiments (afex, v. 0.21-2) (Singmann et al. 2018). Generalized eta squared is more appropriate than eta squared when the study design includes measured factors (as opposed to manipulated factors). $\eta^2 G$ can also be more appropriately compared across studies and can be applied to repeated-measures designs (Bakeman 2005; Lakens 2013; Olejnik and Algina 2003). $\eta^2 G$ is a measure of the magnitude of the additional variance (R^2) explained by a particular variable compared to an otherwise identical model in which it is excluded. Cohen (1988) provides cut off values for η^2 (small effect = 0.01, medium effect = 0.06, and a large effect = 0.14); these values may also be used for the interpretation of $\eta^2 G$ (Olejnik and Algina 2003). The proper use and interpretation of effect sizes is an active area of research, and all measures have some limitations. For example, because $\eta^2 G$ can be biased by sample size (artificially increasing effect size estimates in small samples) (Olejnik and Algina 2003), several authors have argued that generalized omega squared ($\omega^2 G$) is more appropriate to use when comparing effect size across studies because it provides some correction for sample size bias (Bakeman 2005; Lakens 2013; see Levine and Hullett 2002 for a short review). However, because our sample contains > 200 respondents in our smallest analysis, and because of the substantial complexity of $\omega^2 G$ calculations, Lakens (2013) recommends using $\eta^2 G$ until $\omega^2 G$ is more broadly utilized and provided by statistical packages. In sum, we use $\eta^2 G$ to estimate the magnitude of significant effects.

To address RQ4 (*To what extent do GAENE measures align with the most widely-used evolution acceptance instrument?*), we examined the strength of the association between Rasch-converted GAENE measures and Rasch-converted MATE measures using data from the same study participants in the Fall 2016 semester. We fit the MATE dataset to a one-dimensional and a two-dimensional (i.e., a "facts" and "credibility" dimension as described above) Rasch model as recommended by Romine et al. (2017) and used a likelihood ratio test and AIC values to determine which model of dimensionality was a better fit to the data. We quantified the association between GAENE and MATE measures by comparing the nature and magnitude of: (1) The effect of instruction (pre- vs. post-course) on GAENE measures versus MATE measures. To this end, we analyzed pre- and post-course MATE measures using the same linear mixed-effects model used for the GAENE in RQ2 (model 1) and $\eta^2 G$ to calculate effect size; (2) The effect of race, gender, and plan on GAENE versus MATE measures. We analyzed MATE measures using the same regression models that we used for the GAENE (models 2 and 3), and calculated effect size using $\eta^2 G$; and (3) The association between Rasch GAENE and Rasch MATE measures using a Pearson correlation. A very high correlation between instrument measures (> 0.70) indicates that the two instruments are measuring acceptance in a similar manner and provides convergent validity evidence; moderate (0.50–0.70) or low correlations (< 0.50) indicate that the two instruments are measuring different aspects of the construct, or possibly, different constructs. We report

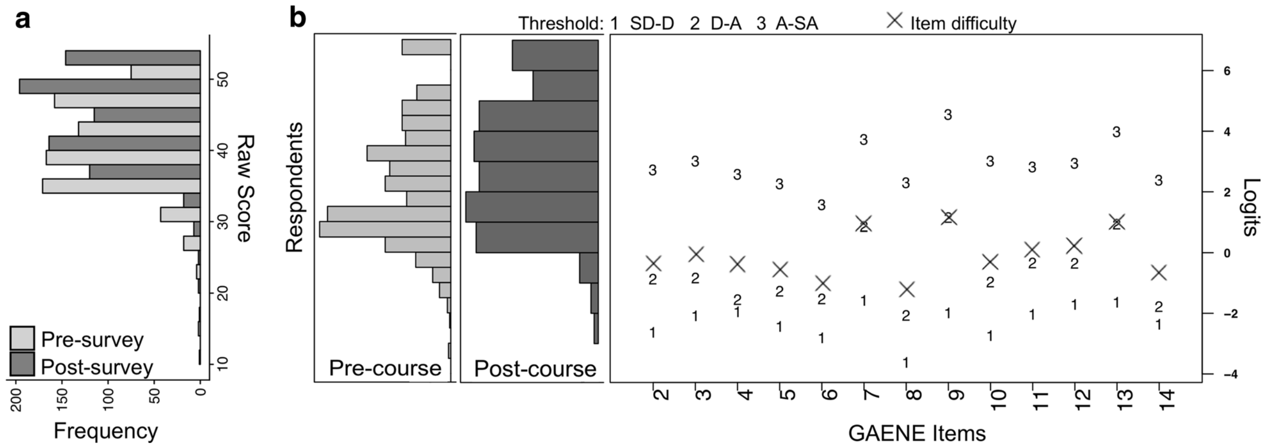

Fig. 1 **a** Raw pre- and post-course scores for the GAENE. Raw scores are out of 52 points. **b** Wright map showing Rasch-transformed measures pre and post-course. The pre-course item difficulties and step parameters were anchored to the post-course at model estimation

correlation coefficients that are both uncorrected and corrected (i.e., disattenuated) for measurement error so that our results may be compared to those of Metzger et al. (2018). Disattenuated correlation coefficients can be calculated by dividing the uncorrected correlation coefficient by the square root of the sum of the Rasch personreliabilities. We used this formula to convert Metzger et al.'s disattenuated correlation coefficients to uncorrected correlation coefficients. Because of their more widespread use in the literature, we focus our discussion on the uncorrected coefficients.

Results

RQ1

Raw GAENE scores were high in both the pre- and the post-course samples (Fig. 1a). The mean by-student pre-course score was 42.22/52 ± 6.23 ($\bar{x}$ by-item = 3.25/4 ± 0.23) and the mean by-student post-course score was 44.30/52 ± 6.05 ($\bar{x}$ by-item = 3.41 ± 0.21). The vast majority of respondents in this study selected the normative "agree" or "strongly agree" options for most items (e.g., items 2–6, 8, 10–12, and 14); very few respondents (< 12%) selected the non-normative "disagree" or "strongly disagree" options (Fig. 2a). By contrast, items 7, 9, and 13 had more than double the respondents (28–33%) select one of the non-normative options (Fig. 2a), making these the most difficult items to agree with (Table 2).

The residuals of the one-dimensional Rasch model had an eigenvalue of the first contrast less than 2 (1.84), suggesting that a unidimensional model captured an acceptable proportion of the variance in the dataset. The overall EAP/PV item separation and WLE person separation reliabilities were high (Table 3). When pre-course Rasch person abilities and item difficulties were plotted on a Wright map, the vast majority of participants were placed at or above the location of each item, indicating that these participants had a greater than 50% probability of selecting evolution-accepting answer options for most items (Fig. 1b). There is also a substantial gap where items did not align with respondent abilities (Fig. 1b).

We evaluated the functioning of the GAENE items by assessing their fit to the Rasch model, and the effectiveness of the rating scale at predictably separating respondents of different abilities. We summarize the results for each item in Table 4. Using the four-option response format of GAENE v. 1.0, items 2, 3, 5, 6, and 8, 10–12 were acceptable in the pre-course dataset using both metrics (see "Methods" for details). Specifically, these items had weighted MNSQ fit statistics within the acceptable range (although not always within the most conservative range) (Table 2). They also displayed a correspondence between participants' answer choices on these items and their overall Rasch person measures (Fig. 2b); these items meaningfully separated respondents based on their evolutionary acceptance levels at the pre-course. Similarly, the Rasch-Andrich thresholds displayed no disorder and thus acceptable separation (Fig. 2c). In the post-course, these items displayed acceptable weighted MNSQ fit statistics and a correspondence between participants' answer choices and their overall Rasch person measures; however, nearly all items (with the exception of item 5) displayed disorder of the Andrich thresholds for SD-D (Fig. 2f). Nevertheless, because very few participants (fewer than in the pre-course) chose the non-normative disagree answer options (Fig. 2b), and because the fit statistics were acceptable, these patterns of disorder are not likely indicative of problematic rating scale functioning.

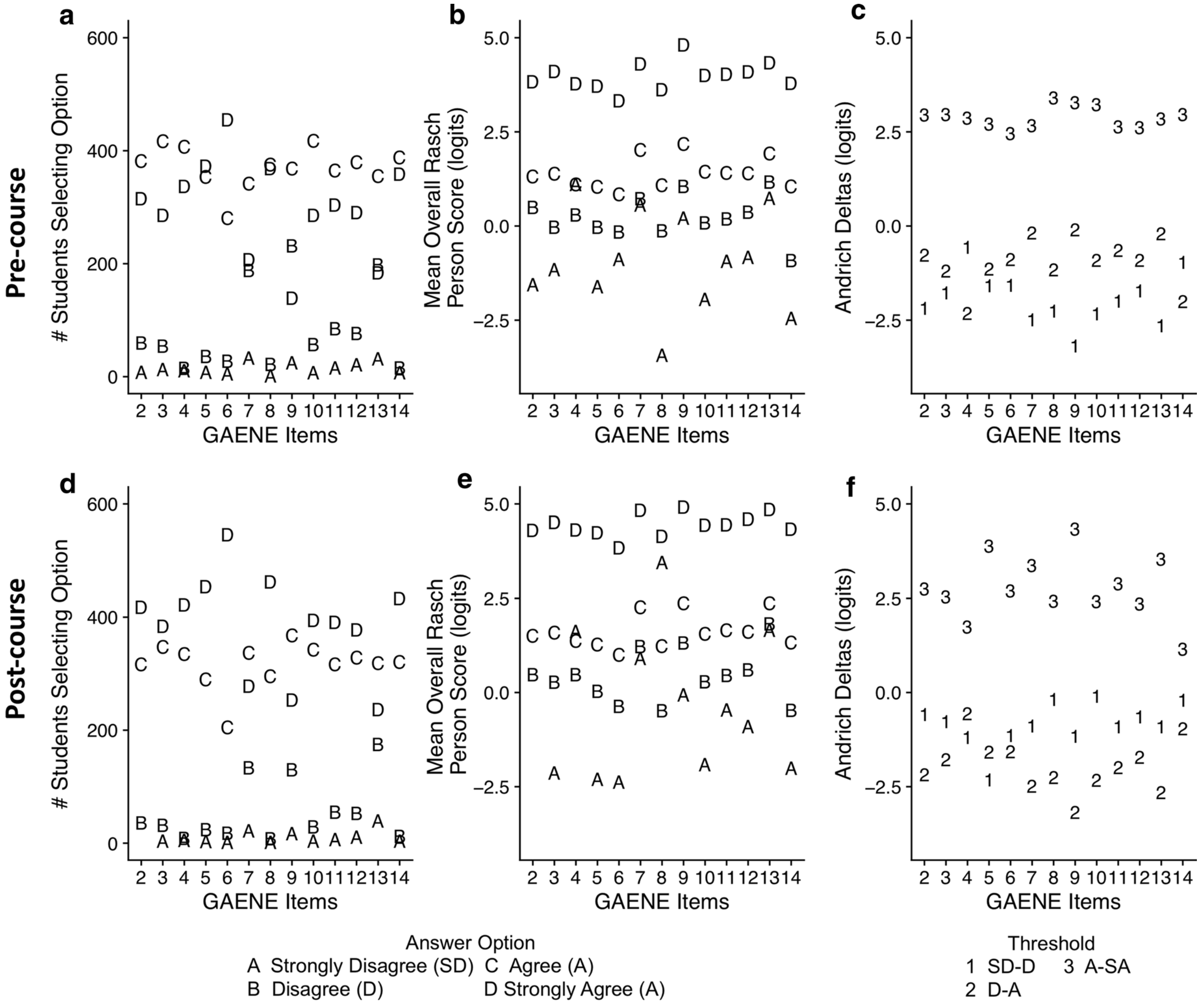

Fig. 2 Item functioning patterns for the GAENE. **a** The frequency of participants selecting each answer option. **b** The mean overall Rasch person measures as a function of the answer option selected for each item. **c** Andrich deltas (also called Rasch-Andrich thresholds or step parameters). Analogous statistics were used by Smith et al. (2016) to establish separation patterns. **d**, **e**, and **f** illustrate post-course results in parallel to the pre-course results shown in **a**, **b**, and **c**

Items 4 and 14 showed disorder in the rating scale (i.e., the Rasch-Andrich thresholds) in the pre- and post-course datasets (Fig. 2c, f). Furthermore, item 4 showed a poor correspondence between respondents' answer choices and their overall Rasch person measures (Fig. 2b, e). However, the low number of participants selecting the non-normative disagree options (Fig. 2a, d) and the sufficiency of the item fit statistics (Table 2) indicate that the rating scale of these items is likely not problematic.

In contrast, the patterns for GAENE items 7, 9 and 13 (see Table 5 for item text) *were* indicative of problematic rating scale functioning. First, in the pre- and post-course samples, these items had a poor correspondence with their overall Rasch person measures (Fig. 2b, e). Specifically, these items did not clearly distinguish the abilities of students that selected the non-normative options "strongly disagree" (option A) vs. "disagree" (option B). This pattern is not explained by low response frequencies for the problematic answer options. Rather, for these items, many more respondents selected the non-normative "strongly disagree" or "disagree" answer options in the pre- and post-course surveys than for the other items. For example, although 28.6–33.5% of respondents selected the non-normative "strongly disagree" or "disagree" for these items in the pre-course survey (Fig. 2a), they had relatively high mean overall Rasch person measures (Fig. 2b). The post-course survey showed similar patterns for these items (Fig. 2d, e). Thus,

Table 2 Item difficulties, and weighted (infit) and unweighted (outfit) MNSQ fit statistics of the GAENE

Item	Item difficulty	Unweighted MNSQ (outfit MNSQ)		Weighted MNSQ (infit MNSQ)	
		Pre-course	Post-course	Pre-course	Post-course
GAENE 2	− 0.25	0.97	0.84	1.00	0.89
GAENE 3	0.04	0.78	0.68	0.83	0.74
GAENE 4	− 0.30	1.12	1.02	1.10	0.98
GAENE 5	− 0.47	0.77	0.73	0.80	0.72
GAENE 6	− 0.91	0.83	0.60	0.93	0.79
GAENE 7	1.01	1.23	1.38	1.21	1.32
GAENE 8	− 1.12	0.85	0.78	0.89	0.90
GAENE 9	1.24	1.23	1.25	1.22	1.25
GAENE 10	− 0.22	0.86	0.74	0.88	0.80
GAENE 11	0.16	0.80	0.82	0.84	0.88
GAENE 12	0.30	0.84	0.73	0.86	0.78
GAENE 13	1.10	*1.51*	*2.24*	1.46	*2.10*
GAENE 14	− 0.57	0.62	0.61	0.70	0.73

Unweighted MNSQ is more sensitive to outliers, but weighted MNSQ has more of an impact on instrument functioning. Underlined numbers indicate those outside of the most conservative range (0.7–1.3) but within the less conservative range (0.5–1.5) of MNSQ values. Italic numbers indicate those outside of any acceptable MNSQ range and are considered to have poor fit

Table 3 Item and person separation reliabilities for the GAENE

	Pre-course		Post-course	
	EAP/ PV item separation	WLE person separation	EAP/ PV item separation	WLE person separation
GAENE (2–14)	0.90	0.88	0.89	0.86

Table 4 Summary of item functioning for the GAENE

Item	Item fit	Rating scale separation
GAENE 2	Acceptable	Poor in post[a]
GAENE 3	Acceptable	Poor in post[a]
GAENE 4	Acceptable	Poor in pre and post[a]
GAENE 5	Acceptable	Acceptable
GAENE 6	Acceptable	Poor in post[a]
GAENE 7	*Borderline*	*Poor in pre and post*
GAENE 8	Acceptable	Poor in post[a]
GAENE 9	*Acceptable*	*Poor in post*
GAENE 10	Acceptable	Poor in post[a]
GAENE 11	Acceptable	Poor in post[a]
GAENE 12	Acceptable	Poor in post[a]
GAENE 13	*Not acceptable*	*Poor in pre and post*
GAENE 14	Acceptable	Poor in post[a]

[a] Items for which poor separation coincided with a low frequency of responses. Italic items show patterns indicative of problems with their ability to reliably measure evolution acceptance.

these items (particularly 7 and 13) failed to consistently and meaningfully separate a large number of participants based on their evolutionary acceptance measures. Furthermore, like most of the items in the post-course survey, items 7, 9, and 13 displayed evidence of rating scale disorder at the end of the semester (Fig. 2f). However, although rating scale disorder for the other items was associated with a low frequency of responses, this was not the case for items 7, 9, and 13. Specifically, for these items, 19–27.8% of the respondents selected answer options with disordered Rasch-Andrich thresholds, indicating that the rating scale functioned poorly for a large fraction of the population. Items 7 and 13 had post-course fit statistics that were outside of the most conservative range of acceptable values (Table 2). Item 13's fit statistics were also outside of the less conservative range, indicating it had a larger than expected amount of unmodeled variation (Wright and Linacre 1994).

RQ2

Controlling for all student demographic and background variables, raw and Rasch GAENE measures increased significantly from the pre- to the post-course (Raw: $b = 2.44$, $df = 739$, $t = 4.38$, $p < 0.001$; Rasch: $b = 0.68$, $df = 739$, $t = 7.33$, $p < 0.001$) (Fig. 1) (see Table 6 for a summary). The $\eta^2 G$ between instruction and GAENE measures was small (Raw: $\eta^2 G = 0.02$, $p < 0.001$; Rasch: $\eta^2 G = 0.03$, $p < 0.001$) (Fig. 3). This same model revealed that acceptance of evolution did not vary significantly across semesters.

Table 5 Text for items that show evidence of problematic item functioning

Item #	Item text
Item 7	I would be willing to argue in favor of evolution[a] in a public forum such as a school club, church group, or meeting of public school parents
Item 9	Nothing in biology makes sense without evolution
Item 13	Evolution is a scientific fact

[a] Item 7 in Smith et al. (2018) used the word "evolutionary", which we consider to be a typo. We modified it to "evolution" here and in the survey administered to students

Table 6 Summary of regression results for the GAENE and the two dimensions of the MATE

Variable	GAENE	MATE facts	MATE credibility
Pre-survey			
Gender	M > F	M > F	M > F

Race	White > URM;
	White > Asian
		White > URM;
		White = > Asian
			White > URM;
			White > Asian

Major	n.s.	n.s.	n.s.
Previous bio courses	n.s.	n.s.	n.s.
Gains			
Instruction (pre to post)	Pre < Post	Pre < Post	Pre < Post
Gender	n.s.	n.s.	n.s.
Race	n.s.	White > URM	n.s.
Major	n.s.	n.s.	n.s.
Previous bio courses	n.s.	Bio-STEM > non-STEM[a]	n.s.

Critical p-value set at 0.016

n.s. indicates not significant

[a] Only raw scores were significant

RQ3

The demographic and background variables explained between 8.1 and 8.8% of the variation in pre-course GAENE measures for raw and Rasch data, respectively (Raw: F(21,717) = 4.09, $p < 0.001$; Rasch: F(21,717) = 4.39, $p < 0.001$). Controlling for these variables, males had a significantly higher evolution acceptance than females in the pre-course (Raw: $b = 1.97$, df = 717, t = 4.32, $p < 0.001$; Rasch: $b = 0.59$, df = 717, t = 4.24, $p < 0.001$) (Table 6). The unique variance explained by gender was small (Raw: $\eta^2 G = 0.02$, $p < 0.001$; Rasch: $\eta^2 G = 0.02$, $p < 0.001$; Cohen's d: 0.22) (Fig. 4a, b). When controlling for pre-course measures as well, males and females did not differ significantly in their post-course measures, indicating that they had a similar magnitude of gains in acceptance associated with evolution instruction (Fig. 4a, b).

Again controlling for demographic and background variables, White respondents had a significantly higher evolution acceptance than Asian and URM respondents in the pre-course sample (Raw: *b*Asian vs. White = 1.85, t = 3.25, *b*URM vs. White = 2.87, df = 717, t = 4.66, $p < 0.001$; Rasch: *b*Asian vs. White = 0.68, df = 717, t = 3.91, *b*URM vs. White = 0.89, df = 717, t = 4.78, $p < 0.001$) (Fig. 4c, d; Table 6). The unique variance explained by race was also small but remained the most important predictor (Raw: $\eta^2 G = 0.05$, $p < 0.001$; Rasch: $\eta^2 G = 0.05$, $p < 0.001$; Cohen's d: White vs. Asian = 0.44, White vs. URM = 0.49, Asian vs. URM = 0.07). The unique variance explained by the interaction between race and gender was not significant (Raw: $\eta^2 G = 0.002$, Rasch: $\eta^2 G = 0.002$). When controlling for pre-course measures as well, White, Asian, and URM respondents did not differ significantly in their post-course measures, indicating that a similar magnitude of evolution acceptance gains (Fig. 4c, d; Table 6). The unique variance explained by the interaction between instruction, race, and gender was not significant for any comparison.

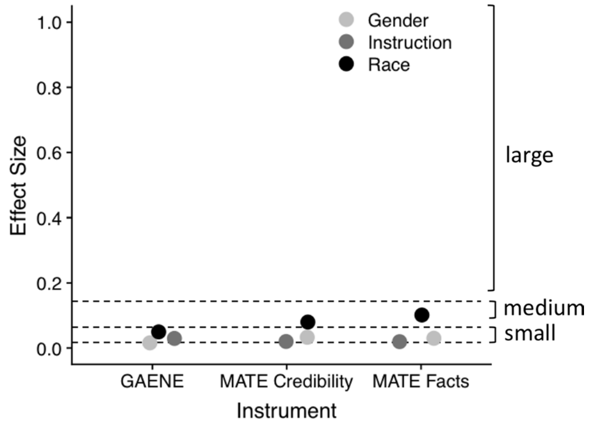

Fig. 3 Generalized Eta squared ($\eta^2 G$) for gender, race, and instruction (pre- to post-course) (range from 0 to 1). $\eta^2 G$ measures the unique variance in Rasch evolution acceptance that is explained by each variable as measured by the GAENE and the two dimensions of the MATE instruments. All $\eta^2 G$ were significant at $p < 0.016$

Surprisingly, there were no significant differences in the pre-course measures among respondents with different degree plans (Fig. 5a, b) or different histories of prior biology coursework (Fig. 5c, d) (controlling for demographic and background variables). When controlling for pre-course measures, there was no difference in post-course measures for either of these variables, indicating similar gains for respondents with different degrees plans or previous coursework (Fig. 5a–d; Table 6).

RQ4

The raw mean pre-course MATE score was 80.28 (Facts: $\bar{x}$ by-person = 41.01 ± 5.56, $\bar{x}$ by-item = 4.10 ± 0.11; Credibility: $\bar{x}$ by-person = 39.26 ± 6.07; $\bar{x}$ by

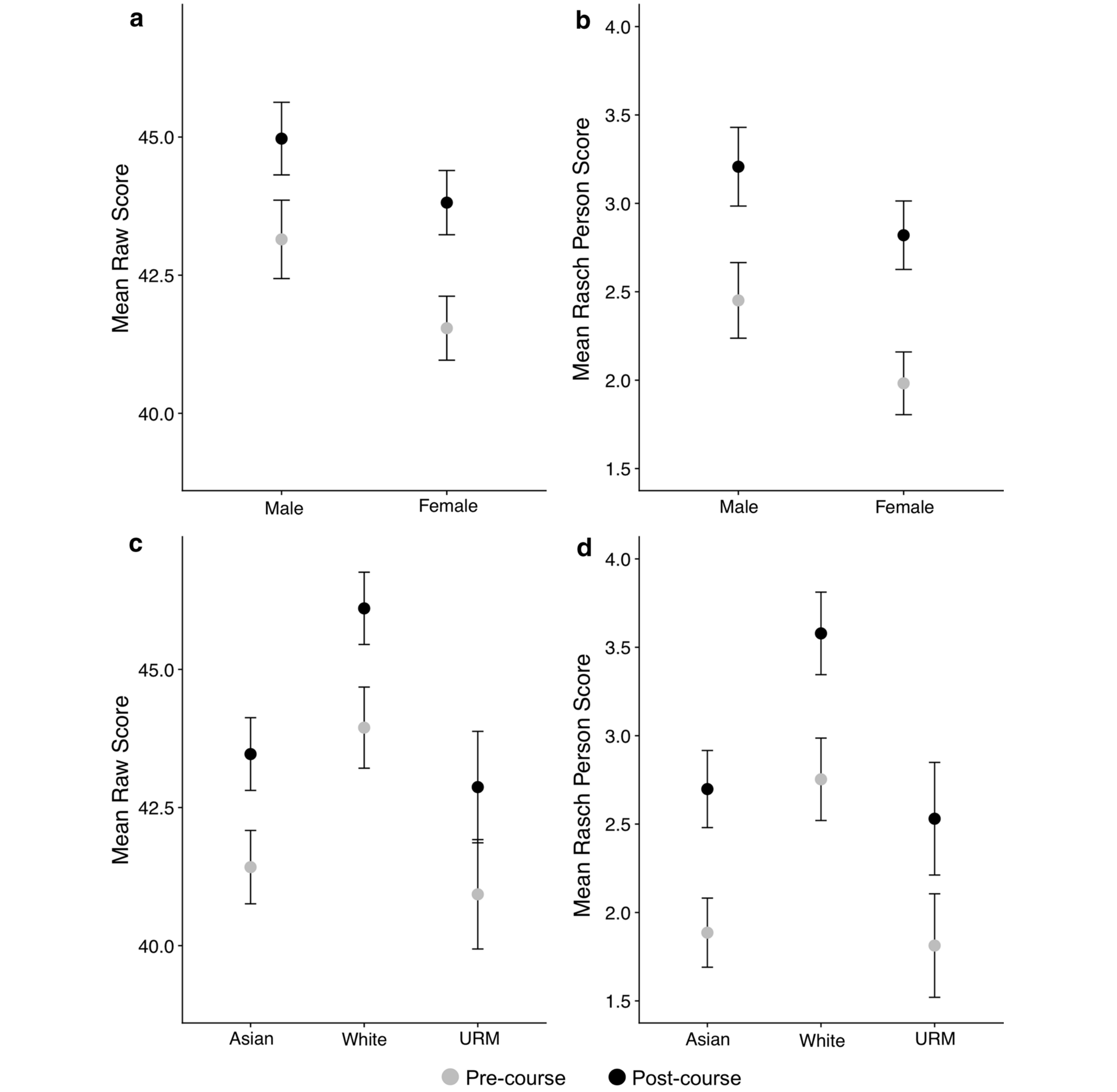

Fig. 4 Raw scores and Rasch measures for the pre- and post-course GAENE across genders (**a**, **b**) and races (**c**, **d**)

item = 3.93 ± 0.31) and the post-course mean score was 84.22 (Facts: x̄ by-person = 42.88 ± 5.28, x̄ by-item = 4.29 ± 0.07; Credibility: x̄ by-person = 41.34 ± 6.0; x̄ by item = 4.13 ± 0.21). The MATE data fit a two-dimensional model significantly better than a one-dimensional model ($\chi^2 = 58.14$, df = 2, $p < 0.001$, $AIC_{uni} = 10{,}941$[81 parameters], $AUC_{multi} = 10{,}887$[83 parameters]) and a PCA of the Rasch residuals indicated that the eigenvalues of the first contrast for each dimension was < 2.0 (Facts = 1.82; Credibility = 1.81), indicating that each item set was unidimensional. The weighted MNSQ fit statistics and the person and item reliabilities were acceptable (Additional file 1: Figure S1, Additional file 2: Figure S2).

Correlation between instruments. The GAENE and the facts dimension of the MATE were strongly correlated with one another in both the pre- and post-course. The GAENE and the credibility dimension of the MATE were moderately correlated with one another at both time points (Table 7).

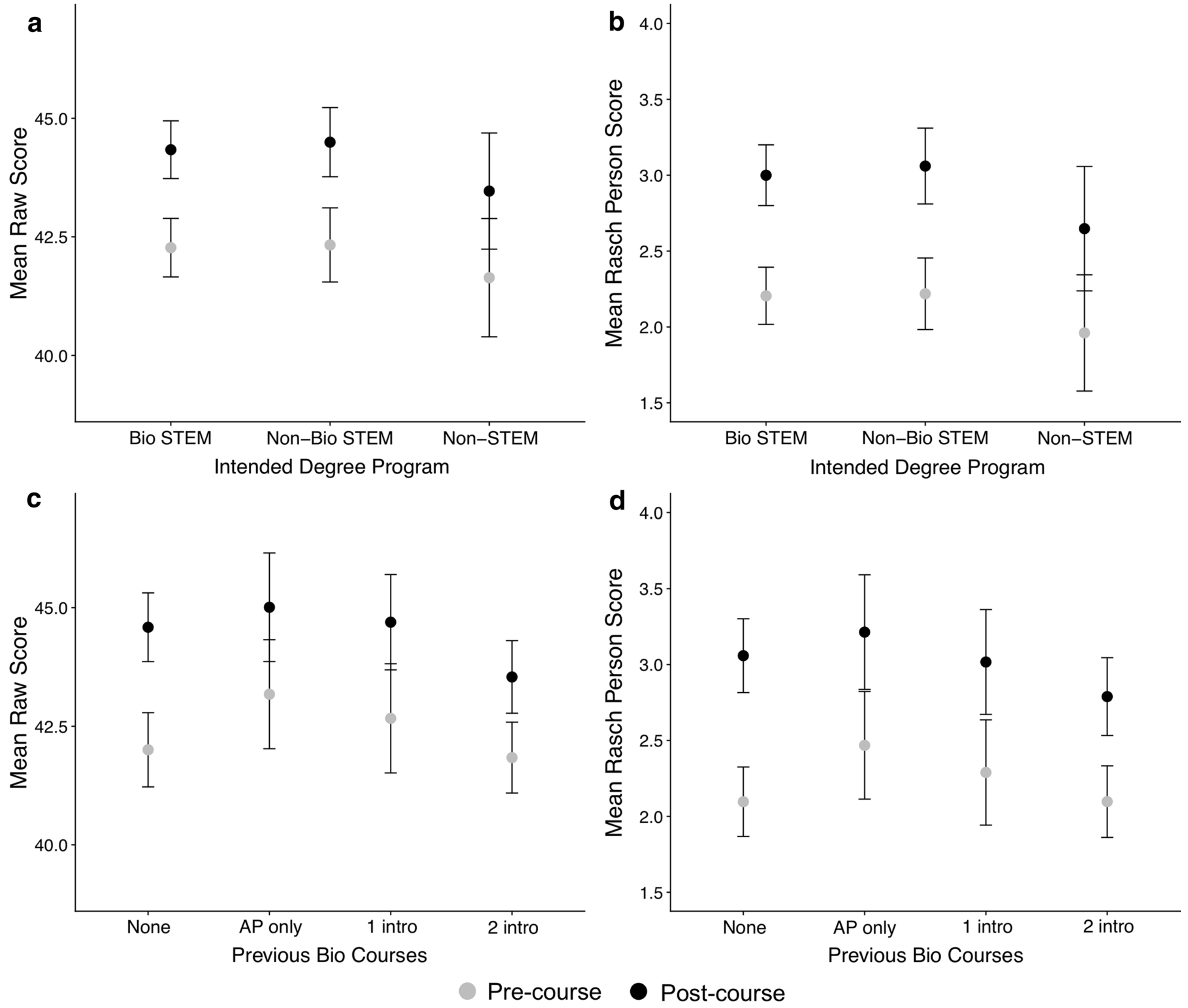

Fig. 5 Raw scores and Rasch measures for the pre- and post-course GAENE across intended degree programs (**a**, **b**) and previous biology courses (**c**, **d**)

Comparison of the effects of each variable on acceptance. As compared to the GAENE, the demographic and background variables explained nearly double the variation in pre-course MATE measures (R^2 = 18.4–19% and 15.9–19.4% for MATE facts and credibility dimensions, respectively) (Facts: Raw: F(20, 252) = 4.05, $p < 0.001$; Rasch: F(20,252) = 4.20, $p < 0.001$; Credibility: Raw: F(20, 252) = 4.28, $p < 0.001$; Rasch: F(21,252) = 3.57, $p < 0.001$).

As with GAENE measures, MATE measures increased significantly from the pre- to the post-course for the facts dimension (Raw: $b = 2.21$, df = 273, $t = 3.13$, $p < 0.001$; Rasch: $b = 1.11$, df = 273, $t = 4.16$, $p < 0.001$) and the credibility dimension (Raw: $b = 2.34$, df = 273, $t = 2.69$, $p < 0.01$; Rasch: $b = 0.93$, df = 273, $t = 4.20$, $p < 0.001$) (Table 6). The unique variance explained by instruction was small (Facts: Raw: $\eta^2G = 0.02$, $p < 0.001$; Rasch: $\eta^2G = 0.02$, $p < 0.001$; Credibility: Raw: $\eta^2G = 0.02$, $p < 0.001$; Rasch: $\eta^2G = 0.02$, $p < 0.001$) and similar for both instruments (Fig. 3).

As was the case for the GAENE, males had significantly higher pre-course MATE measures than females for the facts dimension (Raw: $b = 2.25$, df = 252, $t = 3.49$, $p < 0.001$; Rasch: $b = 0.99$, $t = 4.39$, df = 252, $p < 0.001$) and the credibility dimension (Raw: $b = 2.44$, df = 252, $t = 3.51$, $p < 0.001$; Rasch: $b = 0.62$, df = 252, $t = 3.65$, $p < 0.001$), as well as a similar magnitude of gains after evolution instruction (Table 6). The unique variance explained by gender was small (Facts: Raw: $\eta^2G = 0.02$,

Table 7 Pearson's correlation coefficients between the Rasch person measures for the GAENE and the two dimensions of the MATE

	GAENE	MATE facts	MATE Cred
GAENE	–	*0.79 (0.93)*[b] 0.82 (0.90)[c]	*0.68 (0.80)*[b] 0.77 (0.87)[c]
MATE Facts	*0.78 (0.89)*[a]	–	*0.80 (0.96)*[b] 0.80 (0.92)[c]
MATE Credibility	*0.63 (0.74)*[a]	*0.83 (0.98)*[a]	–

Italic values indicate correlation coefficients from the present study

[a] Pre-course correlation coefficients. Significant at a p-value of < 0.001

[b] Post-course correlation coefficients. Significant at a p-value of < 0.001

[c] Correlation coefficients reported in, or derived from Metzger et al. (2018). Parenthesis indicate disattenuated correlation coefficients

$p < 0.01$; Rasch: $\eta^2G = 0.03$, $p < 0.001$; Credibility: Raw: $\eta^2G = 0.02$, $p < 0.01$; Rasch: $\eta^2G = 0.03$, $p < 0.001$) and similar for both instruments (Fig. 3).

As with the GAENE, both dimensions of the MATE showed that White respondents had significantly higher pre-course MATE measures than URM respondents (Facts raw: bURM vs. White = 2.66, df = 252, t = 2.98, $p < 0.01$; Facts Rasch: bURM vs. White = 0.84, df = 252, t = 2.67, $p < 0.01$; Credibility raw: n.s.; Credibility Rasch: bURM vs. White = 0.58 df = 252, t = 2.48, $p < 0.016$). Conversely, while White respondents also had significantly higher pre-course MATE measures than Asian respondents for the Credibility dimension (Raw: n.s.; Rasch: bAsian vs. White = 0.53, df = 252, t = 2.55, $p < 0.016$), they did not differ significantly for the facts dimension (Table 6). As with the GAENE, the gains in MATE measures from pre- to post-course were equivalent across races for the credibility dimension. However, for the facts dimension of the MATE, White respondents had significantly higher pre-to post-course gains compared to URM respondents (Raw: n.s.; Rasch: bURM vs. White = 0.64, df = 251, t = 2.53, $p < 0.016$) (Table 6). The unique variance explained by race was medium for the MATE facts dimension (Raw: $\eta^2G = 0.09$, $p < 0.001$; Rasch: $\eta^2G = 0.08$, $p < 0.001$) and the MATE credibility dimension (Raw: $\eta^2G = 0.11$, $p < 0.001$; Rasch: $\eta^2G = 0.110$, $p < 0.001$), and about three times as large as compared to the GAENE (Fig. 5). The unique variance explained by the interaction between instruction, race, and gender was not significant for any comparison in either dimension.

As we found using GAENE measures, degree plan and the number of previous biology courses were not associated with significant differences in MATE measures. The one exception (from the raw data) was that bio-STEM respondents had significantly higher raw pre-course MATE scores for the facts dimension than did non-STEM respondents (Raw: $b = 2.39$, df = 252, t = 2.45, $p < 0.016$; Rasch: n.s.) (Table 6). All other comparisons among respondents with different degree plans and different numbers of previous biology courses, had similar pre-course MATE measures and similar pre- to post-course gains (Table 6).

Discussion

GAENE fit and function

The GAENE has been administered and the results published in three studies (i.e., Metzger et al. 2018; Rachmatullah et al. 2018; Smith et al. 2016). The raw scores reported in the present study are the highest mean levels of evolution acceptance described in undergraduate students using this instrument (see Table 8; Note that Rachmatullah et al. studied pre-service teachers in Indonesia). Studies in more populations across the US are needed in order to provide evidence in support of the generalizability of the inferences produced by the GAENE (cf. Campbell and Nehm 2013; Messick 1995). Moreover, given that significant demographic impacts have been documented in two different studies, it is also important that the demographic composition of the study sample be described and examined (Table 8). Notably, this was not addressed in the original GAENE study (Smith et al. 2016).

There were consistent patterns in the psychometric properties of the GAENE across the two prior studies with American undergraduates and the present study (Table 8). Specifically, the instrument was found to be one-dimensional, the item and person reliabilities were acceptable, and the items generally fit the Rasch model well. The Wright map demonstrated that the items were generally easy to agree with for most respondents, and those items that were most difficult to agree with were consistently difficult across studies (i.e., items 7, 9, and 13).

There were several inconsistencies across studies (Table 8). First, we found that item 13 had fit values well above the acceptable range in the post-course survey, indicating that it underfit the model after instruction. Although Smith et al. (2016) reported acceptable fit for this item, they reported it for a single time point and not in the context of a relevant biology course. In fact, their reported fit statistics for this item (infit: 1.43; outfit: 1.42) are similar to the pre-course fit statistics that we report (infit: 1.46; outfit: 1.51). In our study, *post-course* GAENE measures demonstrated model underfit for item 13. However, Smith et al. did report other potential problems with this item. Specifically, they found significant differential item functioning (DIF) between high school and undergraduate students, indicating that the item might be influenced by different levels of knowledge (Smith et al. 2016), which may be problematic because

Table 8 Summary of GAENE studies on undergraduate students and recommendations for future work

	Smith et al.	Metzger et al.	This study	Recommendation
Sample characteristic				
Sample size	155 High school; 516 undergraduate	105 Undergraduate	770 Undergraduate	None
Geographic region	Across U.S.	Midwestern U.S.	Northeastern U.S.	More studies in other geographic regions
Demographics	–	80% female; 37% URM	57% female; 22% URM	More studies reporting and analyzing demographics
Mean GAENE scores	By-item (5 max): 3.78 i.e. 3.02/4	By-person (65 max): 51.7 [post] i.e. 41.36/52	By-item (4 max): 3.25 [pre], 3.41 [post] By-person (52 max): 42.22 [pre], 44.3 [post]	Report by-item and by-person scores for ease of comparison across studies
Analysis				
Dimensionality (Internal structure validity)	One-dimensional	One-dimensional	One-Dimensional	None
Item fit	Items 2–14 acceptable at single time-point	Items 2–14 acceptable in post-survey; Pre-survey not present	Items 2–14 acceptable in pre-survey; Item 13 poor fit in post-survey	Investigate pre- post- dynamics of item 13, esp. in course with NOS unit. Remove or modify item
Rating scale functioning (RSF) of items	GAENE 2.1: Unclear how per-item RSF assessed. 7 and 13 had noise	GAENE 2.1: Per-item RSF not assessed	GAENE 1.0: RSF poor for items 7, 9, and 13	Investigate RSF for each item. Remove or modify items 7, 9, and 13
Overall reliabilities	Acceptable Person: 0.91 Item: 0.99	Acceptable Person: 0.93 Item: 0.86	Acceptable Person: 0.86–0.88 Item: 0.89–0.9	None
Wright map	Large gap at high end	–	Large gap at high end	Add more difficult items
Impact of instruction	–	–	Pre < post	More pre- post- data in large samples
Impact of gender and race	–	URM < Non-URM F = M No effect size reported	URM < White Asian < White F < M $\eta^2 G = 0.02–0.05$	More studies with gender and race, and effect sizes for each
Impact of degree plan	–	–	Bio STEM = Non-Bio-STEM = Non-STEM	More studies of degree plan to corroborate findings
Variance explained (R^2) by all modeled variables	–	Post-course R^2 ~ 11%	Pre-course R^2 ~ 9%	More reporting of variance explained by student variables
External structure validity	–	High correlation with both dimensions of the MATE	High correlation with MATE facts; moderate correlation with MATE credibility	Correlate GAENE with other acceptance instruments like the I-SEA and MATE

the instrument was designed to measure acceptance only. We have related concerns with GAENE item 13. Specifically, it is possible that instruction in the course did not align with the expected normative answer. As part of our nature of science unit, we teach that evolution is both a pattern (e.g., observation, fact) and a process (e.g., explanation, theory). Therefore, item 13's assertion that "evolution is a scientific fact" could have confused students given that evolution was discussed in the course as both a pattern *and* a process. Finally, it is not clear if experts would provide the expected normative answer for item 13. The US National Academy of Sciences, for example, and many textbooks refer to evolution as a theory (http://www.nas.org, Futuyma and Kirkpatrick 2018). Clearly, further investigations of the pre- to post-course dynamics of item 13, especially in courses that contain NOS instruction, are needed to corroborate our explanation for these item response patterns (Table 8).

Our analysis of the functioning of the GAENE included an item-level assessment of the rating scale. We found that while the overall person and item reliabilities were acceptable, the rating scale functioned poorly for three items: 7, 9, and 13. These items had a poor correspondence between respondents' answer choices and their overall Rasch person measures in the pre- and post-course survey, and they displayed rating scale disorder in the post-survey. These patterns suggest that the items failed to consistently and meaningfully separate participants based on their levels of evolutionary acceptance. The finding that overall reliabilities were acceptable but some individual items had rating scale issues highlights the importance of a clear item-level analysis of rating scale functioning. It is not clear how or if Smith et al. (2016) analyzed the rating scale of each GAENE item; these authors did report that "Items 7 and 13 exhibit[ed] slightly more noise in their response patterns than would be expected and will be examined in subsequent administrations of the scale" (Smith et al. 2016, p. 17). Therefore, even though we used a slightly different rating scale (GAENE 1.0) than Smith et al. (GAENE 2.1), both scales uncovered similar rating scale concerns for similar items (Table 8).

It is notable that items 7 and 9 had acceptable fit statistics even though they displayed rating scale anomalies that were not accounted for by low response frequencies of the relevant answer options. We have not generated evidence to explore the causes of these rating scale anomalies, but we hypothesize that these two items may introduce construct-irrelevant variation. Item 7 states, "I would be willing to argue in favor of evolution in a public forum such as a school club, church group, or meeting of public school parents" (Smith et al. 2016, p. 16). This question may capture latent traits beyond evolution acceptance, such as a willingness to engage in argumentative acts in public settings. Item 9 states, "Nothing in biology makes sense without evolution," which may trigger a test-taking behavior that some students utilize when engaging in multiple-choice tests. Specifically, students are often advised to take note of all-or-nothing language (e.g., "always", "nothing", "never", "only") in test-preparation guides (e.g., The Pennsylvania State University 2017). Interviews with students and experts will help to elucidate the causes of the problematic rating scales for these items. Overall, our analyses of the fit and rating scale functioning of the GAENE generated comparable results to those of Smith et al. (2016), including the finding that some of the same items displayed psychometric limitations. Therefore, we recommend that items 7, 9, and 13 be modified or removed from the instrument (Table 8).

Race and gender

Understanding the roles that race and gender play in STEM educational outcomes has emerged a major research topic (e.g., Gender: Creech and Sweeder 2012; Lauer et al. 2013; Willoughby and Metz 2009; Wright et al. 2016; Race: Creech and Sweeder 2012; Ma and Liu 2015; Nehm and Schonfeld 2008). STEM fields continue to suffer from a substantial lack of diversity compared to the overall population (PCAST 2012). The roles of race and gender on acceptance of evolution and its possible impacts on attrition in STEM fields has rarely been explored in the literature. We report that all of the demographic and background variables that we included in our model explained up to 9% of the variation in pre-course, Rasch-converted GAENE measures. Male and White respondents had the highest GAENE measures in our population, which corroborates findings by Metzger et al. (2018) using this instrument in a Midwestern sample (Table 8). The magnitude of the unique variation in GAENE measures that can be explained by gender and race was small, but importantly, larger than the variation explained by instruction.

We also measured evolution acceptance using the MATE. The pre- and post-course MATE raw scores reported here are among the highest reported for any student population (Metzger et al. 2018, Table 5; Rachmatullah et al. 2018, p. 348–349). For example, undergraduate health science students in the Midwestern US had a pre-course GAENE score of 78.68 and a post-course score of 81.72 (Metzger et al. 2018, Table 5). Like the GAENE, MATE scores increased from the pre- to the pre-course, and White and male respondents had the highest evolution acceptance. However, the size of the effect of race was nearly three times as large for both dimensions of the MATE as compared to the GAENE. In fact, White students not only had higher baseline scores, but they also had higher gains from pre- to post-course than URM students for the MATE facts dimension. Furthermore, the entire model, which included all student demographic and background variables, explained almost double the variation in MATE measures (for the facts and credibility dimensions) as compared to GAENE measures. These patterns provide some convergent evidence for the contributions of gender and race to evolution acceptance measures (Table 8), but it is unclear if the differences in the impact of race reflect meaningful distinctions in the operation of the instrument. For example, it is possible that assessing evolution acceptance in the presence of a specified context or scale (as is the case with the MATE) may generate different response patterns among students than when it is assessed in a generalized format (as is the case with the GAENE). More research is needed to better understand the impact of demographic and background variables on evolution acceptance measures.

Degree plan and previous biology courses

Surprisingly, using both the GAENE and the MATE, we did not find significant differences in evolution acceptance using Rasch measures among respondents with different degree plans or among those with different histories of prior biology coursework (Table 8). Other studies have shown that biology majors and non-majors did not differ substantially in other metrics of STEM ability including evolution misconceptions (Nehm and Reilly 2007), exam grades, and overall course performance (Sundberg and Dini 1993). More studies on the roles of degree plan and previous coursework are necessary in order to corroborate these findings (Table 8). However, this finding adds to a growing body of work questioning the impact of biology knowledge on evolution acceptance (Ha et al. 2012).

Assessing evolution acceptance

The GAENE was developed to address the purported limitations of other evolution acceptance instruments, including the MATE. However, although it appears to have some significant limitations (see Romine et al. 2017; Smith et al. 2016), the MATE remains the most commonly used acceptance measure, appearing in dozens of peer-reviewed studies. Surprisingly, the authors of the GAENE did not analyze how their new and improved instrument compared to the MATE or discuss if the use of the new instrument would lead to different conclusions about the patterns of evolution acceptance in a population. We report that the GAENE and MATE generate similar patterns of pre-course evolution acceptance and we recommend that when reporting raw data, authors include both the by-item and by-student statistics for ease of comparison across studies (Table 8). We also report that both instruments displayed similar magnitudes of acceptance change in response to instruction, and in terms of the impact of certain student variables on this trait. However, demographic and background variables predicted almost double the variation in MATE measures as compared to GAENE measures, and the magnitude of the impact of race may differ between the instruments. Furthermore, while the Rasch measures for the GAENE and the MATE facts dimension were strongly correlated, the GAENE was only moderately correlated with the MATE credibility dimension.

Our study suggests that overall measures of acceptance change will be similar using the MATE or the GAENE in most cases. Therefore, if a researcher's goal is to measure overall levels of acceptance, or acceptance change through time, then both instruments may lead to similar conclusions. Although we report some differences in the impact of demographic variables, this is one of only a few studies to do so, and it is unclear if these patterns will generalize to other populations, especially those with lower evolution acceptance. Few studies have assessed the effect of race or gender on evolution acceptance and even fewer have estimated the magnitude of this effect using statistics that are comparable across studies. We report effect sizes using generalized eta squared ($\eta^2 G$) in a repeated-measures design, which both accounts for the non-independence of pre- to post-course testing, and permits appropriate comparisons across studies, including in meta-analyses (Lakens 2013). However, because of the lack of comparable data reported in the literature, it is difficult to interpret the effect sizes of race and gender on many outcome variables (comparisons of effect sizes is the preferred method of interpreting the magnitude of an effect; Lakens 2013). A more consistent reporting of appropriate and comparable effect sizes is needed to best diagnose the magnitude of the effect of these variables (Table 8). Furthermore, more studies that address the roles of race and gender on evolution acceptance using the GAENE and other instruments such as the MATE and the I-SEA would help determine if the patterns identified here are generalizable across populations, and if the differences in the instruments are meaningful or if they are evidence of psychometric or conceptual limitations (Table 8).

Limitations

It is critical to establish robust measures of latent traits that can be utilized consistently across populations (NRC 2001). Although our study is an important step in evaluating the relative quality of two evolution acceptance instruments, our work alone cannot be used to determine whether the MATE or the GAENE are "better" measurement tools. There are several reasons for this claim. First, the theoretical rationale for how to measure evolution acceptance and the practical application of that theory in the form of an appropriate measurement instrument is still in its infancy. Several authors have argued that the definition of evolution acceptance must distinguish it from evolutionary knowledge, belief, and understanding (Smith and Siegel 2004; Wagler and Wagler 2013), which is one of the major criticisms of the MATE (Smith et al. 2016). However, others have suggested that the belief that something is true is an essential component of acceptance (Ha et al. 2012). More recently, McCain and Kampourakis (2016) pointed out the distinction between "belief in" evolution (i.e., valuing its unifying and explanatory power) versus "belief about" evolution (i.e., accepting that it is true). Some authors also argue that the definition should address the distinct scales and contexts it is hypothesized to encompass (e.g., biological diversity, micro- and macroevolution; see Nadelson and Southerland 2012; Nehm and Ha 2011; Nehm 2018). The authors of the GAENE put forth one of the few formal definitions of *generalized* evolution acceptance, which they define as "the mental act or policy of deeming, positing, or postulating that the current theory of evolution is the best current available scientific explanation of the origin of new species from preexisting species" (Smith et al.

2016, p. 8). However, given that the instrument was only proposed recently, the authors' theoretical conceptualization of acceptance has not been robustly evaluated by the science education community. Indeed, the definition is notable for its singular focus on *macroevolutionary* phenomena (i.e., speciation) despite the fact that most of the items are not specifically about this mode of evolution.

Second, there are many criteria for evaluating the degree to which evidence supports inferences drawn from instrument scores (Campbell and Nehm 2013; Messick 1989; Messick 1995). Our study addressed several of the criteria including internal structure evidence (dimensionality), external structure evidence (correlations with other instruments), and generalization evidence across some contexts (student background and demographic variables). However, these analyses were conducted in only one population, and as such, cannot lead to generalizable inferences or well-informed actions. As emphasized by many authors, validity is not a property of an instrument, but rather a property of the inferences derived from these instruments and the actions those inferences entail (Messick 1992, 1995). Messick (1992, 1995) has described validation as a *continuing* process marked by *degrees* of validity, as opposed to an all or nothing designation. Reaching the highest degree of validation will require the determination that several forms of validity evidence are consistent with one another as well as with our inferences (cf. Messick 1992). Therefore, although the inference that MATE and GAENE scores reflect magnitudes of evolution acceptance is supported by psychometric evidence, there is still much work to be done. At present, more studies are needed that address the patterns and functioning of these instruments across populations, especially using pre- to post-course study designs that generate large, replicated data sets and include the reporting of appropriate effect sizes (Table 8).

A consensus on the quality and meaning of the measures generated from these instruments does not exist and any conclusions about which instrument is superior for the measurement of evolution acceptance are premature. Despite this, Metzger et al. (2018) claimed that the GAENE was better than the MATE for measuring evolution acceptance in their population because it displayed less measurement error. However, because the theoretical constructs used to operationalize evolution acceptance remain open to criticism (see above), using measurement error or other psychometric qualities alone is insufficient to support claims about the best way to measure this trait.

Although we report effect sizes for various demographic and background variables on evolution acceptance, questions remain about how these variables impact evolution acceptance, which in turn limits the inferences that can be drawn from GAENE and MATE scores. Gathering further evidence from DIF studies, substantive validity studies, and ethnographic research will be needed. Finally, our study was not designed a priori to test for the impacts of demographic and background variables on evolution acceptance. Future studies should be designed with this goal in mind, and generate a more balanced data set across racial categories, and collect information on additional, relevant variables (e.g., socioeconomic status, parental education level, and religiosity).[1]

Additional files

Additional file 1. Item and person separation reliabilities for the MATE facts and MATE credibility dimensions.

Additional file 2. Item difficulties, and weighted (infit) and unweighted (outfit) MNSQ fit statistics for the MATE facts and MATE credibility dimensions.

Authors' contributions

Both authors designed the research questions and collected the data. GS performed all of the psychometric analyses, drafted the figures, and wrote the results. Both authors contributed to the writing of the final manuscript. Both authors read and approved the final manuscript.

Author details

[1] Department of Ecology and Evolution, Stony Brook University, Stony Brook, USA. [2] Institute for STEM education and Department of Ecology and Evolution, Stony Brook University, Stony Brook, USA.

Acknowledgements

We acknowledge John True for allowing access to the class in which this study took place. We also acknowledge Bill Boone, Xiufeng Liu, and Knut Neumann for their insights on Rasch analysis. We acknowledge William Romine for his comments on the manuscript and suggestions for improvement, especially regarding the interpretation of the results. Finally, we acknowledge an anonymous reviewer for their thoughtful comments.

Competing interests

The authors declare they have no competing interests.

Availability of data and materials

Available from authors upon request.

Consent to publish

We have received consent from all study participants to collect and publish the data they shared with us.

Ethics approval and consent to participate

Our study complies with the ethical requirements of the Institutional Review Board at Stony Brook University (Protocol #504271). We have received consent for participation from all participants of this study.

Funding

G.S. was funded by a National Science Foundation TUES Grant (#1322872). Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. R.H.N. was funded by a Howard Hughes Medical Institute Science Education Grant (Inclusive Excellence program).

[1] We thank William Romine for this helpful idea.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 7 August 2018 Accepted: 26 November 2018
Published online: 27 December 2018

References

Adams RJ, Wu ML, Wilson M. The Rasch rating model and the disordered threshold controversy. Educ Psychol Meas. 2012;72(4):547–73.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME). The standards for educational and psychological testing. Washington, DC: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME); 2014.

Andrich D. An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any "threshold disorder controversy". Educ Psychol Meas. 2013;73(1):78–124.

Andrich D, De Jong JHAL, Sheridan BE. Diagnostic opportunities with the Rasch model for ordered response categories. In: Rost J, Lange Heine R, editors. Applications of latent trait and latent class models in the social sciences. Munster: Waxman, Verlag Gmbh; 1997. p. 59–70.

Bakeman R. Recommended effect size statistics for repeated measures designs. Behav Res Methods. 2005;37(3):379–84.

Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. New Jersey: Lawrence Erlbaum Associates, Inc.; 2001.

Boone WJ, Staver JR, Yale MS. Rasch analysis in the human sciences. Dordrecht: Springer; 2014.

Brewer CA, Smith D. Vision and Change in undergraduate biology education: a call for action. Washington, DC: Directorate for Biological Sciences, American Association for the Advancement of Science; 2011.

Campbell CE, Nehm RH. A critical analysis of assessment quality in genomics and bioinformatics education research. CBE Life Sci Educ. 2013;12(3):530–41.

Cohen J. Statistical power analysis for the behavioral sciences. New York: Routledge Academic; 1988.

Creech LR, Sweeder RD. Analysis of student performance in large-enrollment life science courses. CBE Life Sci Educ. 2012;11(4):386–91.

Futuyma DJ, Kirkpatrick M. Evolution. 4th ed. Oxford: Oxford University Press; 2018.

Grigg K, Manderson L. The Australian racism, acceptance, and cultural-ethnocentrism scale (RACES): item response theory findings. Int J Equity Health. 2016;15:49. https://doi.org/10.1186/s12939-016-0338-4.

Ha M, Haury DL, Nehm RH. Feeling of certainty: uncovering a missing link between knowledge and acceptance of evolution. J Res Sci Teach. 2012;49:95–121.

Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol. 2013;4(863):1–12.

Lauer S, Momsen J, Offerdahl E, Kryjevskaia M, Christensen W, Montplaisir L. Stereotyped: investigating gender in introductory science courses. CBE Life Sci Educ. 2013;12(1):30–8.

Levine RT, Hullett CR. Eta squared, partial eta squared, and misreporting of effect size in communication research. Hum Commun Res. 2002;28(4):612–25.

Linacre JM. Category disordering (disordered categories) vs. threshold disordering (disordered thresholds). In: Rasch Measurement Transactions. Institute for Objective Measurement. 1999. https://www.rasch.org/rmtbooks.htm. Accessed 6 Nov 2018.

Ma Y, Liu Y. Race and STEM degree attainment. Soc Compass. 2015;9(7):609–18.

McCain K, Kampourakis K. Which questions do polls about evolution and belief really ask, and why does it matter? Public Underst Sci. 2016;27(1):2–10.

Messick S. Validity. In: Linn RL, editor. Educational measurement (3rd ed.). New York: Macmillan; 1989. p. 13–103.

Messick S. Validity of test interpretation and use. In: Linn RL, editor. Encyclopedia of educational research (6th ed.). New York: MacMillan; 1992. p. 1487–95.

Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50(9):741–9.

Metzger K, Montplaisir D, Haines D, Nickodem K. Investigating undergraduate health sciences students' acceptance of evolution using MATE and GAENE. Evol Educ Outreach. 2018;11:10. https://doi.org/10.1186/s12052-018-0084-8.

Nadelson LS, Southerland S. A more fine-grained measure of student's acceptance of evolution: development of the inventory of student evolution acceptance—I-SEA. Int J Sci Educ. 2012;34(11):1637–66.

National Research Council (NRC). Knowing what students know. Washington, D.C: National Academies Press; 2001.

Nehm RH. Evolution. In: Reiss M, Kampourakis K (eds). Teaching biology in schools, Chap 14. New York: Routledge; 2018.

Nehm RH, Ha M. Item feature effects in evolution assessment. J Res Sci Teach. 2011;48(3):237–56.

Nehm RH, Reilly L. Biology majors' knowledge and misconceptions of natural selection education. Bioscience. 2007;57(3):263–72.

Nehm RH, Schonfeld IS. Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. J Res Sci Teach. 2008;45(10):1131–60.

Nijsten T, Sampogna F, Chren M, Abeni D. Testing and reducing Skindex-29 using Rasch analysis: Skindex-17. J Investig Dermatol. 2006;126(6):1244–50.

Olejnik S, Algina J. Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol Methods. 2003;8(4):434–47.

Pew Research Center. Chapter 4: evolution and perceptions of scientific consensus. American, Politics and Science Issues. July 1, 2015. http://www.pewinternet.org/2015/07/01/chapter-4-evolution-and-perceptions-of-scientific-consensus/. Accessed 14 Nov 2018.

President's Council of Advisors on Science and Technology. Engage to excel: producing one million additional college graduates with degrees in science, technology, engineering and mathematics. 2012. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf. Accessed 20 Feb 2018.

Rachmatullah A, Nehm RH, Ha M, Roshayanti F. Evolution education in Indonesia: pre-service biology teachers' evolutionary knowledge levels, reasoning models, and acceptance patterns. In: Deniz H, Borgerding L, editors. Evolution education around the globe. Dordrecht: Springer; 2018. p. 335–55.

Robitzsch A, Kiefer T, Wu M. Test analysis modules (TAM). v. 2.10-24; 2018.

Romine WL, Walter E, Todd A. Understanding patterns of evolution acceptance—a new implementation of the measure of acceptance of the theory of evolution. J Res Sci Teach. 2017;54(5):642–71.

Rutledge ML, Warden MA. Development and validation of the measure of acceptance of the theory of evolution instrument. Sch Sci Math. 1999;99(1):13–8.

Sbeglia GC, Nehm RH. Do you see what I-SEA? A Rasch analysis of the psychometric properties of the inventory of student evolution acceptance. Sci Educ (**in press**).

Singmann H, Bolker B, Westfall J, Aust F, Højsgaard S, Fox F, et al. Analysis of factorial experiments (afex). V. 0.21-2; 2018.

Smith MU, Siegel H. Knowing, believing, and understanding: what goals for science education? Sci Educ. 2004;13(6):553–82.

Smith MU, Snyder SW, Devereaux R. The GAENE—generalized acceptance of evolution evaluation: development of a new measure of evolution acceptance. J Res Sci Teach. 2016;53(9):1289–315.

Sundberg MD, Dini ML. Science majors vs. nonmajors: is there a difference? J Coll Sci Teach. 1993;22(5):299–304.

The Pennsylvania State University. Test-taking tips. In: Penn state learning. Office of Undergraduate Education. 2017. https://pennstatelearning.psu.edu/test-taking-tips. Accessed 14 Nov 2018.

Wagler A, Wagler R. Addressing the lack of measurement invariance for the measure of acceptance of the theory of evolution. Int J Sci Educ. 2013;35(13):2278–98.

Willoughby SD, Metz A. Exploring gender differences with different gain calculations in astronomy and biology. Am J Phys. 2009;77(7):651–7.

Wilson M. Constructing measures: an item response modeling approach. Mahwah: Erlbaum; 2005.

Wright BD. Rack and stack: time 1 vs. time 2 or pre-test vs. post-test. Rasch Meas Trans. 2003;17(1):905–6.

Wright BD, Linacre M. Reasonable mean-square fit values. Rasch Meas Trans. 1994;8(3):370.

Wright CD, Eddy SL, Wenderoth M, Abshire E, Blankenbiller M, Brownell SE. Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. CBE Life Sci Educ. 2016. https://doi.org/10.1187/cbe.15-12-0246.

Yang Y, He P, Liu X. Validation of an instrument for measuring students' understanding of interdisciplinary science in grades 4–8 over multiple semesters: a Rasch measurement study. Int J Sci Math Educ. 2017. https://doi.org/10.1007/s10763-017-9805-7.