RESEARCH ARTICLE

Investigating undergraduate health sciences students' acceptance of evolution using MATE and GAENE

Kelsey J. Metzger[1]*📷, Darian Montplaisir[1], David Haines[1] and Kyle Nickodem[2]

Abstract

Background: Despite the overwhelming agreement among scientists regarding the fundamental importance of evolution to all areas of biology, a lack of evolution understanding and acceptance has been reported in studies of students, educators, and members of society. In the present study, we investigate and report evolution acceptance in a population of undergraduate health sciences students enrolled in a first-year foundational biology course. Two published instruments—The Measure of Acceptance of the Theory of Evolution (MATE) and the Generalized Acceptance of Evolution Evaluation (GAENE)—were used to quantify evolution acceptance. A confirmatory factor analysis (CFA) was performed on both instruments to test whether the items measured the underlying construct sufficiently. Additionally, Rasch scaling was used to investigate fit between the data and the measurement model, and to determine if the MATE should be treated as a unidimensional or bidimensional instrument. Using correlation and regression analysis, we examined the relationships between the two measures of evolution acceptance, and between measures of evolution acceptance with other student variables of interest.

Results: The health sciences students in this study demonstrated high acceptance of evolution at the start of term, as well as a significant increase in evolution acceptance from pre- to post-test. CFA and Rasch scaling provided some evidence that the MATE is a bidimensional instrument, but considering MATE as a bidimensional instrument provided little additional insight compared to treating MATE as a unidimensional instrument. Measures of evolution acceptance resulting from the MATE and GAENE instruments were significantly and strongly correlated. Multiple regression modeling identified underrepresented minority status as a demographic variable predictive of evolution acceptance, and provided further evidence of the strong association between the MATE and GAENE instruments.

Conclusions: The undergraduate health sciences students in this study demonstrated a significant increase in evolution acceptance from pre- to post-test after one semester of instruction in general biology. Measures of evolution acceptance from the MATE and GAENE instruments were strongly correlated whether MATE was treated as a unidimensional or bidimensional instrument. This work provides initial indications that the MATE and GAENE instruments perform comparably as measures of evolution acceptance. Although the instruments are closely related, this work found more psychometric evidence for interpreting and using GAENE scores than MATE scores as a measure of evolution acceptance.

Keywords: Evolution, Acceptance, Biology education, Assessment, MATE, GAENE, Undergraduate

---

*Correspondence: kmetzger@umn.edu
[1] Center for Learning Innovation, University of Minnesota Rochester, 300 University Square, 111 South Broadway, Rochester, MN 55904, USA
Full list of author information is available at the end of the article

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 2 of 18

Background

The importance of evolution education

Evolution is the unifying principle of biology (Armstrong 1929, p. 135; Dobzhansky 1973, p. 125; Mayr 1982), and, more broadly, "an essential concept for anyone who considers science to be the best way to understand the natural world" (Fishman 2008, p. 1586). Moreover, evolutionary principles are routinely and effectively leveraged for practical applications in medicine, public health, agriculture, conservation biology, natural resource management, and environmental science (Catley and Novick 2009, p. 313; Hendry et al. 2011; Novick and Catley 2012). A lack of evolution understanding and acceptance prevents informed decision making regarding biological issues that may have personal ramifications (Nadelson and Hardy 2015). As such, competence in evolutionary biology is universally recommended as a core outcome for students of biology (Marocco 2000; NRC BIO2010 2010; AAAS Vision and Change 2011; Quinn et al. 2011; NRC 2012, 2013), and informed members of our society (AAAS 1989, 2001). Emphasizing the importance of evolution understanding for all members of society, Smith (2010) states that "omitting evolution from basic instruction of our citizenry would constitute the equivalent of educational malpractice" (p. 544). Thus, an understanding of evolution at all levels of the curriculum is foundational to the discipline of biology.

Despite the overwhelming agreement among scientists regarding the centrality of evolutionary concepts to all areas of biology (Pew Research Center 2015), many students of biology have a limited understanding of evolution, as demonstrated by difficulty in (1) correctly identifying the patterns, processes, and outcomes of evolution (Mayr 1982; Clough and Driver 1986; Good et al. 1992; Scharmann and Harris 1992; Cummins et al. 1994; Anderson et al. 2002; Smith 2010), and (2) interpretation of phylogenetic relationships represented graphically (Baum et al. 2005; Meir et al. 2007; Baum and Offner 2008; Naegle 2009). Students and non-scientists also frequently employ the misconception that evolutionary processes are direct rather than emergent (Chi 2005, p. 174) and view evolution as a static entity rather than a dynamic—and generally lengthy—process (Smith 2010, p. 543). A concurrent lack of familiarity with the timescale of evolution (i.e. "deep time") likely exacerbates this disconnect (Metzger 2011). These principles and skills are part of a broader set of complex concepts that are widely considered difficult to teach, and even more difficult to teach well (Anderson 2007; Gregg et al. 2003). Furthermore, many science teachers—particularly high school science teachers—often lack the disciplinary knowledge and confidence to teach evolution effectively (Bishop and Anderson 1990; Glaze and Goldston 2015).

Acceptance of evolution

In addition to widespread difficulty in, and lack of, understanding regarding evolutionary biology, low rates of *acceptance* of evolution have been reported in the general population (Miller et al. 2006; Gallup 2016), in pre-service educators (Romine et al. 2016), high school biology educators (Moore and Kraemer 2005; Moore and Cotner 2009; Glaze and Goldston 2015), university professors (Romine et al. 2016), and in various student populations (Rice et al. 2011; Romine et al. 2016). Although belief and acceptance may be closely related, recent investigators of evolution acceptance have separated these constructs, with acceptance being more closely related to "believing that" rather than "believing in" (Smith et al. 2016, pp. 1291–1292), and that acceptance "…is more voluntary than belief, and involves a commitment to use what is accepted; belief is less voluntary, and need not be used as a basis for inference or action." (Smith et al. 2016, p. 1292). For example, a biology student may interpret phylogenetic data from an experiment using an evolutionary framework (belief that), but simultaneously hold religious or cultural views (belief in) that are kept distinct from the scientific process. Thus, it may be useful to investigate the extent to which students *accept* evolution as distinct from student *belief* in evolution.

Several specific variables have been identified as having significant associations with acceptance of evolution, although with somewhat varying consistency in the literature. Perhaps the two most commonly identified variables identified as having a significant association with evolution acceptance are religiosity (Mazur 2004; Nehm and Schonfeld 2007; Evans 2008; Moore et al. 2011; Heddy and Nadelson 2013; Yousuf et al. 2011; Barone et al. 2014; Carter and Wiles 2014; Rissler et al. 2014), and performance on biology or evolution knowledge assessments (Nadelson and Southerland 2010; Yousuf et al. 2011; Walter et al. 2013; Carter et al. 2015; Mead et al. 2017).

Other variables that have been identified as having a significant association with evolution acceptance include age (Evans 2000), gender, academic standing, college major, prior study in biology and/or philosophy (Ingram and Nelson 2006; Rutledge and Mitchell 2002), trust in science and scientists (Nadelson and Hardy 2015), attitudes toward science and technology, attitudes toward life (Miller et al. 2006), and high school biology experience (Moore and Cotner 2009). Studies investigating populations in Minnesota specifically have reported that approximately 25–30% of high school biology teachers believe that creationism has a valid scientific foundation (Moore and Kraemer 2005; Moore and Cotner 2009); 63% of high school biology teachers teach evolution and not creationism while an additional 20% of high school

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 3 of 18

biology teachers teach both evolution and creationism (Moore and Cotner 2009, p. 98). Moore and Cotner (2009) found that the nature of biology education students experience in high school significantly impacts students' later attitudes toward evolution when they are in college: students taught evolutionary theory in high school exhibit a significantly higher degree of acceptance of evolution as compared to students who were taught both evolutionary theory and creationism, or only creationism (Moore and Cotner 2009; Rissler et al. 2014). Thus, the incorporation of creationism in high school biology instruction significantly increases the likelihood that students accept creationism and reject evolution when they arrive at college. Interestingly, students who were not taught evolutionary theory nor creationism are more likely to accept the scientific validity of evolutionary theory and related concepts upon entering college as compared to peers who experienced high school biology classes that included creationism (Moore and Cotner 2009, p. 97; Rissler et al. 2014), leading to the recommendation that "omission of evolution from high school biology courses may be preferable to a mixed approach that validates nonscientific explanations of diversity" (Moore and Cotner 2009, p. 99).

A recent multifactorial analysis (Dunk et al. 2017) reported that the greatest predictive variable for evolution acceptance (as measured by the MATE instrument, see "Measuring evolution acceptance" below) was student responses to a validated measure of an understanding of the nature of science, "Understanding of Science," (Johnson and Peeples 1987). Additional predictive variables included religiosity (measure obtained using three items from "Evolution Attitudes and Literacy Survey-Short Form (EALS-SF), Short and Hawley 2012), openness to experience (measure obtained by "Big five inventory," John et al. 2008), religious denomination, number of biology courses previously taken, and knowledge of evolutionary biology terms ("Familiarity with Evolutionary Terms," Barone et al. 2014). These variables together accounted for nearly a third of the variation in the measurement of acceptance of evolution, indicating that other variables unidentified in the model contribute significantly to the measure of acceptance in that population. Thus, there are a wide variety of factors, likely some as yet unidentified, that contribute to evolution acceptance.

Measuring evolution acceptance

A number of instruments for measuring acceptance of evolution have been developed, including the Measure of Acceptance of the Theory of Evolution (MATE) (Rutledge and Warden 1999, 2000; Rutledge and Sadler 2007), the Inventory of Student Evolution Acceptance (I-SEA) (Nadelson and Southerland 2012), the Evolutionary

Attitudes and Literacy Survey (EALS) (Short and Hawley 2012), and the Generalized Acceptance of EvolutioN Evaluation (GAENE) (Smith et al. 2016).

The I-SEA is a 24-item Likert-type scale questionnaire designed to capture students' potential differential responses toward the acceptance of microevolution, macroevolution, and human evolution (Nadelson and Southerland 2012, p. 1657), with items evenly distributed across those three subcategories. Confirmatory factor analysis supports the contention that I-SEA has the potential to consistently and reliably measure evolution acceptance overall and differentially across the three constructs (Nadelson and Southerland 2012). However, the instrument may be of more limited utility in differentiating acceptance in microevolution vs. macroevolution constructs in populations with a lower level of evolution understanding, in which microevolution (i.e. variation within a species) and macroevolution (i.e. speciation) are likely to be more conflated (Nadelson and Southerland 2012, pp. 1657, 1659).

The EALS was initially developed as a 104-item instrument developed to measure a wide array of factors related to acceptance of evolution including political ideology, moral objections to evolution, religious identity, activity distrust of scientific enterprise, exposure to evolutionary theory, young earth creationist beliefs, attitudes toward life, intelligent design fallacies, scientific, genetic, and evolutionary literacy, relevance of evolutionary theory, social objections, and demographics (Hawley et al. 2011). Recognizing that 104 items may be cumbersome for implementation by researchers and educators, Short and Hawley (2012) developed an EALS-short form version (EALS-SF) consisting of 62 items that maintains the original instrument's structure and validity.

The MATE is a 20-item Likert-type scale instrument designed to measure acceptance of fundamental evolutionary concepts (Rutledge and Sadler 2007). Although the authors of the MATE instrument separate the 20 items into six evolution concepts, the MATE has generally been considered a unidimensional measure of evolution acceptance. Reliability measures of the MATE indicate the instrument produces high reliability (c. f. Romine et al. 2016, Table 1) and Rutledge and Sadler (2007) report a high test–retest consistency. However, the instrument has been criticized on several fronts, including lacking a clear definition of "acceptance," potential conflation of evolution acceptance with knowledge and/or religious beliefs, as well as inadequate construct validation and unresolved dimensionality (Wagler and Wagler 2013; Romine et al. 2016, p. 2; Smith et al. 2016, p. 1293). As MATE has been widely used, we selected it for our study because it would provide a means of comparing

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 4 of 18

Table 1 Fit and reliability for confirmatory factor analysis models

Model	n	CFI	RMSEA	SRMR	ω_h
GAENE	105	0.98	0.10	0.05	0.96
Pre-MATE—Uni	104	0.93	0.12	0.10	0.97
Pre-MATE—Bi	104	0.93	0.11	0.09	0.96
Post-MATE—Uni	102	0.90	0.14	0.10	0.97
Post-MATE—Bi	102	0.91	0.13	0.09	0.97

CFI: comparative fit index; RMSEA: root mean squared error of approximation; SRMR: standardized root mean squared residual; ω_h: McDonald's hierarchical omega

outcomes in our population to other populations that have been investigated using the MATE.

The GAENE (Smith et al. 2016) is a recently published instrument that has not yet been widely used. However, adoption of this rigorously developed instrument may yield an improved measure of evolution acceptance that does not conflate evolution understanding and evolution acceptance (Smith et al. 2016). GAENE Version 1.0 consisted of 16 Likert-type items; extensive psychometric testing and refinement resulted in GAENE Version 2.0, a 14-item instrument, and GAENE Version 2.1, a psychometrically superior 13-item instrument recommended for use in most settings. Smith et al. (2016) include a comparison of the characteristics of development for the EALS, I-SEA, MATE, and GAENE evolution acceptance instruments (p. 1290).

Still other instruments seek to measure evolution *understanding*, such as the Conceptual Inventory of Natural Selection (CINS), a 20-item multiple choice instrument targeting understanding of natural selection (Anderson et al. 2002), and the Measure of Understanding of Macroevolution (MUM), a 27-item instrument targeting understanding of macroevolution, with 26 multiple choice items and one free-response item (Nadelson and Southerland 2010).

Study objectives

In this study, we sought to investigate students' level of acceptance of the theory of evolution, with a null hypothesis that there would be no change in students' level of acceptance from the beginning of the term to end of term. Topics in evolutionary biology were the focus of instruction both early in the term and late in the term, representing a "book-end" approach that reinforced the foundational nature of evolution understanding for a coherent and unifying lens through which to view all biological knowledge. Early in the term, evolutionary biology topics included an investigation of the history of life on earth with an emphasis on developing students' sense

of deep time (Metzger 2011), and an understanding of the evolutionary relatedness of all life on earth, including familiarity with visual representations of evolutionary relationships and interpretation of evolutionary relationships presented in phylogenetic trees. Later in the term, evolutionary biology topics included patterns and processes of evolution incorporating concepts learned from population genetics and molecular genetics modules earlier in the course.

Since evolution acceptance had not previously been measured at our institution, this study establishes a "baseline" to which future curricular interventions could be compared. Students in our program experience a one-semester foundational biology course with lab in the context of a Health Sciences undergraduate degree program; many of the populations in which evolution acceptance has been studied are students in a biology major with a two-semester introductory biology sequence, or are "non-majors" students. It was therefore of interest to us to investigate our students in comparison to students of other major designations, and assess if our population's level of acceptance was more closely aligned to biology majors or non-majors from other institutions. At other institutions, a review of the published literature demonstrates that some populations experience little or no gain in measures of evolution acceptance (Romine et al. 2016, p. 3), while others demonstrated significant gains pre-to-post instruction (Smith 2010; Romine et al. 2016, p. 3). In many studies in which marked gains are reported, the instructional methods focused on intensive instruction in evolutionary topics (c.f. Wiles and Atler 2011). Our course design did not employ an explicit intervention to promote evolution acceptance, but understanding of key evolutionary principles is a primary course learning objective.

A further objective of our study was to determine which characteristics and performance variables serve as predictors for evolution acceptance in our study population. As evolution acceptance has been demonstrated to have significant relationships with a number of other student characteristics and performance variables, our study includes a consideration of variables for which data were available.

To investigate evolution acceptance, we utilized two independent measures of evolution acceptance: the MATE and the GAENE, and performed an analysis to determine the association between the scores obtained by each instrument. The MATE has been used as a measure of evolution acceptance in at least 25 studies previously, while the GAENE is a recently published instrument (Smith et al. 2016). We are aware of no other study that presents a comparison of scores obtained in a single population for these two evolution acceptance instruments.

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 5 of 18

As previous research (Romine et al. 2016) provided evidence that the MATE instrument may be more appropriately considered as a bidimensional instrument that captures two different constructs of evolution acceptance—Facts and Credibility—we also wished to perform psychometric analyses to determine the most appropriate way to treat the MATE scores (i.e. as a single score or as two separate scores).

Methods

Demographics and incoming performance metrics of study population

This study took place within the context of a health science undergraduate degree program (Bachelor of Science in Health Sciences, BSHS) at a small liberal arts university in the Midwest. Students entering the program were mostly traditional-aged college students. According to institutional data, approximately 75% of students in the program identified as female, and 27% identified as institutionally underrepresented minorities (URM), a designation which includes the categories American Indian, Asian, Black, and Hispanic. All participants included consented to participate in this research in accordance with University of Minnesota IRB protocol #1008E87333.

The total number of students enrolled in the course was 127. A total of 105 students completed all three assessments (pre-MATE, post-MATE, post-GAENE) satisfactorily (participation rate = 105/127 = 82.67%). Of the students participating in the study, 85 (80%) identified as female, and 38 (36.5%) identified as an underrepresented minority (URM). The average number of college credits completed prior to enrolling in the course was 31.33, with an average college GPA of 3.06. The average ACT Math score for this population was 24.64.

Course context

Study subjects were students enrolled in two sections of a 5-credit first-year foundational biology course with lab. Instruction took place in an active learning classroom (Dori and Belcher 2005; Beichner et al. 2007; Walker et al. 2011) with a flipped pedagogy model in which students were expected to, and were held accountable for, engaging with assigned material prior to classroom instruction. The physical classroom environment and curricular design facilitated regular implementation of a variety of teaching and learning activities and Classroom Assessment Techniques (CATs) (Angelo and Cross 1993). In preparation for classroom instruction and activities, students were assigned pre-instruction reading with corresponding preparation questions (i.e. study guide questions). Additionally, students completed a low-stakes pre-class quiz consisting of five questions related to the material in the assigned reading. Students were

allowed two attempts on the pre-class quiz and were able to see which items they answered correctly or incorrectly immediately after submitting the quiz. Additional files posted on the course website included slides, links to online conceptual animations, practice questions, and other resources.

Schedule of course topics

Understanding of the centrality and importance of evolution in the biological sciences was a key learning objective in the course. As such, evolutionary topics were not relegated to one unit and then set aside for the remainder of the term. Rather, the semester began and ended with explicit instruction in evolutionary biology, referred to here as a "bookend" approach. The intervening instruction, while primarily addressing other topics, would also incorporate connections to evolutionary biology as a unifying principle. For example, the unit focusing primarily on metabolism incorporated a consideration of the homologous relationship between the cytochrome proteins of mitochondria and the cytochrome proteins of the chloroplast. Thus, evolutionary principles were reiterated throughout the course as an organizing theme.

Early instruction emphasized deep time as a way of viewing the history of the earth and life on earth, along with evidence for evolution (e.g. fossil record, biogeography, anatomical homologies) and easily recognizable evolutionary processes, such as response to predation selection pressure, with which students likely had some previous exposure or knowledge. In addition to connecting with students' prior knowledge, to extend students' breadth of evolution understanding early in the course, we also included neutral evolutionary processes such as genetic drift, which are less familiar and accessible to students, but which are increasingly prominent in our modern understanding of evolution at the molecular level (Kimura 1977; Bromham and Penny 2003). Later instruction in evolution topics included a more in-depth consideration of sources of genetic variation, molecular evolution, and population genetics. A molecular perspective of evolution is more accessible to students following instruction in other topics such as DNA replication, meiosis, the genetic code, and gene expression, which were addressed between the bookends of evolution instruction in the course. Previous research has demonstrated that placing instruction in genetics prior to instruction in evolution improved students' evolution *understanding*, but did not significantly impact evolution *acceptance* as compared to instruction that places instruction in evolution prior to instruction in genetics (Mead et al. 2017).

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 6 of 18

Calculation of overall course grade

As our study did not employ a separate measure of knowledge in evolution, we chose to use students' final course grades (%) as a measure (albeit, an imperfect measure) of biology knowledge. A student's final course grade was comprised of grade requirements in the following categories, weighted in calculation of the final grade as indicated in parentheses: pre-class quizzes and in-class activities (15%), formal and informal writing assignments (20%), exams (40%) laboratory activities (20%), and reflection exercises (5%).

Additional measures of broad student knowledge—ACT Math score and cumulative college GPA at the start of term—were also included in our investigation.

Implementation of MATE and GAENE instruments

To assess students' acceptance of evolution, we utilized two published instruments: the 20-item Measure of Acceptance of the Theory of Evolution (MATE) (Rutledge and Sadler 2007), and the Generalized Acceptance of EvolutioN Evaluation (GAENE), Version 2.1 (Smith et al. 2016). The GAENE Version 2.1 is a 13-item instrument, which we implemented with random order presentation and 5-point Likert-type scale as per the authors' recommendations (Smith et al. 2016).

The MATE instrument was implemented as a pre- and post-test measure to investigate the level of acceptance in our undergraduate health sciences students before and after instruction, while the GAENE instrument was implemented as a post-test only. In all cases, student responses were gathered via our online course management system. Students completed the assessments outside of class time and were awarded nominal completion points for submitting responses to the instruments. Our online assessment allowed students to enter a numeric character as a response to each Likert-scaled item; instances in which a student entered a non-numeric or multiple numeric characters of different value were deemed ambiguous responses and thus removed from the dataset prior to analysis. Instances in which a student entered the same numeric character multiple times (e.g. 11) were considered non-ambiguous errors of entry and were replaced with a single numeric character of that value. If an individual student had more than one ambiguous character entry for an assessment, that individual was removed from the dataset.

Building validity evidence

Dimensionality—confirmatory factor analysis

The GAENE and the MATE are both intended to be unidimensional measures of evolution acceptance. To contribute evidence for the valid use of these instruments, a confirmatory factor analysis (CFA) was performed to examine the dimensionality of both instruments based on responses from the current sample. A unidimensional model was fit for the GAENE. For both the pre- and post-measures of the MATE a unidimensional and a bidimensional model were tested with the bidimensional model examining whether items loaded onto Romine et al's (2016) proposed Facts and Credibility dimensions. The fit of the uni- and bidimensional models were then compared using the likelihood ratio test, which tests whether the addition of a second dimension significantly improves model fit. Items on both the GAENE and the MATE are five-category Likert-type items and were treated as ordered categorical variables rather than continuous variables in the CFA estimation (Flora and Flake 2017; Flora et al. 2012). As categorical variables, the association of the items and the underlying factor(s) was nonlinear. Consequently, all of the CFA models were estimated with a diagonally weighted least squares estimator, which makes no assumptions about the distribution of the item responses and uses the polychoric, rather than product-moment, correlation matrix (Li 2016; Rhemtulla et al. 2012). The full weight matrix, however, was used to compute robust standard errors and a mean- and variance-adjusted Chi square test statistic. The CFA models were run using the *lavaan* package (v. 0.6-1) in R (Rosseel 2012). The comparative fit index (CFI), root means squared error of approximation (RMSEA), and standardized root mean squared residual (SRMR) were used to assess model fit for the CFA analyses. CFI evaluates incremental fit assessing whether the tested model fits better than the null model that treats all items as completely unrelated to each other. Absolute fit—the degree to which the relationships between variables implied by the model are similar to the relationships actually found in the data—are measured by RMSEA and SRMR with RMSEA including a penalty for greater model complexity. Simulation studies suggest acceptable model fit should have a $CFI \geq 0.95$, $RMSEA \leq 0.06$, and $SRMR \leq 0.08$ (Hu and Bentler 1999). Additionally, hierarchical omega reliability (ω_h) (McDonald 1999) was calculated to evaluate the proportion of total variance in item responses explained by the factor model.

Item calibration and person scores—Rasch scaling

Rasch scaling tests whether data from an instrument fit a theoretical measurement model (Rasch 1960). The Rasch model assumes the instrument is unidimensional, which is why CFA is more useful for examining the dimensionality of instruments. CFA, however, is based on classical test theory and has weaker assumptions than the Rasch model, which is rooted in the item response theory framework (Smith et al. 2002). Thus, when data fit the Rasch model and its stronger assumptions, it provides

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 7 of 18

more appropriate person score and item calibration estimates which contain a number of properties that are beneficial for making both norm-referenced and criterion-referenced score interpretations: (a) item locations and person scores are placed on the same scale (the logit scale). In the current study, a person's score is a measure of their level of acceptance of evolution whereby a person with a high level of acceptance will have a high score on the logit scale. Each item in the instrument is placed on the same logit scale whereby an item reflective of a high level of evolution acceptance when endorsed also has a high score on the scale; (b) the common metric for person and item parameters allows for calculating the probability a person with a certain evolution acceptance will endorse an item at a given location, which is useful for making predictions about person responses and evaluating whether the items on the instrument adequately cover the variability in respondents' acceptance of evolution; (c) despite the ordinal nature of the Likert-type items used in the instrument, when the data fit the Rasch model, the resulting scaled scores are on an interval, linear scale enabling the use of scores in parametric statistical tests; (d) item location estimates are independent of the distribution of person scores and the person score estimates are independent of the item location distribution, thus enabling greater generalizability of the person and item estimates. In contrast, summed scores for the raw item responses are dependent both on the sample of items and the sample of persons, thus making it difficult to predict how persons would respond to a different set of items or how well the items would measure a different sample of persons. Readers interested in learning more about Rasch analysis can consult Wright and Masters (1982) and Bond and Fox (2015) and for the use of Rasch in instrument development see Smith et al. (2002) and Boone (2016).

For the current analysis, all Rasch models were run with the Rasch partial credit model (Masters 1982) using the *mirt* (v. 1.28) package in R (Chalmers 2012). The fit of GAENE and MATE instrument data to the Rasch partial credit model was evaluated using outfit mean square, infit mean square, and marginal reliability for the item location and person scores. Outfit measures how sensitive the item (person) estimates are to outliers while infit measures the difference between the observed score patterns and the model expected score patterns with poor infit being a greater threat to validity than poor outfit for the interpretation and use of scores (Linacre 2002). Outfit and infit are expected to close to 1.0 with values between 0.5 and 1.5 being acceptable.

Unlike classical test theory conceptions of reliability, such as Cronbach's α (Cronbach 1951), that assume an instrument measures people on the underlying construct

equally well across the construct's entire spectrum, the Rasch and other item response theory models do not make this assumption and estimate a reliability for each observed score (Bond and Fox 2015). The average of the reliability estimates from across all observed scores is the marginal reliability. Given that Cronbach's α was commonly reported by others using the MATE and GAENE, it was also calculated for each instrument to allow for comparing the reliability of the instruments on the present sample with previous administrations.

To further examine Romaine et al.'s (2016) Facts and Credibility subscales of the MATE, three separate unidimensional Rasch models were run on the pre-MATE responses: (1) all pre-MATE items, (2) Fact items only, and (3) Credibility items only. Using the item fit approach discussed in Smith (1996), we compared the fit of each item to the Rasch model when it was used with all pre-MATE items or when used only as part of the separate Fact or Credibility dimension. If items tended to fit better in the model with all pre-MATE items this was evidence the pre-MATE is a unidimensional instrument; whereas if the items tended to fit better in the Fact or Credibility models this was evidence the pre-MATE is a bidimensional instrument. The process of running three separate unidimensional Rasch models and using the item fit approach to compare the models was repeated with the post-MATE responses. After evaluating item fit, the stacking procedure outlined by Wright (1996, 2003) was then used to fit three more Rasch partial credit models (an all item model, a Fact item only model, and a Credibility item only model) using both the pre-and post-MATE responses simultaneously in order to estimate comparable pre- and post-MATE person scores. The person scores from these three simultaneously estimated models were used for all subsequent correlation and regression analyses. For the GAENE, a single Rasch partial credit model was run from which the item fit was evaluated and the person scores were used in the correlation and regression analyses.

Change in pre and post MATE responses

Changes in student responses to the MATE instrument from the pre-to post-administrations were investigated in three ways:

1. Change in the simultaneously estimated pre- and post-MATE Rasch-scaled scores were compared with a paired t-test.
2. Change in the raw summed scores calculated as originally proposed by Rutledge and Sadler (2007) were compared using mean normalized change (*c*; Marx and Cummings 2007). Normalized change calculates the *mean of the change* in raw summed score from

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 8 of 18

pre- to post-test, rather than the *change in the mean* raw summed score from pre- to post-test. In keeping with Marx and Cummings (2007) recommendations, students who scored 100% on both the pre- and post-MATE instrument were removed from the analysis of normalized change, as those students' performance was beyond the scope of the instrument's measurement (Marx and Cummings 2007, p. 87).

3. At the item level, the association between raw ordinal responses for all 20 items of the pre- and post-MATE were compared using Cramer's V, an effect size measure from the association based family of effect sizes (Cramer 1946; Cohen 1988). Values for Cramer's V range from 0 to 1, with larger values indicating a stronger association. Cohen's (1988) standard was used to interpret the strength of association, where V values between 0.1 and 0.29 represent a small association, values between 0.3 and 0.49 represent a medium association, and values above 0.5 represent a large association.

Association between MATE and GAENE Rasch-scaled scores

To measure the degree of association between the MATE and GAENE Rasch-scaled scores, bivariate Pearson product-moment correlations were calculated between the GAENE Rasch scores and the pre- and post-MATE Rasch scores from each of the three MATE Rasch models (all items, Fact items only, Credibility items only) when the pre- and post-MATE scores were estimated simultaneously. The correlation coefficients were then disattenuated of (i.e. corrected for) measurement error using the formula first presented by Spearman (1904). Estimates of reliability quantify the extent to which variance in Rasch scores on the evolution acceptance instruments was due to measurement error. Thus, attenuated (i.e. uncorrected) correlations not only measure the association between students' true evolution acceptance as measured by the MATE or GAENE, but also any measurement error. By correcting the correlation by the score reliability of the two instruments, measurement error can be removed from the estimation of the association between the two instruments' measurement of evolution acceptance. The disattenuated correlations also provide evidence for whether the MATE is a unidimensional or bidimensional instrument: if the Rasch scores from the Fact and Credibility dimensions are highly correlated with each other and with the scores from the all-items model, then we can conclude that having separate Fact and Credibility dimension scores does not provide any unique information about students' acceptance of evolution beyond what a unidimensional MATE score provides. As with Cramer's V, Cohen's standard was used to determine the strength of the associations (Cohen 1988).

Multiple regression

The Rasch scores from the all item pre-MATE, all item post-MATE, and GAENE were used as the outcome variable in three separate multiple regression models to investigate variables possibly predictive of evolution acceptance: gender, ethnicity, college GPA, and Math ACT. Two additional regression models were run to further investigate the association between the MATE and GAENE instruments while controlling for the other variables in the regression model. First, with the post-MATE as the outcome variable, the pre-MATE and GAENE Rasch scores were added to the initial model, and second, with the GAENE as the outcome variable, the pre- and post-MATE Rasch scores were added to the initial model.

Results

Validity evidence

Dimensionality

Confirmatory factor analysis was used to evaluate whether the MATE should be considered a unidimensional measure of evolution acceptance or bidimensional instrument measuring separate Facts and Credibility dimensions for a sample of health science undergraduate students. A likelihood ratio test directly comparing the unidimensional and bidimensional models was significant for both the pre-MATE ($\chi^2 = 7.32$, $df = 1$, $p = 0.01$) and the post-MATE ($\chi^2 = 29.04$, $df = 1$, $p < 0.01$), indicating that the bidimensional model significantly improved model fit. The fit statistics (Table 1) for both the unidimensional model and the bidimensional model at pre- and post-administration, however, fall outside Hu and Bentler's (1999) criteria that an acceptable model should have a CFI ≥ 0.95, RMSEA ≤ 0.06, and SRMR ≤ 0.08. This suggests that although the bidimensional model is the better fitting model, it is still a poor fit for the data. In contrast, all of the MATE models had high hierarchical omega reliability with values > 0.95, indicating in a classical test theory sense that a large proportion of variation in the observed raw summed scores on the MATE was true variation in the summed scores as opposed to measurement error. Given that CFA is sample dependent, the high reliability yet poor model fit suggest the MATE raw summed scores were measured with precision, but the summed score was a weak measure of the underlying construct for this sample of health science undergraduate students.

For the GAENE, the unidimensional model met the fit criteria on the CFI and SRMR for acceptable fit with values of 0.98 (≥ 0.95) and 0.05 (≤ 0.08), respectively, but the RMSEA of 0.10 was above the criteria of ≤ 0.06. Hu and Bentler (1999) note, however, that when CFI > 0.96 models can still have acceptable fit when RMSEA and SRMR > 0.09. Therefore, the GAENE unidimensional

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 9 of 18

model can be considered an adequate fit for the data from the sample. Taken in conjunction with the high hierarchical omega reliability ($\omega_h = 0.96$), these results are evidence that the raw summed score from the GAENE is a precise measure that can be interpreted as an adequate indication of evolution acceptance for a student in the sample. Although the GAENE and MATE models cannot be compared directly, the model fit from the CFA provide evidence that the GAENE is a better instrument for measuring evolution acceptance than the MATE for health science undergraduate students in the sample.

Item calibration and person scores from Rasch scaling

Results from the Rasch analysis on the pre-MATE and post-MATE data provide ambiguous evidence for whether the MATE should be considered a unidimensional or bidimensional instrument. Data fit the Rasch model when the item and person outfit mean square and infit mean square are close to 1.0. Additionally, the item location and person score estimates on the shared logit scale are more precise as the marginal reliability (ρ) approaches 1.0. The person outfit, infit, and marginal reliability were 0.99, 1.02, and 0.93 for the pre-MATE with all items model (Table 2). These were closer to 1.0 than the corresponding values for the pre-MATE Facts (outfit = 0.90, infit = 0.93, $\rho = 0.90$) or Credibility models (outfit = 0.91, infit = 0.94, $\rho = 0.86$). The item marginal reliability was similar for the three pre-MATE models (All items: $\rho = 0.87$, Facts: $\rho = 0.86$, Credibility: $\rho = 0.87$). The item outfit and infit values for each item, displayed in Fig. 1, are more indicative of fit than the scale-level values (Linacre 2002). Acceptable values are between 0.5 and 1.5 (see Additional file 1 for full item-level statistics). For the pre-MATE with all items model, items 2, 15, and 19 had outfit and/or infit > 1.5, indicating these items did not fit the Rasch model and inclusion of these items deteriorates

the quality of the instrument. For the pre-MATE Facts and Credibility models only item 15 had an outfit > 1.5, meaning that this item increased the instruments measurement error because it was overly sensitive to outliers in the person responses.

The post-MATE with all items model had person outfit, infit, and marginal reliability of 1.05, 1.08, and 0.93, respectively, which were equal or closer to 1.0 than the corresponding values for the post-MATE Facts (outfit = 0.95, infit = 0.91, $\rho = 0.89$) or Credibility models (outfit = 0.91, infit = 0.94, $\rho = 0.86$). The item marginal reliability was similar for the three post-MATE models (All items: $\rho = 0.78$, Facts: $\rho = 0.80$, Credibility: $\rho = 0.77$). As shown in Fig. 1, items 11, 17, and 19 had outfit > 1.5 for the post-MATE with all items model, indicating these items did not fit the Rasch model due to their sensitivity to person response outliers. Only item 11 had outfit > 1.5 for the post-MATE Facts and Credibility models.

For the last set of Rasch models on the MATE, the pre-MATE and post-MATE data were used simultaneously to estimate the item locations and person scores primarily for the purpose of creating comparable pre- and post-MATE person scores. The all items model had person outfit, infit, and marginal reliability of 1.01, 1.05, and 0.93, respectively, which were closer to 1.0 than the corresponding values for the Facts (outfit = 0.91, infit = 0.92, $\rho = 0.90$) or Credibility models (outfit = 0.91, infit = 0.94, $\rho = 0.86$). The item marginal reliability was similar for the three MATE models with pre- and post-responses estimated simultaneously (All items: $\rho = 0.91$, Facts: $\rho = 0.91$, Credibility: $\rho = 0.92$). As shown in Fig. 1, items 2, 15, and 19 had outfit > 1.5 in the all items model while items 11 and 15 had outfit > 1.5 in the Facts and Credibility models. These items did not fit the Rasch model as a result of over sensitivity to person response outliers. In consideration of whether the MATE is a unidimensional

Table 2 Fit and reliability for Rasch scales

Model	Item			Person			
	Outfit	Infit	ρ	Outfit	Infit	ρ	α
Pre-MATE All items	0.99	0.98	0.87	0.99	1.02	0.93	0.93
Pre-MATE Facts	0.90	0.93	0.86	0.90	0.93	0.90	0.90
Pre-MATE Credibility	0.91	0.96	0.87	0.91	0.94	0.86	0.83
Post-MATE All items	1.05	0.98	0.78	1.05	1.08	0.93	0.92
Post-MATE Facts	0.95	0.96	0.80	0.95	0.91	0.89	0.90
Post-MATE Credibility	0.91	0.97	0.77	0.91	0.94	0.86	0.84
MATE All items	1.01	0.98	0.91	1.01	1.05	0.93	0.93
MATE Facts	0.91	0.95	0.90	0.91	0.92	0.90	0.90
MATE Credibility	0.91	0.96	0.92	0.91	0.94	0.86	0.84
GAENE	0.94	0.95	0.86	0.94	0.98	0.93	0.93

ρ: marginal reliability; α: Cronbach's alpha

Metzger *et al. Evo Edu Outreach (2018) 11:10*

Page 10 of 18
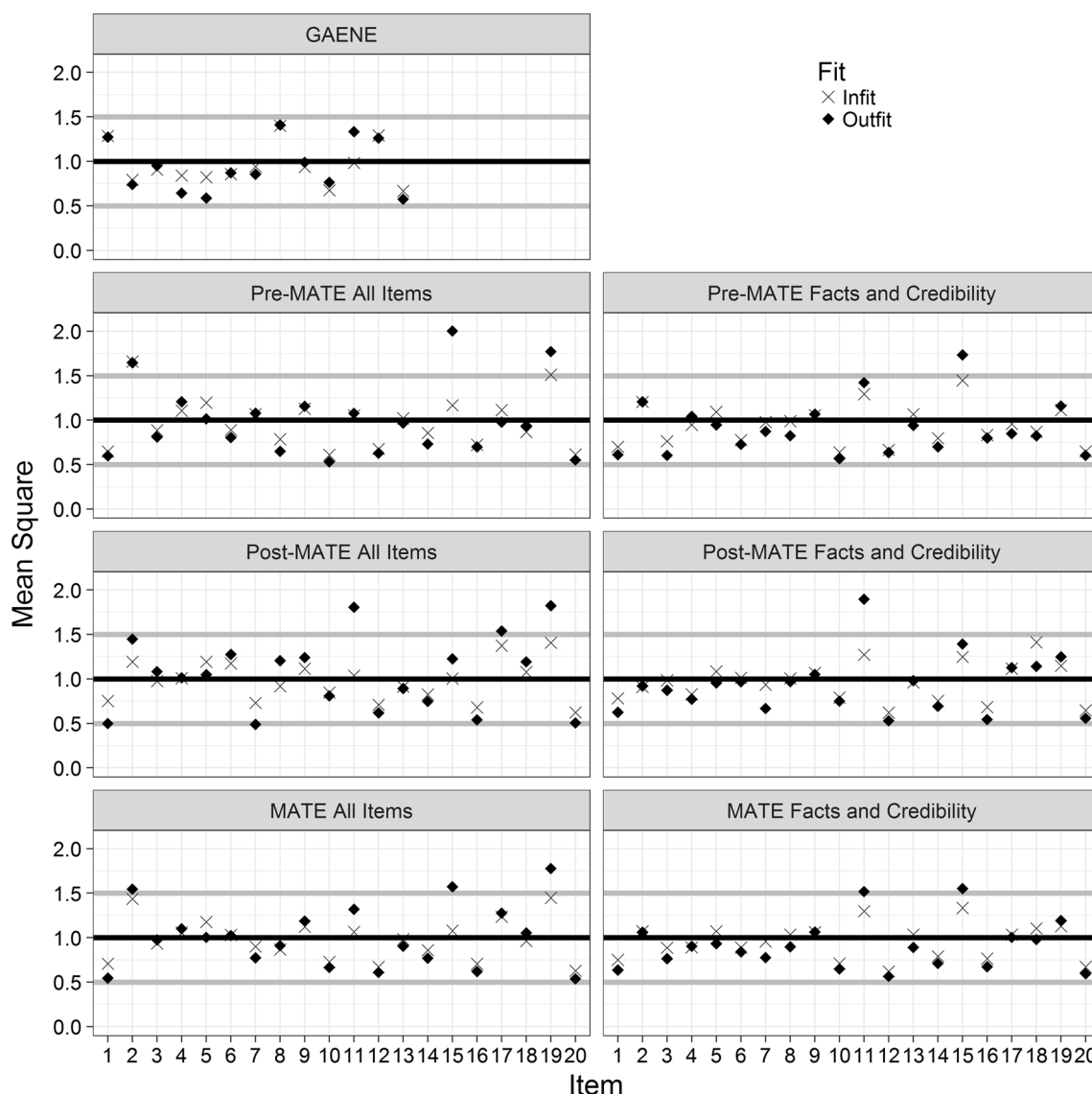
Fig. 1 Item level outfit mean square and infit mean square for Rasch scales. Item outfit and infit should be close to 1, designated in the figure by the black line, and values between 0.5 and 1.5, designated by the grey lines, are considered acceptable

or bidimensional instrument the results do not provide clear support. For the pre-MATE, post-MATE, and simultaneously estimated MATE the models with all items produced equal or better person fit and reliability, but the Facts and Credibility models demonstrated better item fit.

The Rasch model using the pre- and post-MATE data simultaneously was used for comparison with the Rasch model fit to the GAENE data. The person outfit, infit, and marginal reliability for the GAENE were 0.94, 0.98, and 0.93, respectively. Despite containing seven fewer items, the person fit and reliability for the GAENE was better than the MATE Facts and Credibility models and similar to all items MATE model. The item reliability for the GAENE ($\rho = 0.86$) was lower than the MATE models (All items: $\rho = 0.91$, Facts: $\rho = 0.90$, Credibility: $\rho = 0.92$); however, this result is unsurprising given that a larger person sample leads to higher item reliability and the MATE models used both pre- and post-responses, and thus, had twice the sample size as the GAENE (Bond and Fox 2015). Regarding item fit, while both the MATE with all items and the MATE Fact and

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 11 of 18

Credibility models had multiple items with high outfit, all of the items on the GAENE demonstrated acceptable infit and outfit.

Results from the Rasch analysis suggest that the GAENE data fit the Rasch model meaning the resulting item locations and person scores are placed on the same linear and interval-level scale with the item locations independent of the person score distribution and, unlike the raw summed scores, the person scores are independent of the item location distribution. Although the person fit was acceptable, regardless of whether the MATE was estimated with all items or with the Facts and Credibility dimensions estimated separately, some items demonstrated high outfit suggesting that the MATE is sensitive to person response outliers. Poor outfit, however, is less of a threat to validity for the interpretation and use of scores than infit (Linacre 2002), so as a whole the MATE can be considered an adequate fit for the Rasch model. Nonetheless, the Rasch analysis provides evidence that the GAENE more appropriately measures evolution acceptance than the MATE.

Changes in pre- and post-MATE scores

When the Rasch-scaled pre-MATE and post-MATE scores were estimated simultaneously for comparison, students demonstrated a significant change in scores ($t(104) = 3.94$, $p < 0.01$) from pre- ($M = -0.18$, $SD = 1.26$) to post-assessment ($M = 0.18$, $SD = 1.41$). The effect size of the Rasch score change of 0.36 from pre- to post-MATE was $d = 0.38$, considered a small to medium effect size (Cohen 1988). Using the raw scores, the mean pre-MATE score was 78.68 ($SD = 12.44$) and mean post-MATE score was 81.72 ($SD = 12.41$), with a mean normalized change (*c*) of 14.21%. According to categories of acceptance developed by Rutledge (1996) and reported in Rutledge and Sadler (2007), MATE raw scores between 77 and 88 represent "High Acceptance".

At the item level, a Cramer's V association of 1.00 signifies that a student's response to an item on the pre-MATE was a perfect indicator of the student's response to the item on the post-MATE whereas a Cramer's V of 0.00 means a student's pre-MATE response was unrelated to their post-MATE response. The Cramer's V associations ranged from 0.31 to 0.59 (Fig. 2) suggesting there was a medium to large association (i.e. effect size) in pre- and post-MATE raw ordinal responses for all items, but also that there was some change in response patterns between pre- and post-administrations of the MATE.

Investigating the association between MATE and GAENE scores

GAENE scores

Students in our study obtained a mean Rasch scaled score of -0.01 ($SD = 1.79$) and mean summed raw score of 51.70 ($SD = 9.02$) on the GAENE instrument. Unlike for the MATE, the authors of the GAENE instrument elected not to propose cutoff scores to delineate what GAENE score constitutes low acceptance, moderate acceptance, and high acceptance (Smith et al. 2016, pp. 1309–1310).
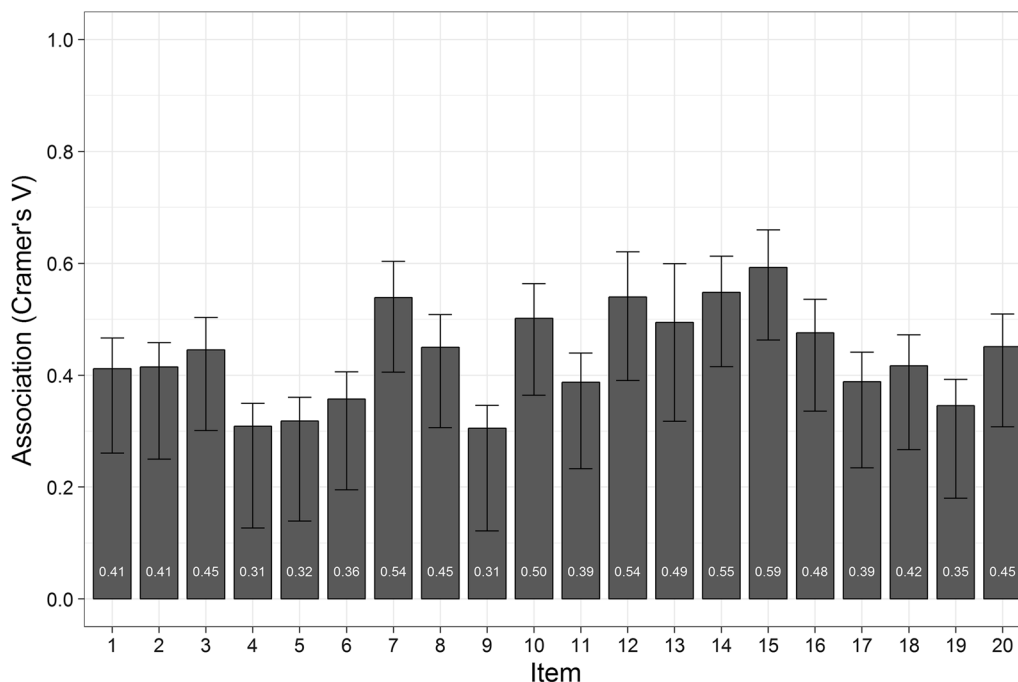
Fig. 2 Cramer's V Association between pre- and post-MATE responses. Error bars represent the 95% confidence interval of the V statistic

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 12 of 18

Disattenuated correlations

Disattenuated correlations correcting for measurement error reveal significant, strong correlations between the evolution acceptance Rasch scores produced by the MATE and GAENE instruments (Table 3). The GAENE was only administered at the end of the semester and are most appropriately compared to the post-MATE Rasch scores. Nonetheless, significant associations between the GAENE Rasch scores and both the pre-MATE and post-MATE held when compared with the Rasch scores from all MATE items, or when compared with the Rasch scores from the Facts and Credibility dimensions of the MATE.

The disattentuated correlation between the Facts Rasch score and the Credibility Rasch score was 0.95 and 0.92 for both pre- and post-MATE, indicating that after correcting for measurement error the two scores were largely redundant. Additionally at both pre- and post-administration, the Facts and Credibility Rasch scores were perfectly correlated ($r = 1.00$) with the Rasch score from all MATE items meaning that the Fact and Credibility scores provided no additional information above and beyond what was already provided by the unidimensional scores. Therefore, from a practical standpoint, using and reporting a unidimensional MATE score is more efficient than separate Fact and Credibility scores.

Multiple regression

Multiple regression models were run separately on pre-MATE, post-MATE, and GAENE Rasch scores with gender, URM status, college GPA, course performance, and Math ACT as variables possibly predictive of evolution acceptance (Table 4). Overall the variables explained little of the variation in evolution acceptance scores with R^2 values of 0.07, 0.12, and 0.11 for the pre-MATE, post-MATE, and GAENE, respectively. The R^2 for the pre-MATE model was lower in part because course performance was not included in the model given that course performance was measured after the pre-MATE and therefore could not be a predictor. The regression models with demographic and academic performance as

predictors identified only URM as a significant predictor of Rasch GAENE score ($\beta = -0.37$, $p = 0.03$), but none were significantly associated with pre- or post-MATE Rasch scores. The association between URM and GAENE became non-significant ($\beta = -0.16$, $p = 0.09$), however, after adding pre- and post-MATE scores to the model. In contrast, both the pre-MATE ($\beta = 0.31$, $p < 0.01$) and post-MATE ($\beta = 0.72$, $p < 0.01$) scores were significant indicating that the two time points each explain unique variation in GAENE scores and highlights that there are differences between the pre- and post-MATE scores.

Although the demographic and academic performance variables only explained 11% of the variation in GAENE scores, pre-MATE ($\beta = 0.31$, $p < 0.01$) and post-MATE ($\beta = 0.72$, $p < 0.01$) scores when added to the model explained an additional 62% of the variation in GAENE scores. Similarly, the demographic and academic performance variables only explained 12% of the variation on post-MATE scores with pre-MATE ($\beta = 0.27$, $p < 0.01$) and GAENE ($\beta = 0.56$, $p < 0.01$) scores explaining an additional 62% of the variation in post-MATE scores when added to the model (Table 4).

Discussion

Evolution acceptance in undergraduate health sciences majors

The students in our study reported a high level of evolution acceptance at the start of the semester: the average pre-test value based on the raw score for the MATE in our sample was 78%, which is just above the boundary between 'Moderate Acceptance' (65–76%) and High Acceptance (77–88) using the categories of acceptance developed by Rutledge (1996) and reported in Rutledge and Sadler (2007). This result is strikingly similar to the average raw MATE score of 77.17% obtained by Dunk et al. (2017, Table 5). The demographic composition of the sample in the study by Dunk et al. (2017) was also strikingly similar to ours: "skewed young, white, and female with a high proportion of health majors". By comparison, other studies have reported lower MATE raw scores in college biology majors (Rissler et al. 2014; Ingram and

Table 3 Disattenuated correlations for Rasch-scaled scores

	1	2	3	4	5	6	7
1. GAENE	–						
2. Pre-MATE All items	0.79	–					
3. Pre-MATE Facts	0.83	1.00	–				
4. Pre-MATE Credibility	0.76	1.00	0.95	–			
5. Post-MATE All items	0.90	0.82	0.83	0.79	–		
6. Post-MATE Facts	0.90	0.83	0.87	0.79	1.00	–	
7. Post-MATE Credibility	0.87	0.77	0.77	0.77	1.00	0.92	–

All correlations were significant at $p < 0.05$

Table 4 Unstandardized and standardized coefficients (standard errors) for regression models

Predictor	Pre-MATE		Post-MATE				GAENE			
	b	β	b	β	b	β	b	β	b	β
Intercept	− 1.60 (1.27)	0.00 (0.00)	− 3.32 (2.70)	0.00 (0.00)	− 2.47 (1.51)	0.00 (0.00)	− 0.18 (3.41)	0.00 (0.00)	3.28 (1.92)	0.00 (0.00)
Course performance			2.43 (4.56)	0.07 (0.14)	2.32 (2.53)	0.07 (0.08)	− 0.76 (5.76)	− 0.02 (0.17)	− 3.21 (3.23)	− 0.10 (0.10)
Cumulative GPA	0.24 (0.31)	0.09 (0.12)	0.30 (0.61)	0.07 (0.14)	− 0.04 (0.34)	− 0.01 (0.08)	0.64 (0.77)	0.14 (0.17)	0.39 (0.43)	0.09 (0.10)
Female	− 0.38 (0.34)	− 0.13 (0.12)	− 0.12 (0.38)	− 0.04 (0.14)	0.29 (0.21)	0.10 (0.08)	− 0.61 (0.48)	− 0.22 (0.17)	− 0.38 (0.27)	− 0.14 (0.10)
URM	− 0.16 (0.27)	− 0.07 (0.12)	− 0.49 (0.30)	− 0.22 (0.14)	− 0.05 (0.17)	− 0.02 (0.08)	− 0.81 (0.38)*	− 0.37 (0.17)*	− 0.36 (0.22)	− 0.16 (0.10)
ACT Math	0.04 (0.04)	0.13 (0.12)	0.03 (0.05)	0.10 (0.14)	0.03 (0.03)	0.07 (0.08)	− 0.01 (0.06)	− 0.03 (0.17)	− 0.05 (0.03)	− 0.16 (0.10)
Pre-MATE					0.32 (0.09)*	0.27 (0.08)*			0.38 (0.12)*	0.31 (0.10)*
Post-MATE									0.78 (0.11)*	0.72 (0.10)*
GAENE					0.48 (0.06)*	0.56 (0.08)*				
n	101		101		101		101		101	
R^2	0.07		0.12		0.74		0.11		0.73	
Adj-R^2	0.03		0.07		0.72		0.06		0.71	
AIC	329.2		351.1		233.2		398.2		282.8	

URM underrepresented minority

* $p < 0.05$

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 14 of 18

Table 5 MATE scores reported in the literature for various populations

Description of population	N	Pre-MATE/single measurement	Post-MATE	Citation
Upper level evolution students, Midwest university	225, over 3 offerings	>60%	77, 70, 74	Ingram and Nelson (2006)
Undergraduate medical students—Pakistan	271	58.32	–	Yousuf et al. (2011)
Undergraduate non-science majors	61	Test: 55.8 Re-test: 56.1	–	Rutledge and Sadler (2007)
Undergraduate non-science majors	268	64.9	65.9	Walter et al. (2013)
High school students—intensive evolution instruction	81	72.8	85.8	Wiles and Alters (2011)
High school teachers—Turkey	132	50.95	–	Deniz et al. (2008)
Undergraduate Intro Bio for non-majors—Deep South	138	62.27	74.58	Rissler et al. (2014)
Undergraduate Honors Bio students—Deep South	41	69.48	84.69	Rissler et al. (2014)
Undergraduate Anatomy & Physiology students—Deep South	308	61.65	60.82	Rissler et al. (2014)
Undergraduate Anatomy & Physiology students—Midwest	284	77.17	–	Dunk et al. 2017
Undergraduate Intro Bio for health sciences major—Midwest	105	78.68	81.72	Present study

Nelson 2006), college non-biology majors (Rutledge and Sadler 2007; Deniz et al. 2008), gifted high school students (Wiles and Alters 2011), and biology teachers (Rutledge and Warden 1999). From this, we conclude that evolution acceptance in our population of health sciences students is relatively high when compared to a number of other university student populations, both biology majors and non-biology majors, although not all (Table 5).

Change in evolution acceptance pre- to post-test
Although this study did not investigate the impact of a specific curriculum intervention, students in this study were enrolled in a foundational introductory biology course and experienced instruction in a wide variety of biology topics, including evolution, between the administration of the pre- and post-MATE. A significant increase in students' reported level of evolution acceptance was found between pre and post MATE Rasch scores. Other studies implementing a pre- and posttest design using the MATE as an instrument have similarly reported significant gains pre- to post-test (Rissler et al. 2014; Ingram and Nelson 2006; Wiles and Alters 2011), while others have failed to find a significant different following instruction (Walter et al. 2013). Rissler et al. (2014) reported significant gains in evolution acceptance, but only for the "least religious" students (p. 11). From our results, we conclude that the curriculum design and instruction implemented for our undergraduate introductory biology course is having an impact on student acceptance of evolution. We think this is notable for at least two reasons: (1) change was demonstrated after a single semester of instruction as opposed to a two-semester sequence, and (2) students' level of evolution acceptance was significantly positively impacted despite not having an explicit emphasis or curriculum intervention designed to target

evolution acceptance. This result appears to be consistent with other studies reporting increased evolution acceptance as a result of instruction in general biology and other courses in which evolution is a topic of study (c.f. Wiles and Alters 2011), but in contrast to courses in which topics in evolution are likely absent (e.g. anatomy and physiology) and no change in evolution acceptance is observed (c.f. Rissler et al. 2014, p. 10).

Validity evidence
The confirmatory factor analysis directly compared a unidimensional and bidimensional model for the MATE with the significant likelihood ratio test at both pre- and post-test providing evidence that structurally the MATE is a bidimensional instrument. The Rasch analysis provided additional, albeit limited, evidence for a bidimensional MATE structure as the all item model had more misfitting items than the Fact and Credibility models at pre- and post-test and when the pre- and post-MATE data were used simultaneously. The person fit and reliability, however, favored the all items model and the disattentuated correlations showed that the having two scores for the MATE was redundant. Therefore, results from the present study suggest that while the MATE might more appropriately measures two dimensions of evolutionary acceptance, interpretation and use of a single unidimensional score is equally informative and more practically efficient. The evidence is, however, ambiguous enough to warrant further investigation. The vague dimensionality could be due to measurement error or the MATE might measure evolution acceptance differently under different circumstances or with different groups of people. One future avenue of research to address this quandary would be to perform a differential item functioning analysis to investigate measurement invariance between

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 15 of 18

various groups of respondents, such as students in natural and health science majors versus students in liberal arts majors or people with high and low religiosity.

The GAENE produced adequate fit and reliability in both the CFA and Rasch analysis to provide converging evidence that the GAENE is a unidimensional measure of evolution acceptance. The MATE, in addition to the ambiguity of its dimensionality, had poor model fit in the CFA, and the Rasch analysis showed that some items were over-sensitive to outlier person responses. Therefore, the psychometric evidence points to the GAENE being the superior measure of evolution acceptance.

Measures of student performance and evolution acceptance

Measures of student performance (overall course grade, ACT Math, college GPA at start of term) did not emerge as predictive of MATE and GAENE Rasch scores in regression modeling (see Table 4). While some authors have reported significant associations between knowledge in evolution and acceptance of evolution (Rutledge and Warden 2000; Nadelson and Southerland 2010; Walter et al. 2013; Carter and Wiles 2014), others have found no significant association between knowledge of and acceptance of evolution (Cavallo and McCall 2008; Sinatra et al. 2003). A proposed limitation of the MATE instrument for measuring students' level of acceptance is the possible conflation of knowledge and acceptance by inclusion of items in an acceptance instrument that measure knowledge (Smith et al. 2016, p. 1293), and that considering the MATE as a bidimensional instrument may help to address this issue. However, we found limited evidence for considering the MATE as a bidimensional instrument, and no practical utility for reporting scores beyond a single unidimensional score.

While the average MATE raw score for our sample indicates "High Acceptance" according to the categories developed by Rutledge (1996) and reported in Rutledge and Sadler (2007), the GAENE has no developed cutscores for interpreting relative acceptance using raw scores from this instrument (Smith et al. 2016, p. 1310). As we are aware of no other study that reports scores from both MATE and GAENE in the same sample, it will be of interest to see if future work replicates the significant and strong correlation between the MATE and GAENE evolution acceptance scores as we report here.

There are several additional limitations which may affect the results reported in our study. First, our investigation of evolution acceptance involves a relatively small number of students that is not intended to be representative of all undergraduate populations; our study sample is a reasonable representation of the undergraduate population of health sciences majors at our institution, and thus is informative. In this study, we sought to investigate evolution acceptance in this population in conjunction with other variables that have been reported to co-vary or impact evolution acceptance, including student demographic and performance variables. However, we did not include a formal measure of student knowledge in evolution, but rather used more holistic measures of student knowledge (e.g. overall course performance, ACT Math, college GPA). As such, we are limited to the extent that we can comment on the relationship between student *knowledge* in evolution and *acceptance* of evolution. Further, overall course performance is not a perfect measure or representation of a students' knowledge in biology. Additional measures of broad student knowledge—ACT Math score and cumulative college GPA at the start of term—were also included in our investigation. While geographical location and context may impact students' evolution acceptance (Berkman and Plutzer 2011; Belin and Kisida 2014), we also did not address this as a variable in our study. A final limitation of our study is that we did not include a measure of religiosity, a variable which has been repeatedly reported to have a significant association with evolution acceptance (Smith 2010; Rissler et al. 2014; Dunk et al. 2017).

These identified limitations point to future directions for continued investigation to more thoroughly understand evolution acceptance in this population of undergraduate students, and broadly. We are also quite interested in exploring, as others have done (c.f. Smith 2010) the impact of curricular modifications to determine if differences in instructional approaches will affect either short- and/or long-term measures of evolution acceptance in our students. Further work should also address potential disparities in evolution acceptance between URM and non-URM status students. While the present work identified URM as a variable predictive of evolution acceptance (with URM students having lower acceptance as compared to non-URM students), this variable was not predictive in all models, and thus is as yet of ambiguous importance to evolution acceptance broadly, and evolution acceptance in this population specifically.

Additional files

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 16 of 18

Abbreviations
CINS: Conceptual Inventory of Natural Selection; EALS: Evolutionary Attitudes and Literacy Survey; GAENE: Generalized Acceptance of EvolutioN Evaluation; I-SEA: Inventory of Student Evolution Acceptance; MATE: Measure of Acceptance of Evolution; MUM: Measure of Understanding of Macroevolution.

Authors' contributions
KJM made substantial contributions to conception and design, acquisition of data, analysis and interpretation of data as well as drafting and revising of the manuscript. DM contributed to the experimental design, acquisition of data and drafting and of the manuscript. DH contributed to the acquisition of data and drafting and drafting of the manuscript. KN contributed to data analysis and interpretation, and revision of the manuscript. All authors read and approved the final manuscript.

Author details
[1] Center for Learning Innovation, University of Minnesota Rochester, 300 University Square, 111 South Broadway, Rochester, MN 55904, USA. [2] University of Minnesota Twin Cities, 56 East River Road, 250 Education Sciences Building, Minneapolis, MN 55455, USA.

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

References
American Association for the Advancement of Science. Atlas of science literacy. Washington, DC: AAAS/National Science Teachers Association; 2001. http://www.project2061.org/publications/atlas/. Accessed 23 Jan 2018.

American Association for the Advancement of Science. Science for all Americans: a project report on literacy goals in science, mathematics, and technology. Washington, DC: AAAS; 1989. http://www.project2061.org/publications/sfaa/. Accessed 23 Jan 2018.

American Association for the Advancement of Science. Vision and change in undergraduate biology education: a call to action. Washington, DC; 2011. http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf. Accessed 23 Jan 2018.

Anderson RD. Teaching the theory of evolution in social, intellectual, and pedagogical context. Sci Educ. 2007;91:664–77.

Anderson DL, Fisher KM, Norman GJ. Development and evaluation of the conceptual inventory of natural selection. J Res Sci Teach. 2002;39:952–78.

Angelo TA, Cross KP. Classroom assessment techniques. 2nd ed. San Francisco: Jossey-Bass; 1993.

Armstrong O. Beating the evolution laws. Pop Sci Mon. 1929;1929(11517–19):134–5.

Barone LM, Petto AJ, Campbell BC. Predictors of evolution acceptance in a museum population. Evol Educ Outreach. 2014;7:23.

Baum DA, Offner S. Phylogenetics & tree-thinking. Am Biol Teach. 2008;70:222–9.

Baum DA, Witt SDS, Donovan SS. The tree-thinking challenge. Science. 2005;310:979–80.

Beichner RJ, Saul JM, Abbott DS, Morse JJ, Deardorff DL, Rhett JA, Bonham SW, Dancy MH, Risley JS. The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project. In: Redish EF, Cooney PJ, editors. Research-based reform of university physics. College Park: American Association of Physics Teachers; 2007.

Belin CM, Kisida B. State science standards, science achievement, and attitudes about evolution. Educ Policy. 2014;29:1053–75.

Berkman MB, Plutzer E. Defeating creationism in the courtroom, but not in the classroom. Science. 2011;331:404–5.

Bishop BA, Anderson CW. Student conceptions of natural selection and its role in evolution. J Res Sci Teach. 1990;27:415–27.

Bond TG, Fox CM. Applying the rasch model: fundamental measurement in the human sciences. 3rd ed. New York: Routledge; 2015.

Boone WJ. Rasch analysis for instrument development: why, when, and how? CBE Life Sci Educ. 2016;15(4):rm4.

Bromham L, Penny D. The modern molecular clock. Nat Rev Genet. 2003;4(3):216–24.

Carter BE, Wiles JR. Scientific consensus and social controversy: exploring relationships between students' conceptions of the nature of science, biological evolution, and global climate change. Evol Educ Outreach. 2014;7:6.

Carter BE, Infantini LM, Wiles JR. Boosting students' attitudes & knowledge about evolution sets them up for college success. Am Biol Teach. 2015;77:113–6.

Catley KM, Novick LR. Digging deep: exploring college students' knowledge of macro-evolutionary time. J Res Sci Teach. 2009;46:311–32.

Cavallo AM, McCall D. Seeing may not mean believing: examining students' understandings & beliefs in evolution. Am Biol Teach. 2008;70:522–30.

Chalmers RP. mirt: a multidimensional item response theory package for the R environment. J Stat Softw. 2012;48(6):1–29.

Chi MT. Commonsense conceptions of emergent processes: why some misconceptions are robust. J Learn Sci. 2005;14:161–99.

Clough EE, Driver R. A study of the consistency in the use of students' conceptual frameworks across different task contexts. Sci Educ. 1986;70:473–96.

Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Lawrence Earlbaum Associates; 1988.

Cramer H. Mathematical methods of statistics. Princeton: Princeton University Press; 1946.

Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334. https://doi.org/10.1007/BF02310555.

Cummins CL, Demastes SS, Hafner MS. Evolution: biology education's underresearched unifying theme. J Res Sci Teach. 1994;31:445–8.

Deniz H, Donnelly LA, Yilmaz I. Exploring the factors related to acceptance of evolutionary theory among Turkish preservice biology teachers: toward a more informative conceptual ecology for biological evolution. J Res Sci Teach. 2008;45:420–43.

Dobzhansky T. Nothing in biology makes sense except in the light of evolution. Am Biol Teach. 1973;35:125–9.

Dori YJ, Belcher J. How does technology-enabled active learning affect undergraduate students' understanding of electromagnetism concepts? J Learn Sci. 2005;2005(14):243–79.

Dunk RDP, Petto AJ, Wiles JR, Campbell BC. A multifactorial analysis of acceptance of evolution. Evol Educ Outreach. 2017;10:4.

Evans EM. The emergence of beliefs about the origins of species in school-age children. Merrill-Palmer Q. 2000;46:221–54.

Evans EM. Conceptual change and evolutionary biology: A developmental analysis. In: Vosniadou S, editor. International handbook of research on conceptual change. New York: Routledge; 2008. p. 263–94.

Fishman RS. Evolution and the eye. Arch Ophthalmol. 2008;126:1586.

Flora DB, Flake JK. The purpose and practice of exploratory and confirmatory factor analysis in psychological research: decisions for scale development and validation. Can J Behav Sci. 2017;49(2):78–88.

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 17 of 18

Flora DB, LaBrish C, Chalmers RP. Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. Front Psychol. 2012;3:55.

Gallup. Evolution, creationism, intelligent design. Gallup.com. 2016. https://news.gallup.com/poll/21814/evolution-creationism-intelligent-design.aspx?version=print. Accessed 17 Nov 2017.

Glaze AL, Goldston MJ. US science teaching and learning of evolution: a critical review of the literature 2000–2014. Sci Educ. 2015;99:500–18.

Good RG, Trowbridge JE, Demastes SS, Wandersee JH, Hafner MS, Cummins CL. Toward a research base for evolution education: report of a national conference. In: EDRS conference proceedings, ED 361 183, SE 053 585, evolution education research conference, Baton Rouge, LA. 1992.

Gregg TG, Janssen GR, Bhattacharjee JK. A teaching guide to evolution. Sci Teach. 2003;70:24–31.

Hawley PH, Short SD, McCune LA, Osman MR, Little TD. What's the matter with Kansas?: the development and confirmation of the Evolutionary Attitudes and Literacy Survey (EALS). Evol Educ Outreach. 2011;4(1):117–32.

Heddy BC, Nadelson LS. The variables related to public acceptance of evolution in the United States. Evol Educ Outreach. 2013;6:1–14.

Hendry AP, Kinnison MT, Heino M, Day T, Smith TB, Fitt G, Bergstrom CT, Oakeshott J, Jørgensen PS, Zalucki MP, Gilchrist G, Carroll SP. Evolutionary principles and their practical application. Evol Appl. 2011;4:159–83.

Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model Multidiscip J. 1999;6(1):1–55. https://doi.org/10.1080/10705519909540118.

Ingram EL, Nelson CE. Relationship between achievement and students' acceptance of evolution or creation in an upper-level evolution course. J Res Sci Teach. 2006;43:7–24.

John OP, Naumann LP, Soto CJ. Paradigm shift to the integrative big-five trait taxonomy: history, measurement, and conceptual issues. In: John OP, Robins RW, Pervin LA, editors. Handbook of personality theory and research. 3rd ed. New York: Guilford Press; 2008. p. 114–58.

Johnson RL, Peeples EE. The role of scientific understanding in college: student acceptance of evolution. Am Biol Teach. 1987;49:93–8.

Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267(5608):275–6.

Li CH. Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. Behav Res Methods. 2016;48(3):936–49.

Linacre JM. What do infit and outfit, mean-square and standardized mean? Rasch Meas Trans. 2002;16(2):878.

Marocco DA. Biology for the 21st century: the search for a core. Am Biol Teach. 2000;62:565–9.

Marx JD, Cummings K. Normalized change. Am J Phys. 2007;75:87–91.

Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982;47(2):149–74.

Mayr E. The growth of biological thought: diversity, evolution and inheritance. Cambridge: Harvard University Press; 1982.

Mazur A. Believers and disbelievers in evolution. Trends Cogn Sci. 2004;23:55–61.

McDonald RP. Test theory: a unified treatment. Mahwah: Erlbaum; 1999.

Mead R, Hejmadi M, Hurst LD. Teaching genetics prior to teaching evolution improves evolution understanding but not acceptance. PLoS Biol. 2017. https://doi.org/10.1371/journal.pbio.2002255.

Meir E, Perry J, Herron JC, Kingsolver J. College students' misconceptions about evolutionary trees. Am Biol Teach. 2007;69:1–76.

Metzger K. Helping students conceptualize species divergence events using the online tool "TimeTree: the timescale of life". Am Biol Teach. 2011;73(2):106–8. https://doi.org/10.1525/abt.2011.73.2.9.

Miller JD, Scott EC, Okamoto S. Public acceptance of evolution. Science. 2006;313:765–6.

Moore R, Cotner S. The creationist down the hall: does it matter when teachers teach creationism? Bioscience. 2009;59(5):429–35.

Moore R, Kraemer K. The teaching of evolution and creationism in Minnesota. Am Biol Teach. 2005;67:457–66.

Moore R, Brooks DC, Cotner S. The relation of high school biology courses and students' religious beliefs to college students' knowledge of evolution. Am Biol Teach. 2011;73(4):222–6.

Nadelson LS, Hardy KH. Trust in science and scientists and the acceptance of evolution. Evol Educ Outreach. 2015;8:9. https://doi.org/10.1186/s12052-015-0037-4.

Nadelson LS, Southerland SA. Development and preliminary evaluation of the measure of understanding of macroevolution: introducing the MUM. J Exp Educ. 2010;78:151–90.

Nadelson LS, Southerland S. A more fine-grained measure of students' acceptance of evolution: development of the Inventory of Student Evolution Acceptance—I-SEA. Int J Sci Educ. 2012;34:1637–66.

Naegle E. Patterns of thinking about phylogenetic trees: a study of student learning and the potential of tree thinking to improve comprehension of biological sciences. D.A. dissertation, Idaho State University; 2009.

National Research Council. BIO2010: transforming undergraduate education for future research biologists. Washington, DC: National Academies Press; 2010.

National Research Council. A framework for K-12 science education: practices, crosscutting concepts, and core ideas. Washington, DC: The National Academies Press; 2012. https://doi.org/10.17226/13165. Accessed 13 July 2018.

National Research Council. Next generation science standards. 2013. http://www.nextgenscience.org/. Accessed 27 Nov 2017.

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. 2007;45:1131–60.

Novick LR, Catley KM. Assessing students' understanding of macroevolution: concerns regarding the validity of the MUM. Int J Sci Educ. 2012;34:2679–703.

Pew Research Center. Views of evolution. 2015. http://www.pewresearch.org/fact-tank/2017/02/10/darwin-day/ft_15-02-11_darwin/. Accessed 17 Nov 2017.

Quinn H, Schweingruber H, Keller T. A framework for K-12 science education: practices, crosscutting concepts, and core ideas. Washington, DC: National Academies Press; 2011.

Rasch G. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Oxford: Nielsen & Lydiche; 1960.

Rhemtulla M, Brosseau-Liard PÉ, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods. 2012;17(3):354.

Rice JW, Olson JK, Colbert JT. University evolution education: the effect of evolution instruction on biology majors' content knowledge, attitude toward evolution, and theistic position. Evol Educ Outreach. 2011;4:137–44.

Rissler L, Duncan S, Caruso N. The relative importance of religion and education on university students' views of evolution in the Deep South and state science standards across the United States. Evol Educ Outreach. 2014;7:24. https://doi.org/10.1186/s12052-014-0024-1.

Romine WL, Walter EM, Bosse E, Todd AN. Understanding patterns of evolution acceptance—a new implementation of the Measure of Acceptance of the Theory of Evolution (MATE) With Midwestern University Students. J Res Sci Teach. 2016;54:642–71.

Rosseel Y. Lavaan: an R package for structural equation modeling. J Stat Softw. 2012;48(2):1–36.

Rutledge ML. Indiana high school biology teachers and evolutionary theory: acceptance and understanding. Doctoral Dissertation, Ball State University; 1996.

Rutledge ML, Mitchell MA. High school biology teachers' knowledge structure, acceptance & teaching of evolution. Am Biol Teach. 2002;64:21–8.

Rutledge ML, Sadler KC. Reliability of the Measure of Acceptance of the Theory of Evolution (MATE) instrument with university students. Am Biol Teach. 2007;69:332–5.

Rutledge ML, Warden MA. The development and validation of the Measure of Acceptance of the Theory of Evolution instrument. Sch Sci Math. 1999;99:13–8.

Rutledge ML, Warden MA. Evolutionary theory, the nature of science & high school biology teachers: critical relationships. Am Biol Teach. 2000;62:23–31.

Scharmann L, Harris W. Teaching evolution: understanding and applying the nature of science. J Res Sci Teach. 1992;29:375–88.

Short SD, Hawley PH. Evolutionary Attitudes and Literacy Survey (EALS): development and validation of a short form. Evol Educ Outreach. 2012;5:419–28.

Metzger *et al. Evo Edu Outreach* (2018) 11:10

Page 18 of 18

Sinatra GM, Southerland SA, McConaughy F, Demastes JW. Intentions and beliefs in students' understanding and acceptance of biological evolution. J Res Sci Teach. 2003;40:510–28.

Smith RM. A comparison of methods for determining dimensionality in Rasch measurement. Struct Equ Model Multidiscip J. 1996;3(1):25–40.

Smith MU. Current status of research in teaching and learning evolution: II. Pedagogical issues. Sci Educ. 2010;19:539–71.

Smith EV, Conrad KM, Chang K, Piazza J. An introduction to Rasch measurement for scale development and person assessment. J Nurs Meas. 2002;10(3):189–206.

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;9:1289–315.

Spearman C. The proof and measurement of association between two things. Am J Psychol. 1904;15:72–101.

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;35:2278–98.

Walker JD, Brooker, DC, Baepler P. Pedagogy and Space: empirical research on new learning environments. EDUCAUSE Q. 2011;34. http://er.educause.edu/articles/2011/12/pedagogy-and-space-empirical-research-on-new-learning-environments. Accessed 28 Feb 2018.

Walter EM, Halverson KL, Boyce CJ. Investigating the relationship between college students' acceptance of evolution and tree thinking understanding. Evol Educ Outreach. 2013;6:26.

Wiles JR, Alters B. Effects of an educational experience incorporating an inventory of factors potentially influencing student acceptance of biological evolution. Int J Sci Educ. 2011;18:2559–85. https://doi.org/10.1080/09500693.2011.565522.

Wright BD. Time 1 to time 2 (pre-test to post-test) comparison: racking and stacking. Rasch Meas Trans. 1996;10:478.

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

Wright BD, Masters GN. Rating scale analysis. Chicago: Mesa Press; 1982.

Yousuf A, Daud MA, Nadeem A. Awareness and acceptance of evolution and evolutionary medicine among medical students in Pakistan. Evol Educ Outreach. 2011;4:580–8. https://doi.org/10.1007/s12052-011-0376-8.