Evolution: Education and Outreach

Open Access

CrossMark

A closer look at the items within three measures of evolution acceptance: analysis of the MATE, I-SEA, and GAENE as a single corpus of items

William L. Romine[1*], Amber N. Todd[2] and Emily M. Walter[3]

Abstract

Background: Current direct Likert measures for evolution acceptance include the MATE, GAENE, and I-SEA. Pros and cons of each of these instruments have been debated, and yet there is a dearth of research teasing out their similarities and differences when they are used together in a single context beyond the fact that their measures tend to be highly correlated. We administered these to 452 college students in non-major biology classes at two research-intensive universities from the Midwestern and Western United States to investigate the measurement properties of the items within these instruments when combined as a single corpus.

Results: Factor analysis using exploratory and confirmatory methods, and Rasch analyses, suggested that a two-dimensional factor structure best describes the corpus of items. Whether the item was positively or negatively worded was the key delimiter in its factor assignment. Examination of the highest loading items on the respective factors indicates that the first factor measures *acceptance of the truth of evolution* and the second factor measures *rejection of incredible ideas about evolution*. The correlation of these two factors is 0.73, indicating that they share 53% of their variance with each other. When treated unidimensionally, eleven items exhibited potential misfit with the Rasch model. This number dropped to nine items when the two factors were considered. These items, and implications for future use of the MATE, GAENE, and I-SEA together, are discussed in detail.

Conclusions: This study is the first analysis of the MATE, GAENE, and I-SEA as a single corpus of items, and yet corroborates previous work showing that these instruments yield measures with highly similar quantitative interpretations. This study also corroborates the effect of negative item wording on how college students interpret the item. While this finding can be applied to college-level students taking undergraduate non-majors biology coursework, work with more advanced biology students has demonstrated that this apparent item wording effect tends to disappear as students advance and become more accepting of evolution. We conclude that despite apparent epistemological differences between the MATE, GAENE, and I-SEA, these can be treated as a single set of items measuring a single factor or two factors without significant loss of quantitative interpretability.

Keywords: Evolution, Acceptance, Measurement, Macroevolution, Microevolution, Human evolution, Exploratory factor analysis, Rasch analysis

*Correspondence: romine.william@gmail.com
[1] Department of Biological Sciences, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45431, USA
Full list of author information is available at the end of the article

BMC

Romine *et al. Evo Edu Outreach*   (2018) 11:17

Page 2 of 20

Introduction

Darwin's release of *On the Origin of Species* in 1859 sent a wave of disconcert which surged rapidly across the seas. According to a Pew Research Center survey reported in Masci (2017), only 62% of United States residents agree that humans have evolved over time, and only half of these state that evolution is the sole reason that humans change. These statistics are even more sobering in Latin America and the Middle East, where less than half of the current citizenry accepts evolution (Masci 2017). A 2008 report suggests that as few as 1 in 5 college students agree that humans evolve (Lovely and Kondrick 2008); indeed this nearly 160-year-old conflict has been a continual bane of efforts to teach the biological sciences in an authentic, useful, and accurate way (Smith 2010). Despite a history of anti-evolutionist pressure from individual states (Lerner 2000), current K-16 standards both recognize and emphasize the importance of including evolution as a foundational component of biology instruction (NGSS Lead States 2013; Brewer and Smith 2011).

Increasing evolution acceptance remains an important implicit target for biology instruction across the globe, now more than ever given recent empirical (Romine et al. 2016; Nadelson and Southerland 2010) and theoretical (Deniz et al. 2008; Ha et al. 2012) work documenting the marriage of evolution acceptance and content understanding. Indeed, evolution acceptance may be an important component of meeting the plausibility and fruitfulness conditions suggested by conceptual change theory (Strike and Posner 1992) required before students are willing to accept scientific ideas explaining how species change over time (Deniz et al. 2008). Interpretation of the moderate relationship between evolution knowledge and acceptance (Romine et al. 2016; Nadelson and Southerland 2010) through the lens of conceptual change theory suggests that evolution acceptance may serve as an important intermediary between biology instruction and a student's willingness to actually change his/her ideas about evolution (Deniz et al. 2008). Putting a quantitative measure on evolution acceptance therefore becomes important toward facilitating understanding of how our instruction is impacting the way our students think about evolution.

This call to assess evolution acceptance has been addressed multiple times in the past two decades. Rutledge and Warden (1999) advanced the first effort to provide a valid quantitative measure for evolution acceptance—the Measure of Acceptance of the Theory of Evolution (MATE) (also described in Rutledge and Sadler 2007). Since its development, the MATE has been used extensively to measure evolution acceptance within a variety of instructional settings and populations including science teachers and students at academic levels spanning from elementary to undergraduate (Romine et al. 2017). Through its usage, a number of limitations of the MATE have surfaced (for a detailed discussion, see Hogan 2000; Smith 2010; Smith et al. 2016; Wagler and Wagler 2013; Nadelson and Southerland 2012). Two recent evolution acceptance instruments, the Inventory of Student Evolution Acceptance (I-SEA) (Nadelson and Southerland 2012) and the Generalized Acceptance of Evolution Evaluation (GAENE) (Smith et al. 2016) were developed to provide measures that overcome key limitations of the way in which the MATE measures evolution acceptance, which include the lack of attention paid to the context of the evolutionary event and the conflation with evolution knowledge and religiosity.

Nadelson and Southerland (2012) developed the I-SEA to make measures of evolution acceptance more fine-grained, embracing the possibility that evolution acceptance may comprise multiple related constructs which account for the specific type of evolution being considered. The authors cite that micro- and macroevolution are viewed differently by students (Nehm and Ha 2011). Specifically, many who reject macroevolution may readily accept ideas about microevolution (Scott 2005), and further, even those who accept evolution over long time scales often believe that humans are exempt from the process of evolution (Gallup 2010).

Some items of the MATE involve context and others do not but the role of context in measurement of acceptance was not carefully considered. Hence the I-SEA instrument puts forth a three-dimensional model where acceptance of evolution is assessed along three constructs: (1) microevolution, (2) macroevolution, and (3) human evolution. In this article, we use the word "dimension" to refer to a quantitative representation of a construct which accounts for the correlation between item responses (Kline 2014). In discussion of the methods and results, we will also use the term "factor", which refers to an individual construct or dimension (Kline 2014). In the case of the I-SEA, Nadelson and Southerland (2012) use three dimensions to account for the relationships between the responses, whereas the MATE and the GAENE use a single dimension to account for the correlation between responses on their respective items. In the development of the GAENE, Smith et al. (2016) argue that conflation of acceptance with knowledge, belief, and religious connotation limits the content validity of the MATE, thereby limiting our ability to use the MATE as a valid measure of evolution acceptance. Smith and colleagues henceforth developed set of items which are worded in such a way that they avoid these confounding factors.

All three of the above instruments have undergone validation efforts in undergraduate populations, and the

frameworks underlying their development have clear qualitative differences. This said, it is difficult to overlook the similarity in structure of the MATE, I-SEA, and GAENE, namely that they comprise Likert items asking participants the extent to which they agree or disagree with a statement about evolution. A recent study was undertaken to tease out differences between the MATE and the GAENE within an undergraduate health sciences context (Metzger et al. 2018). Metzger and colleagues found that, while the GAENE shows better fit statistics than the MATE within a confirmatory factor analysis model, the reliability of both instruments is high (r > 0.9 when treated unidimensionally), and agreement between unidimensional MATE and GAENE measures (disattenuated r = 0.9) is high enough to justify that both instruments may generate similar quantitative conclusions within the same study. Can this be taken to imply that measures generated by the MATE, GAENE, and I-SEA actually harbor a similar quantitative interpretation? To what extent can the MATE, I-SEA, and GAENE be used interchangeably as a single corpus of items? In this study, we address these overarching questions. To follow, we provide a review of current evolution acceptance instrumentation, how these measures fit into current theoretical frameworks around evolution acceptance, and then proceed to discuss our research questions.

Validated measures of evolution acceptance

Both qualitative and quantitative methods have been utilized to understand evolution acceptance. Qualitative methods include open response (Nehm and Schonfeld 2007; Robbins and Roy 2007) and interview protocols (Donnelly et al. 2009; Nehm and Reilly 2007). In this study, we focus on quantitative measurement of evolution acceptance, which has been pursued with significant interest for over two decades. Johnson and Peeples (1987) represented one of the first attempts to measure evolution acceptance using a survey with documented psychometric information. Just over a decade later, the Measure of Acceptance of the Theory of Evolution (MATE) was developed (Rutledge and Warden 1999).

The MATE contains 20 total items—10 positively-worded and 10 negatively-worded—which focus primarily on acceptance of evolution over long time scales including common ancestry and human evolution. The MATE has been both utilized and criticized extensively since its development (for a review, see Romine et al. 2017). The MATE has been used as a unidimensional measure throughout its history. By "unidimensional", we are referring to the idea that a single measure or score is sufficient to describe the relationships between students' responses (Kline 2014). This statistical argument is extended to the qualitative interpretation that the MATE measures acceptance of evolution as a single construct. Wagler and Wagler (2013) provided evidence that the MATE is not unidimensional, and further exhibits dynamic dimensionality across different groups, meaning that the MATE measures different constructs within different populations. Using Rasch analysis, Romine et al. (2017) corroborated the claim of multidimensionality in a sample consisting primarily of science non-majors, finding that the MATE works better as a two-dimensional instrument. This means that two scores or measures are needed to capture the relationships between undergraduate students' responses on the items. In Romine and colleagues' study, the MATE was found to measure the following constructs: (1) acceptance of facts and (2) acceptance of credibility of statements related to evolution. This dimensionality was delineated by whether or not the items were negatively worded.

The Inventory of Student Evolution Acceptance (I-SEA) and the Evolution Attitudes and Literacy Survey (EALS) are also constructed multidimensionally, meaning that multiple quantitative measures are used to account for students' observed responses (Kline 2014). The I-SEA (Nadelson and Southerland 2012) was designed to improve on the MATE by disentangling microevolution, macroevolution, and human evolution contexts. This is an important contribution given that microevolutionary events may be easier for students to accept than those related to macro- and human evolution (Alters and Alters 2001; Scott 2005). Given these goals, the I-SEA consists of 24 total items, 9 of which are negatively worded, with 8 items assigned to one of three subscales for constructs defining acceptance of microevolution, macroevolution, and human evolution, respectively. The authors hypothesized that the items related to acceptance of microevolution would be easier than the items measuring acceptance of macroevolution and human evolution, and their analysis bears this out (Nadelson and Southerland 2012). Here, it may be useful to draw a distinction between Nadelson and Southerland's (2012) decision to use 3 dimensions to explain the differences in items, including their difficulty, versus letting the items take a difficulty hierarchy along a single dimension as is often done in Rasch studies (Boone et al. 2013). Nadelson and Southerland show that the 3-dimensional model explains students' responses adequately; however, the efficacy of the comparatively parsimonious approach of defining the items hierarchically along a single unidimensional Rasch scale remains unexplored.

The Evolution Attitudes and Literacy Survey (EALS) is a 104-item instrument designed to measure 16 lower-order and 6 higher-order constructs which have been known to influence acceptance of evolution. Some of these include creationist reasoning, political/religious

conservativism, knowledge, misconceptions, and exposure to evolution (Hawley et al. 2011). Given the length of the EALS, a study to shorten it was undertaken, and a short form (EALS-SF) was thereby developed which the authors claim retains the psychometric integrity of the original (Short and Hawley 2012). The approach taken by the EALS contrasts with other existing instruments which ask students about their acceptance of evolution directly.

The Generalized Acceptance of Evolution Evaluation (GAENE) is the most recent development in the line of evolution acceptance instrumentation. It is claimed to measure evolution acceptance unidimensionally without the potential confounds with knowledge of evolution and religion (Smith et al. 2016). The GAENE consists of 13 items, all of which are positively-worded. Unlike the MATE and the I-SEA, the items on the GAENE have less focus on specific evolutionary contexts. Rather, the items tend to provide more general opinion-related statements including whether one accepts that there is evidence for evolution, whether evolution is important in the study of biology, and whether one is willing to advocate for evolution. Like the MATE, the GAENE has been validated using Rasch methods.

Lack of alignment with theoretical understanding of evolution acceptance

Although current evolution acceptance instruments generate measures which have shown to be useful within specific research studies, we find that these instruments do not address adequately current theoretical perspectives around evolution acceptance. The model put forth by Deniz et al. (2008) considers evolution thinking in terms of cognitive, affective, and contextual domains. Their model conceptualizes evolution acceptance as a product of both cognition and affect around the theory of evolution. While the model of Deniz et al. (2008) focuses on evolution thinking in general, the model of Ha et al. (2012) has a more constrained focus on evolution acceptance specifically, suggesting that this is mediated through both conscious and unconscious pathways. Foremost, people will make a conscious decision on whether or not they accept evolution through their understanding of evolution and their reasoning around what they perceive as evidence for or against evolution. Indeed, facilitation of conscious experiences is an important focus of evidence-based pedagogies focused on evolution (Romine and Todd 2017; Beard et al. 2014). Ha et al. (2012) further propose that these conscious experiences will merge with past experiences, values, and social interactions to generate unconscious feelings which manifest as an underlying feeling of certainty.

Significant work has been done towards exploring the relationship between knowledge of evolution and acceptance of evolutionary theory (Nadelson and Southerland 2010; Romine et al. 2017). However, it is currently difficult to tease out how much of this relationship is due to the fact that evolution acceptance as it is currently being measured reflects knowledge of evolution. The MATE has been criticized on the grounds that it possibly conflates knowledge of evolution with acceptance of evolutionary theory (Hogan 2000; Smith 2010; Wagler and Wagler 2013), and both the MATE and the I-SEA are criticized on the grounds that they are not tied to a theoretical foundation for evolution acceptance (Smith et al. 2016).

Purpose of the Research

What all of these instruments have in common is that they give the student a statement, and then ask for an indication of agreement or disagreement with that statement via a Likert scale. It is clear that these instruments have important qualitative differences; for example, the MATE and GAENE do not have the focus on microevolution which is seen in the I-SEA. It can also be argued that the GAENE is more context-independent than the MATE and I-SEA. However, even a cursory inspection of these instruments reveals that their item stems are quite similar in their wording and/or get at acceptance in a similar way. For example, the GAENE asks for agreement with: "Evolution is a scientific fact". The MATE asks for agreement with: "Evolution is not a scientifically valid theory". As another example, the I-SEA asks: "Species exist today on the same form that they always have", while the MATE asks: "Humans exist today in the same form that they always have". While not all of the items bear this level of similarity, one can reasonably posit that a person agreeing with one statement will tend to agree with other related statements in the corpus regardless of the instrument on which these statements appear. However, this needs to be tested, and hence raises two potential lines of inquiry: (1) do the MATE, I-SEA, and GAENE get at a similar construct, and (2) how are the items on these instruments similar or different in how they measure evolution acceptance? We explore these lines of inquiry through three questions:

1. What is the dimensionality of the corpus of items provided by the MATE, I-SEA, and GAENE, and how can the dimension(s) be interpreted?
2. Taking into account the dimensionality, which items tend to provide the most useful measures of evolution acceptance, and which items could be considered problematic?

3. What do these instruments together tell us about the construct of evolution acceptance as it is currently being measured?

Methods

Context

We administered the MATE, I-SEA, and GAENE 2.1, to 452 college students taking introductory biology classes at two doctoral universities with moderate research activity using an IRB-approved protocol. Of these students, 354 hailed from a university in the Midwestern United States, and 98 were attending a university in the Western United States. One hundred fifty-two students (34%) reported male gender, and 273 (60%) reported female. Two hundred sixty-nine students (60%) reported White ethnicity. However, other ethnicities were reported, including Black (52, 12%), Hispanic (53, 12%), and Asian (29, 7%). Eight students (2%) reported indigenous American heritage, and 25 students (6%) reported other ethnicities not included in the survey. We finally inquired into students' religious preferences. Two hundred ninety-eight students (66%) reported Christian faith. Agnostic (28, 6%) and Atheist (23, 5%) preferences were reported by 51 students. Eighteen students (4%) reported Islamic faith, 8 (2%) were Buddhist, and 3 (1%) where Hindu. The remainder reported other faiths, or chose not to report.

Factor analysis

We used exploratory factor analysis (EFA) in SPSS 21.0 to reconcile how the items on these three instruments define latent dimensions related to evolution acceptance, and then followed Nadelson and Southerland (2012) in proceeding to use confirmatory factor analysis (CFA) to evaluate how well the hypothesized factor solution derived from EFA reproduces the relationships between the item responses. While the dimensionality of these individual instruments has been discussed and explored in previous work, we initially took the position in this study that we did not know the dimensionality of this collection of items when they are administered together, thereby warranting an *a posteori* approach for exploring dimensionality. This can be contrasted with the a priori, or hypothesis-driven, approach that involves study of the items through a pre-specified model. EFA is a variable-centered clustering technique where latent factors among observed variables are extracted mathematically from the data (Collins and Lanza 2013). EFA has been used extensively in science education research, particularly in the context of instrument validation (i.e. Romine et al. 2013; Corwin et al. 2015), and is a technique that is often associated with classical test validation methods and other

situations where the researcher wishes to reduce a larger feature set to relatively few latent constructs or dimensions (Henson and Roberts 2006).

While methods from Rasch, Item Response Theory (IRT), and CFA traditions are confirmatory in nature, meaning that the measurement model is specified a priori, EFA is exploratory in nature, meaning that the measurement model is extracted mathematically from the data *a posteori*. EFA is therefore a useful tool when the researcher intends to extract a factor structure from the data mathematically without invoking prior information (Costello and Osborne 2005). Given its exploratory nature, a key challenge when conducting an EFA was deciding how many factors should be extracted from the data. Many different guidelines exist for making this decision which can result in different conclusions; we used the scree test (Cattell 1966) in combination with "reasoned reflection" (Henson and Roberts 2006, p. 399) based on the structure of the solution. Specifically, we sought a solution that both makes sense and yields dimensions which are conceptually separable, which is referred to as simple structure (Bryant and Yarnold 1995). After factor extraction, we utilized the promax rotation method (Hendrickson and White 1964) to diagonalize the factor solution. This oblique rotation method allowed the factors to be correlated, which assisted us in deriving a conceptually understandable and separable factor solution.

Although the scree criterion and the requirement of simple structure can provide evidence for the suitability of the factor solution derived from EFA, there are nonetheless multiple methods for extracting the optimal number of factors which often yield different conclusions (Fabrigar and Wegener 2011). Further, a decision regarding the extent to which a loading is high or low, important or ignorable, is subjective. Following Nadelson and Southerland (2012), we therefore used a CFA process to quantify the extent to which our factor solutions derived from EFA actually reproduced the data. Our models assumed no cross-loading between items and utilized the probit link to account for the categorical nature of the responses. Estimation was done using the diagonally weighted least squares (WLSMV) estimator in Mplus 7 (Muthén and Muthén 2012). Fit with the data was evaluated using the Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI). An RMSEA close to 0.06 (Hu and Bentler 1999), but not above 0.07 (Steiger 2007) is indicative of acceptable fit. Acceptable fit is also indicated by a CFI and TLI above 0.9 (Bentler 1990; Bentler and Bonnet 1980).

Before proceeding to discuss the use of Rasch methods in modeling construct validity of the items, we would like

to outline some limitations of EFA, and factor analysis in general. From the perspective of test validation, EFA is criticized based on its treatment of ordinal observations as linear measures and its tendency to select for items which are similar in difficulty (Wright 1996), which is contrary to the objective of more contemporary methodologies such as Rasch modeling (Wright 1996; Boone and Scantlebury 2006). Further, factor analysis is a data-fitting approach which can be dubious when validity of the assessment is in question—rather, a data-independent validity standard is needed (Boone and Scantlebury 2006). Given these criticisms, and its discord with the Rasch validation philosophy in particular, we do not use factor analysis as an end to validation, but instead as a starting point to inform us about how to proceed with Rasch modeling.

Rasch analysis
After making a reflective and informed decision on the number of factors to retain, and observing the specific items loading onto particular factors, our attention turned to construct validity of the individual items towards measurement of each dimension, which we call an *acceptance sub-construct* from here on. Given that a student's appraisal of each item was rated on a Likert scale, we utilized the Rasch rating scale model (Andrich 1982) as a criterion for validity. Rasch models differ in philosophy from other models like IRT and classical test theory (CTT) in that Rasch provides a philosophical criterion for validity of the data as opposed to letting the model define the data. This practice of evaluating the quality of the data through a standard model is different from the statistical practice of fitting the model to the data, which the goal of IRT and CTT. Whether to use Rasch or IRT modeling for validation of tests and surveys has been debated contentiously for many years (Andrich 2004). This said, it is straightforward to argue that the Rasch approach better aligns with the laboratory practice of calibrating machines based on a fixed standard before their use. Further, use of a fixed standard better aligns with the scientific practice of falsification (Popper 1957; Lakatos 1976) than use of statistical data-fitting approaches. Beyond these epistemological advantages, the Rasch approach facilitates generation of item-independent estimates of acceptance and person-independent estimates of item difficulty (also true with IRT), generation of linear estimates, and the ability to map student and item measures along a common scale (Boone and Scantlebury 2006).

We used the WINSTEPS software package (Linacre 2006) to fit a unidimensional Rasch rating scale model to each dimension using the items within each acceptance sub-construct informed by the factor analyses. These provided a standard for the validity of the items and subscales. Conformity of the response patterns of each item to expected fit with the Rasch model was evaluated through mean squares fit. We used infit and outfit of the item response patterns with respect to what the model would predict based on the item's difficulty. These measures of fit have expected values of 1, but Wright et al. (1994) suggest that values between 0.5 and 1.5 are indicative of potentially useful items. Other sources suggest a more conservative range of 0.7–1.3 for acceptable agreement with expected fit (Bond and Fox 2015). Values below 0.5 indicate that the item fits the model better than we would expect in real data, possibly indicating bias in the item that favors students with high acceptance (Masters 1988). With regards to measurement validity, misfitting items represent a significant validity concern as they indicate bias in favor of students with weak levels of acceptance which contradicts the intended scale directionality (Bond and Fox 2015). We evaluated items with respect to the mean squares fit thresholds of 1.3 and 1.5 as potentially indicative of moderate and high misfit, respectively, with Rasch model expectations.

In addition to item validity, we were also interested in the usefulness of the scale itself in explaining and quantifying students' acceptance. Of primary concern was unidimensionality of the items in each acceptance sub-construct: the efficacy of the Rasch model in capturing the systematic variance in the item responses within each subscale. As with item fit, we evaluated unidimensionality using a falsificationist perspective. We first invoked the assumption that the items are unidimensional, and then attempted to falsify this by inspecting the residuals in the item responses with respect to the model. If the items measure a single dimension, then the residuals should be random, or homoscedastic. We implemented principal components analysis (PCA) on the residuals, and inspected the first eigenvalue. If the residuals are random, this eigenvalue should fall below 2 items of variance (Linacre and Tennant 2009). A first eigenvalue above 2 indicates some systematic departure from unidimensionality in the set of items with respect to the intended acceptance sub-construct and warrants exploration of which particular items are responsible for multidimensionality of the scale.

Results
Factor structure of the item corpus
The first eigenvalue from the factor solution on 57 total items is 28.9 items of variance, which accounts for 50.8% of the total item variance. This large first eigenvalue provides a compelling argument that the collection of items from these three instruments may prove useful for providing a unidimensional measure for acceptance

of evolution. However, the second eigenvalue accounts for 3.5 items of variance, an additional 6.2% of the total variance. After 2 dimensions, the eigenvalue of the subsequent dimension drops to 1.9 items of variance, for an additional 3.3% of variance. By Cattell's (1966) scree criterion (scree plot in Fig. 1), this indicates that 2 dimensions may provide a more informative description of the structure of this corpus of items. CFA on the factor solution also bears this out. A 1-dimensional CFA model yields an RMSEA of 0.08, a CFI of 0.92, and a TLI of 0.91. This indicates moderate-to-good fit. However, a 2-dimensional CFA model with ignorable cross-loading between items significantly improves fit to the data ($\chi^2 = 98.2$, $df = 1$, p ≪ 0.001), yielding an RMSEA of 0.064, a CFI of 0.95, and a TLI of 0.94. CFI and TLI values indicate that the 2-dimensional model fits well. That the RMSEA is below 0.07, also indicates acceptable fit. It is noteworthy that our RMSEA value exceeds that reported by Nadelson and Southerland (2012) for the I-SEA, and the CFI values in these respective studies match. These indices indicate collectively that a 2-dimensional factor structure is both necessary and sufficient to obtain acceptable fit with the data. If one adopts the 2-dimensional model, the correlation between these two acceptance sub-constructs is 0.73, indicating that they share 53% of their variance with each other. This demonstrates that these two

sub-constructs share significant similarity, but nonetheless have important differences that need explaining.

The items display an interesting pattern in the oblique 2-dimensional factor structure (Table 1); namely that items with positive loadings (measured between 0.37 and 0.91) onto the first dimension (F1) are positively worded while items with positive loadings (measured between 0.50 and 0.92) onto the second dimension (F2) are negatively worded. Further, we observe that these loadings exhibit the simple structure that is coveted in measurement (Bryant and Yarnold 1995)—negatively-worded items have comparatively small loadings onto the first dimension (between −0.26 and 0.28) while positively worded items have comparatively small loadings onto the second dimension (between −0.19 and 0.24). This indicates that the measure of each acceptance sub-construct aligning with a dimension has a unique and precise meaning which is independent of the other dimension—a measure for the first dimension is not confounded by the second dimension, and vice versa. We now consult the factor loadings on the 2-dimensional model (columns 3 and 4 in Table 1) to evaluate the extent to which this solution makes sense, and to gain insight into the structure of these dimensions. We evaluated their qualitative meaning by inspecting the items with the highest factor loadings onto the respective dimensions (Table 2).
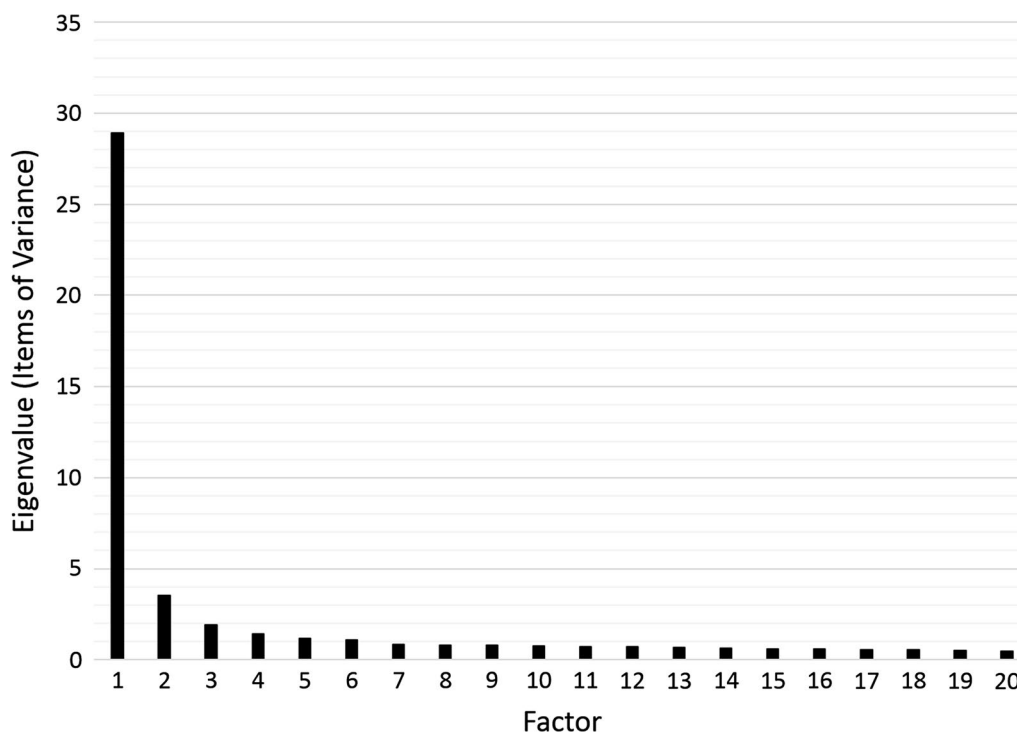

Fig. 1 Scree plot of eigenvalues for exploratory factor analysis on items from the MATE, I-SEA, and GAENE

Table 1 Two-dimensional factor loadings and Rasch item difficulty and mean squares fit statistics for both 1-dimensional and 2-dimensional usages

Item	Wording	EFA dimensionality		1D	1D	1D	2D	2D	2D
		F1 loading	F2 loading	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit
ISEAMacro1	P	0.87	− 0.03	− 0.17	0.67	0.63	− 0.15	0.70	0.68
ISEAMacro2[a]	N	0.12	0.62	0.15	1.25	1.72	0.06	1.05	1.16
ISEAMacro3	P	0.70	0.15	− 0.33	0.62	0.62	− 0.33	0.72	0.75
ISEAMacro4	P	0.77	− 0.05	0.39	0.95	1.07	0.53	1.05	1.17
ISEAMacro5	P	0.72	0.08	− 0.32	0.75	0.78	− 0.33	0.83	0.88
ISEAMacro6	N	0.23	0.58	0.28	0.93	0.95	0.19	0.82	0.84
ISEAMacro7	P	0.71	0.05	− 0.36	0.73	0.73	− 0.38	0.84	0.82
ISEAMacro8[b]	P	0.78	− 0.13	0.27	1.14	1.29	0.39	1.23	1.37
ISEAMicro1[a]	N	− 0.17	0.83	− 0.18	1.22	1.39	− 0.31	0.92	1.15
ISEAMicro2	P	0.61	0.04	− 0.52	0.96	1.03	− 0.57	1.12	1.17
ISEAMicro3[b]	P	0.37	0.24	− 0.68	1.07	0.92	− 0.77	1.32	1.17
ISEAMicro4[a]	N	− 0.26	0.92	− 0.24	1.44	1.74	− 0.36	0.97	1.02
ISEAMicro5	N	− 0.06	0.82	− 0.26	1.11	1.10	− 0.39	0.83	0.95
ISEAMicro6	P	0.50	0.15	− 0.34	0.89	0.91	− 0.35	1.06	1.06
ISEAMicro7[a]	N	− 0.04	0.78	− 0.39	1.33	1.30	− 0.52	0.97	0.91
ISEAMicro8	P	0.72	0.07	− 0.14	0.76	0.72	− 0.11	0.83	0.79
ISEAHuman1	P	0.81	0.01	0.19	0.78	0.80	0.29	0.81	0.84
ISEAHuman2	N	0.24	0.58	0.46	1.08	1.12	0.38	1.01	1.01
ISEAHuman3	N	0.23	0.61	0.26	1.04	1.03	0.18	0.90	0.90
ISEAHuman4	P	0.85	− 0.01	0.17	0.95	0.95	0.27	0.99	0.99
ISEAHuman5	P	0.73	0.08	− 0.02	1.01	0.98	0.03	1.18	1.17
ISEAHuman6	N	0.11	0.67	0.09	1.15	1.23	− 0.01	0.95	1.04
ISEAHuman7	P	0.84	− 0.02	0.08	0.90	0.95	0.16	0.92	0.96
ISEAHuman8	P	0.63	− 0.10	− 0.15	1.00	1.08	− 0.11	1.15	1.18
MATEfacts1	P	0.82	0.07	− 0.13	0.68	0.62	− 0.09	0.71	0.65
MATEcred2	N	0.07	0.65	0.27	1.21	1.22	0.18	1.02	1.06
MATEfacts3	P	0.87	− 0.01	0.09	0.80	0.75	0.18	0.84	0.79
MATEcred4	N	0.12	0.66	0.26	1.08	1.17	0.16	0.94	0.97
MATEcred5[b]	P	0.46	0.05	− 0.50	1.12	1.23	− 0.56	1.33	1.39
MATEcred6	N	0.28	0.50	0.46	0.91	1.00	0.40	0.85	0.90
MATEcred7[a]	N	0.07	0.59	− 0.31	1.48	1.46	− 0.43	1.24	1.23
MATEfacts8	P	0.73	0.09	− 0.15	0.68	0.65	− 0.13	0.77	0.74
MATEcred9[a]	N	0.03	0.73	− 0.07	1.29	1.47	− 0.18	0.99	0.99
MATEcred10	N	0.21	0.63	− 0.03	1.00	1.02	− 0.13	0.90	0.87
MATEfacts11	P	0.64	0.10	− 0.15	1.05	1.06	− 0.12	1.21	1.16
MATEfacts12	P	0.70	0.09	0.04	0.69	0.72	0.11	0.78	0.94
MATEfacts13	P	0.79	− 0.01	− 0.04	0.71	0.84	0.01	0.76	0.91
MATEcred14	N	0.14	0.69	0.13	1.22	1.26	0.03	1.04	1.03
MATEfacts15[a]	N	0.10	0.69	0.20	1.28	1.32	0.11	1.02	1.01
MATEfacts16	P	0.73	0.10	0.11	0.78	0.76	0.20	0.85	0.84
MATEcred17[a,b]	N	− 0.22	0.69	0.22	1.58	1.79	0.12	1.31	1.53
MATEfacts18	P	0.80	0.06	− 0.11	0.61	0.59	− 0.07	0.66	0.63
MATEcred19[a,b]	N	− 0.05	0.56	0.58	1.43	2.01	0.52	1.25	1.54
MATEfacts20	P	0.64	0.24	− 0.01	0.72	0.70	0.05	0.87	0.88
GAENE1[a,b]	P	0.60	− 0.03	− 0.24	1.23	1.36	− 0.23	1.42	1.52
GAENE2	P	0.66	0.14	− 0.03	0.72	0.74	0.02	0.84	0.85
GAENE3[b]	P	0.54	0.07	− 0.70	1.04	1.28	− 0.80	1.21	1.36

Romine *et al. Evo Edu Outreach* (2018) 11:17

Page 9 of 20

Table 1 (continued)

Item	Wording	EFA dimensionality		1D	1D	1D	2D	2D	2D
		F1 loading	F2 loading	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit
GAENE4	P	0.81	0.01	− 0.27	0.96	0.89	− 0.27	1.03	0.96
GAENE5	P	0.69	0.02	− 0.51	0.98	0.98	− 0.57	1.08	1.08
GAENE6[b]	P	0.73	− 0.05	0.84	1.20	1.29	1.08	1.28	1.51
GAENE7	P	0.53	0.05	− 0.73	1.06	1.06	− 0.83	1.23	1.21
GAENE8[a,b]	P	0.79	− 0.19	0.92	1.18	1.50	1.18	1.30	1.69
GAENE9	P	0.81	− 0.03	0.15	0.84	0.85	0.24	0.89	0.87
GAENE10	P	0.73	0.08	0.52	1.03	1.05	0.69	1.11	1.17
GAENE11	P	0.91	− 0.07	0.36	0.92	0.92	0.49	0.93	0.94
GAENE12	P	0.78	− 0.07	0.71	1.09	1.15	0.93	1.15	1.19
GAENE13	P	0.79	0.08	− 0.10	0.69	0.64	− 0.07	0.73	0.68

[a] Potential item misfit with the Rasch model in a unidimensional treatment

[b] Potential item misfit with the Rasch model in a two-dimensional treatment

Table 2 The highest loading items onto the first (F1) and second (F2) dimensions

Name	F1	F2	Wording	Stem
GAENE11	0.91	− 0.07	P	Evolution is a good explanation of how humans first emerged on the earth
ISEAMacro1	0.87	− 0.03	P	I think that new species evolved from ancestral species
MATEfacts3	0.87	− 0.01	P	Modern humans are the product of evolutionary processes that have occurred over millions of years
ISEAHuman4	0.85	− 0.01	P	I think that humans and apes share an ancient ancestor
ISEAHuman7	0.84	− 0.02	P	The many characteristics that humans share with other primates (i.e., chimpanzees, gorillas) can best be explained by our sharing a common ancestor
ISEAMicro4	− 0.26	0.92	N	Species were created to be perfectly suited to their environment, so they do not change
ISEAMicro1	− 0.17	0.83	N	I think that organisms, as they exist now, are perfectly adapted to their natural environments and so will not continue to change
ISEAMicro5	− 0.06	0.82	N	I don't accept the idea that a species of organism will evolve new traits over time
ISEAMicro7	− 0.04	0.78	N	Species exist today in exactly the same shape and form in which they always have
MATEcred9	0.03	0.73	N	Organisms exist today in essentially the same form in which they always have

The five items which load most highly onto the first dimension (loadings between 0.84 and 0.91) address acceptance of evolution as a mechanism for the origin of life as we know it today. Four out of these five items address human evolution specifically, meaning that a student's score on this acceptance sub-construct may be influenced most highly by their acceptance of human evolution. For example, item GAENE11 is the highest loading item (loading = 0.91), and states: *Evolution is a good explanation for how humans emerged on Earth.* Item MATEfacts3 is quite similar, stating: *Modern humans are the product of evolutionary processes that have occurred over millions of years.* ISEAhuman4 and 7 solicit agreement with the idea that humans and apes share a common ancestor. And ISEAmacro1 asks for agreement with the idea that *new species evolved from ancestral species.* In sum, these items solicit appraisal of evolution as a true, useful, and valid theory. It is interesting that the highest loading items address human evolution. Conceptually,

a factor loading for an item is derived from a regression equation where a student's measure for a latent construct or dimension is a predictor of his/her response for a certain item. This addresses the question: how well do students' measures for a certain construct explain their responses for an item defining that construct? Based on this conception, the high factor loadings for the human evolution items indicate that students' measures for this construct do very well in predicting how they will respond on these items. This implies that students who accept evolution as true, useful, and valid will tend to report very high acceptance for these items, whereas students who reject evolution will tend to report very low acceptance with high predictability. From here on, we label the construct underlying this dimension: *acceptance of the truth of evolution.*

After reverse coding, the five items which load most highly onto the second dimension (loadings between 0.73 and 0.92) address rejection of incredible or false

ideas underlying hesitancy against the theory of evolution. These five highest loading items get at the idea that species do not change. For example, the highest loading item (loading = 0.92), ISEAmicro4, states: *Species were created to be perfectly suited to their environment, so they do not change.* ISEAmicro7 and MATEcred9 are worded nearly identically, stating that organisms (or species) *exist in the same form they always have.* ISEAmicro5 solicits non-acceptance of the idea that organisms will *evolve new traits over time.* Finally, ISEAmicro1 addresses a possible root for these above conceptions: that organisms are currently *perfectly suited to their environments, and so do not change.* Acceptance of change (or in this case, rejection of the idea that organisms do not change over time), is required in order to find the theory of evolution credible and useful. While these represent the highest loading items, a vast majority of the negatively worded items present false ideas that must be rejected in order to find the theory of evolution credible and useful. We therefore label the construct underlying this dimension: *rejection of incredible ideas about evolution.*

Rasch analysis
1-dimensional usage
Rasch analysis gives compelling evidence for the usefulness of these items as a unidimensional measure, but as with the EFA, also suggests that a two-dimensional construction may increase the usefulness of the measure. When the 57 items are treated as measuring a single factor, *acceptance of evolution*, they yield measures with a reliability of 0.98 (separation = 6.71). Further, 46 out of the 57 items (81%) fit the Rasch model well, indicating that most of the items provide useful measures for *acceptance of evolution* as a single dimension. Five of these items displayed at least one fit index (mean squares infit or outfit) above 1.50, while the other six items displayed a fit index above 1.30.

The eigenvalue of the first factor derived from PCA on the residuals with respect to this model contains 2.11 items of variance. This is very close to 2, but nonetheless suggests a small departure from unidimensionality (Linacre and Tennant 2009). A closer look at items loading onto this factor corroborates our EFA results that whether the item is positively or negatively worded is to blame for this departure from unidimensionality. Inspection of Table 3 shows that sixteen negatively-worded items have positive loadings of 0.2 or above onto this residual factor, and also tend to be the items which misfit the Rasch model. On the other hand, seven items display negative loadings of 0.2 or above onto this residual factor. These items display positive wording and good fit with the Rasch model. In summary, while one can argue that the MATE, I-SEA, and GAENE are useful in measuring

evolution acceptance as a single construct, introduction of negative wording into the MATE and I-SEA gets at a different aspect of evolution acceptance than the positively worded items. Instead of measuring acceptance that evolution is true, the negatively worded items focus on ability or willingness to reject incredible ideas about evolution which lead one to not accept the theory. For example, those students scoring high (after reverse coding) on MATEcred19 (infit = 1.48, outfit = 2.01) reject the idea that *organisms came into existence at about the same time.* Those with a high score on ISEAmicro4 (infit = 1.44, outfit = 1.74) reject the idea that *species were created to be perfectly suited to their environment, and therefore do not change.* Those scoring high on MATEcred17 (infit = 1.58, outfit = 1.79) reject the idea that *much of the scientific community doubts if evolution occurs.* That these items misfit the Rasch model indicates that students' acceptance of evolution as a single latent variable does not explain the response patterns on these items.

2-dimensional usage
When the 57 items are broken into two separate subconstructs consisting of the 38 positively worded and 19 negatively worded items, respectively, excellent reliability is maintained and the measures become unidimensional. The 38 positively-worded items provide a measure for *acceptance of the truth of evolution* with a Rasch person reliability of 0.97 (separation = 5.95). The first eigenvalue from PCA on residuals from the Rasch rating scale model fitting these items consists of 1.51 items of variance, which is well below 2. The 19 negatively worded items generate measures for *rejection of incredible ideas about evolution* with a Rasch person reliability of 0.94 (separation = 4.09). The Rasch model also shows that this scale is unidimensional, with a first eigenvalue of 1.15 items of variance from PCA on the residuals with respect to the model.

Seven of the 38 items measuring *acceptance of the truth* display at least one mean squares fit index of 1.30 or above. Those items with the greatest misfit (a mean squares fit index above 1.50) come from the GAENE (items GAENE1, GAENE6, and GAENE8 from GAENE 2.1). GAENE1 states *everyone should understand evolution.* GAENE6 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.* GAENE8 states *nothing in biology makes sense without evolution.* These are of moderate-to-high difficulty, indicating that even students who accepted that evolution is true tended to mark lower levels of acceptance on these items. These items are getting at other factors outside of acceptance of the truth of evolution.

Only two of the 19 items measuring *rejection of incredible ideas* displayed significant misfit with the Rasch model. These items come from the MATE (items MATEcred17 and MATEcred19), and display mean squares fit indices greater than 1.50. MATEcred17 states *much of the scientific community doubts if evolution occurs* and MATEcred19 states *with few exceptions, organisms came into existence about the same time.* These items are also of moderate-to-high difficulty in comparison with the other items measuring this construct, indicating that even students who tend to reject the other erroneous or incredible ideas, thus getting high measures on the *rejection of incredible ideas* construct, tend to accept these ideas.

Discussion

Parametrizing evolution acceptance

The purpose of this study was to examine the measurement properties of the items contained in three quantitative evolution acceptance instruments, each of which captures evolution acceptance differently in terms of dimensionality and context. The MATE was designed as a unidimensional measure which used macroevolution and human evolution contexts in its items (Rutledge and Warden 1999). The I-SEA was developed to capture these contexts in separate constructs; it was designed to measure acceptance of microevolution, macroevolution, and human evolution as three respective dimensions (Nadelson and Southerland 2012). The GAENE was designed to provide a unidimensional measure of evolution acceptance which is independent of knowledge of evolution and religious orientation (Smith et al. 2016). Given the high similarity of wording between many of the items on the three instruments and the fact that they share a common Likert measurement scheme, we hypothesized that these instruments may share more similarities than differences and actually provide similar quantitative information about evolution acceptance. We found that this is the case. Putting the 57 items from the three instruments together to form an instrument-independent scaling results in useful unidimensional and two-dimensional parametrizations of evolution acceptance.

Given the two parametrizations, a big-picture question that comes forth immediately is: should evolution acceptance be treated unidimensionally as it has in the vast majority of quantitative work on evolution acceptance, or should it be treated as a two-dimensional construct which is apparently delineated by the positive or negative orientation of the wording of the items (also found in a recent study by Romine et al. 2017)? Our position is that a unidimensional usage of these items may provide a useful measure for evolution acceptance. Evidence for this includes: (1) the high variance of the first dimension

in the EFA (28.9 items of variance out of 57 total items) in comparison to the other dimensions (Fig. 1), (2) the closeness of the first eigenvalue from PCA on Rasch residuals to 2 for a unidimensional construction (2.11 items of variance), and (3) the moderate relationship between *acceptance of truth* and *rejection of incredible ideas* ($r = 0.73$). In fact Metzger et al. (2018) found that much of the apparent two-dimensionality of the MATE found by Romine et al. (2017) disappears in more advanced science students. However, deciding not to recognize evolution acceptance as a two-dimensional construct may also result in a missed opportunity—CFA and Rasch analysis demonstrate that a unidimensional construction does not capture the entire story regarding how these three instruments measure evolution acceptance.

Our analyses collectively (Tables 1, 2, 3) suggest that observing a conceptual distinction between positively- and negatively-worded items is empirically useful, and we argue that recognizing the distinction between *acceptance of truth* and *rejection of incredible ideas* is supported by conceptual work around evolution acceptance

Table 3 Items with factor loadings above 0.2 on the first residual dimension with respect to a unidimensional Rasch model for the collective I-SEA, MATE, and GAENE 2.1 item corpus

Item	Loading[a]	Infit	Outfit	Wording
ISEAmicro4	0.34	*1.44*	*1.74*	Negative
ISEAmicro7	0.32	*1.33*	*1.30*	Negative
ISEAmicro5	0.30	1.11	1.10	Negative
MATEcred17	0.30	*1.58*	*1.79*	Negative
MATEcred9	0.29	1.29	*1.47*	Negative
ISEAmicro1	0.27	1.22	*1.39*	Negative
ISEAmacro6	0.26	0.93	0.95	Negative
MATEfacts15	0.26	1.28	*1.32*	Negative
MATEcred7	0.25	*1.48*	*1.46*	Negative
ISEAhuman6	0.24	1.15	1.23	Negative
MATEcred19	0.24	*1.43*	*2.01*	Negative
MATEcred2	0.24	1.21	1.22	Negative
MATEcred14	0.23	1.22	1.26	Negative
ISEAhuman3	0.20	1.04	1.03	Negative
ISEAmacro2	0.20	1.25	*1.72*	Negative
ISEAhuman2	0.20	1.08	1.12	Negative
ISEAhuman4	− 0.32	0.95	0.95	Positive
GAENE11	− 0.27	0.92	0.92	Positive
MATEfacts1	− 0.22	0.68	0.62	Positive
ISEAhuman7	− 0.22	0.90	0.95	Positive
MATEfacts3	− 0.22	0.80	0.75	Positive
ISEAmacro8	− 0.21	1.14	1.29	Positive
ISEAhuman1	− 0.21	0.78	0.80	Positive

Mean squares fit statistics of 1.30 or higher are indicated in italics

[a] First residual factor

Romine *et al. Evo Edu Outreach*   (2018) 11:17

Page 12 of 20

(Deniz et al. 2008; Ha et al. 2012). For example, Deniz et al. (2008) propose a model in which evolution acceptance is constructed by an integration of cognitive, affective, and contextual factors. Further, they nest evolution acceptance within the theory of conceptual change (Strike and Posner 1992) in their suggestion that acceptance of evolution is necessary in order to find new ideas about evolution plausible and fruitful (Deniz et al. 2008). Our findings in this study not only support this assertion, but also allow us to expand upon it in proposing that the dimensionality by item valence aligns with specific elements of conceptual change theory. The theory of conceptual change is rooted in the idea that students construct new ideas through agreement based on their existing knowledge, experience, and social interactions (Posner et al. 1982). This implies that different individuals may interpret evolutionary phenomena in different ways based on their prior knowledge, metaphysical beliefs, and sociocultural background (Hewson 1981). It follows that one mechanism for learning involves agreement to abandon previously-held ideas and replace them with scientifically acceptable ideas (Posner et al. 1982). For a learner to agree to do this, he/she must find dissatisfaction in a previously-held idea and henceforth find the new idea more intelligible, plausible, and fruitful than the previously-held idea (Posner et al. 1982; Treagust and Duit 2008). Our first sub-construct, *acceptance of the truth of evolution,* aligns with the intelligibility and plausibility conditions of conceptual change, similar to that suggested by Deniz and colleagues. A student's measure for *rejection of incredible ideas about evolution* aligns directly with a student's dissatisfaction with non-scientific ideas and in turn the likelihood that he/she will find a scientifically-acceptable idea fruitful. *Rejection of incredible ideas* therefore more directly influences the likelihood that instruction will be successful in facilitating conceptual change. Although we did not measure knowledge of evolution in this study, previous work (Romine et al. 2017) supports this assertion in its finding that the correlation of acceptance of the credibility of evolution (aligned closely with *rejection of incredible ideas* in this study) with knowledge of macroevolution was significantly higher than the correlation between acceptance of facts about evolution (aligned closely with *acceptance of truth* in this study) with knowledge of macroevolution.

We also find that the two-dimensional parametrization aligns well with the more recent model of Ha et al. (2012), which describes evolution acceptance as manifesting through both conscious and unconscious neurological pathways. Our *acceptance of truth* dimension aligns most closely with the unconscious pathway which is based upon a student's intuitive feeling of certainty. 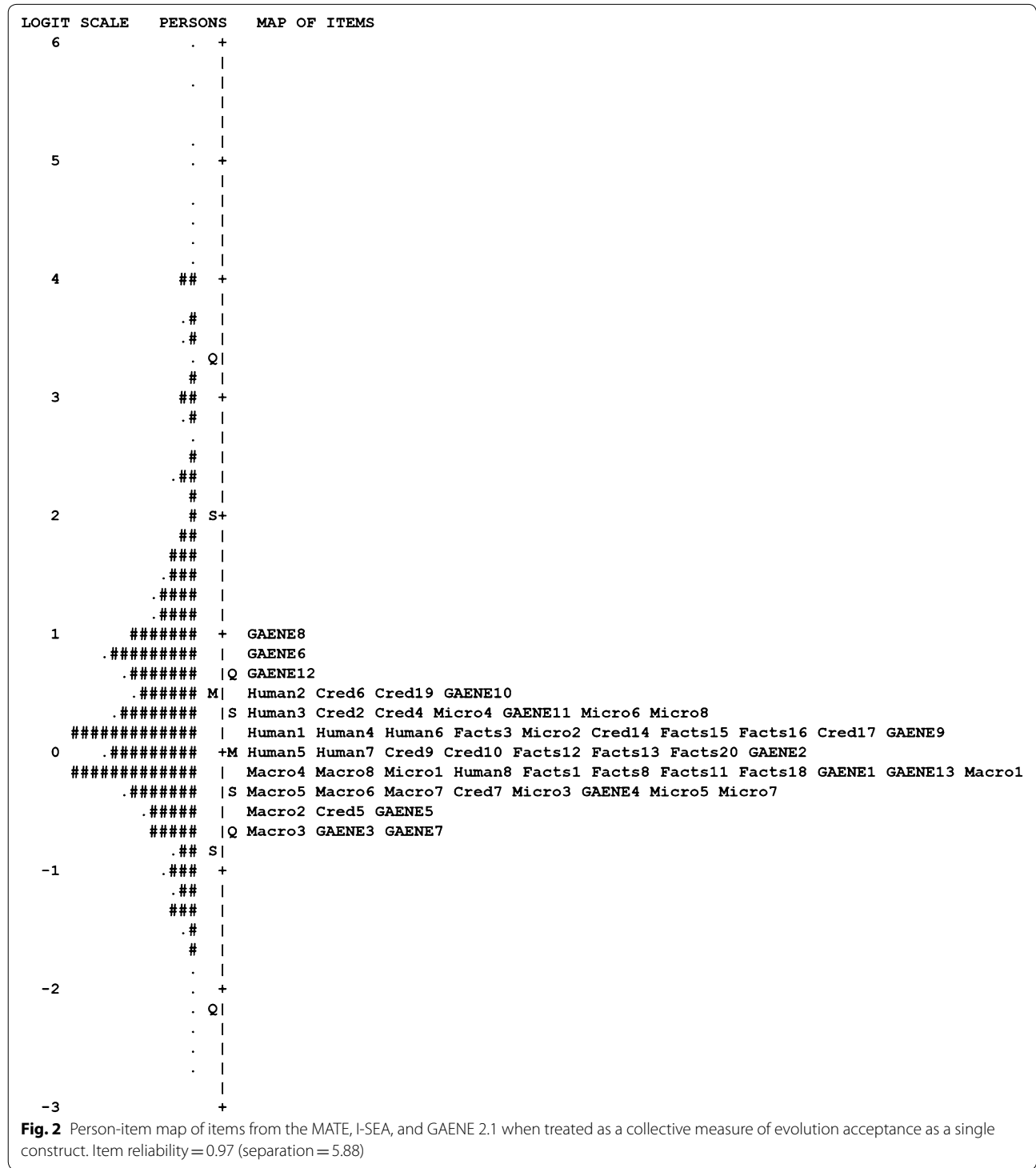For example, *evolution is a good explanation for how humans first emerged on the earth* (from the GAENE) solicits an appraisal based on one's feeling of certainty built from extracurricular experience, and not necessarily analysis of the credibility of an idea based on logic. On the other hand, the *rejection of incredible ideas* dimension aligns most closely with the student's conscious, reflective thinking supported by his/her process of understanding and logical reasoning around conceptual ideas underlying evolutionary theory. For example, expressing disagreement with the statement, *species were created to be perfectly suited to their environment, so they do not change* (from the I-SEA) requires a student to first consider the specific idea and then decide to reject that idea. Although the decision to reject an idea is affective in nature, it is in itself a comparatively logical and reflective process.

In addition to understanding how our current measures of evolution acceptance fit with conceptual work, this study also yields insight into the empirical implications for parametrizing evolution acceptance. Various parametrizations have been explored. Nadelson and Southerland (2012) utilized a three-dimensional model for evolution acceptance in their construction of the I-SEA, suggesting that the dimensions of evolution acceptance should be delineated by the type of evolution: microevolution, microevolution, and human evolution. What this study suggests is that, from a quantitative perspective, topic is a determinant of the difficulty of an item along the same sub-construct (Figs. 3 and 4), but it does not seem to serve as the key delimiter in terms of the unique sub-constructs. In other words, while acceptance of macroevolution, human evolution, and microevolution may be distinct in their difficulty, it may not be necessary to treat them as distinct sub-constructs. Rather, the data show that differences between students' responses on items across contexts are accounted for by the expected difficulty hierarchy imposed by the Rasch model (Boone 2016), making it unnecessary to define new sub-constructs to account for the different response patterns across contexts.

Previous literature supports the idea that acceptance of evolution varies depending on context (i.e. Scott 2005; Nadelson and Hardy 2015). Figures 3 and 4 suggest that this comprises a progression of acceptance similar in structure to what we have seen in research applying Rasch models to learning progressions (Romine et al. 2016; Todd et al. 2017; Todd and Romine 2016). From Figs. 3 and 4, it appears that undergraduate students look at evolution acceptance as the act of accepting that evolution is true and rejecting incorrect ideas, and that within these constructs, students tend to progress from accepting evolution on short time scales, then on long time scales, and then ultimately that evolution happens in humans.

An additional motivation for constructing the I-SEA by the topics of microevolution, macroevolution, and human evolution is that students who are reluctant to accept that organisms, including humans, evolve on long time scales may nonetheless accept that organisms evolve on short, observable time scales. The Wright map for the unidimensional construction (Fig. 2) shows that the items addressing human evolution are indeed the most difficult. However, the items at the bottom of the scale are macroevolution items, and as we go up the

```
LOGIT SCALE    PERSONS    MAP OF ITEMS
   6               .   +
                       |
                   .   |
                       |
                       |
                   .   |
   5               .   +
                       |
                   .   |
                   .   |
                   .   |
                   .   |
   4              ##   +
                       |
                  .#   |
                  .#   |
                  .  Q |
                   #   |
   3              ##   +
                  .#   |
                   .   |
                   #   |
                 .##   |
                   #   |
   2               #  S+
                  ##   |
                 ###   |
                .###   |
               .####   |
               .####   |
   1          ######   +  GAENE8
            .########  |  GAENE6
             .#######  |Q GAENE12
              .##### M|   Human2 Cred6 Cred19 GAENE10
             .####### |S  Human3 Cred2 Cred4 Micro4 GAENE11 Micro6 Micro8
          #############|   Human1 Human4 Human6 Facts3 Micro2 Cred14 Facts15 Facts16 Cred17 GAENE9
   0     .######### +M  Human5 Human7 Cred9 Cred10 Facts12 Facts13 Facts20 GAENE2
         ############|   Macro4 Macro8 Micro1 Human8 Facts1 Facts8 Facts11 Facts18 GAENE1 GAENE13 Macro1
          .####### |S  Macro5 Macro6 Macro7 Cred7 Micro3 GAENE4 Micro5 Micro7
           .##### |   Macro2 Cred5 GAENE5
           ##### |Q Macro3 GAENE3 GAENE7
           .## S|
  -1       .### +
           .## |
          ### |
           .# |
           # |
           . |
  -2       . +
           . Q|
           . |
           . |
           . |
             |
  -3         +
```

Fig. 2 Person-item map of items from the MATE, I-SEA, and GAENE 2.1 when treated as a collective measure of evolution acceptance as a single construct. Item reliability = 0.97 (separation = 5.88)

```
    LOGIT SCALE    PERSONS   MAP OF ITEMS
        7            .#   +
                          |
                     #    |
                          |
                          |
        6                 +
                     .    |
                          |
                     .    |
                     .    |
        5            .    +
                     .#   |
                          |
                     .    |
                     .##  |
        4                Q+
                     .#   |
                     .#   |
                     ##   |
                     .##  |
        3             #   +
                      #   |
                      #   |
                     .## S|
                     .##   |
        2             .#   +
                    ####    |
                   #####     |
                 .######     |
                 .#######    |   GAENE8
        1       .########   +Q  GAENE6  GAENE12
                 #######     |
                .####### M|   Macro4  GAENE10
               .######### |S  GAENE11 Macro8
                .######## |   Human1 Human4 Human7 Facts3 Facts12 Facts16 GAENE9
        0       ########  +M  Human5 Facts1 Facts13 Facts18 Facts20 GAENE2 GAENE13
                ######### |   Micro8 Macro1 Human8 Facts8 Facts11 GAENE1 GAENE4
                .####### |S  Micro6 Macro3 Macro5  Macro7
                 .###### |   Micro2 Cred5 GAENE5
                 .##### |   Micro3 GAENE3 GAENE7
       -1        .### S+Q
                  ###  |
                 .##   |
                 .###  |
                  .#   |
       -2          #   +
                  .#   |
                  .#   |
                  .    |
                  .  Q|
       -3         .   +
                      |
                      |
                  .   |
                      |
       -4             +
```
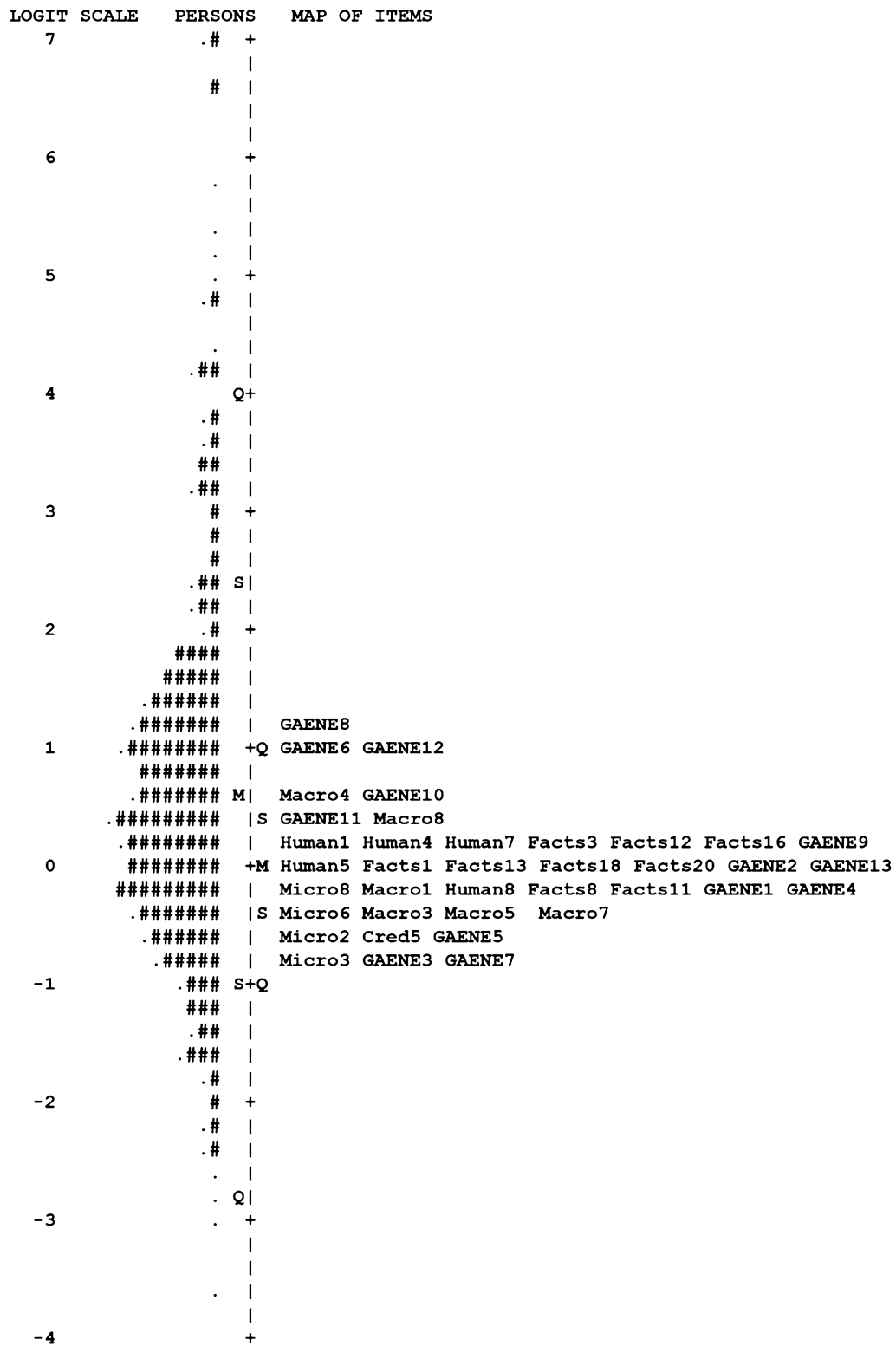
Fig. 3 Person-item map of the items from the MATE, I-SEA, and GAENE 2.1 measuring *acceptance of the truth of evolution*. Item reliability = 0.98 (separation = 6.97)
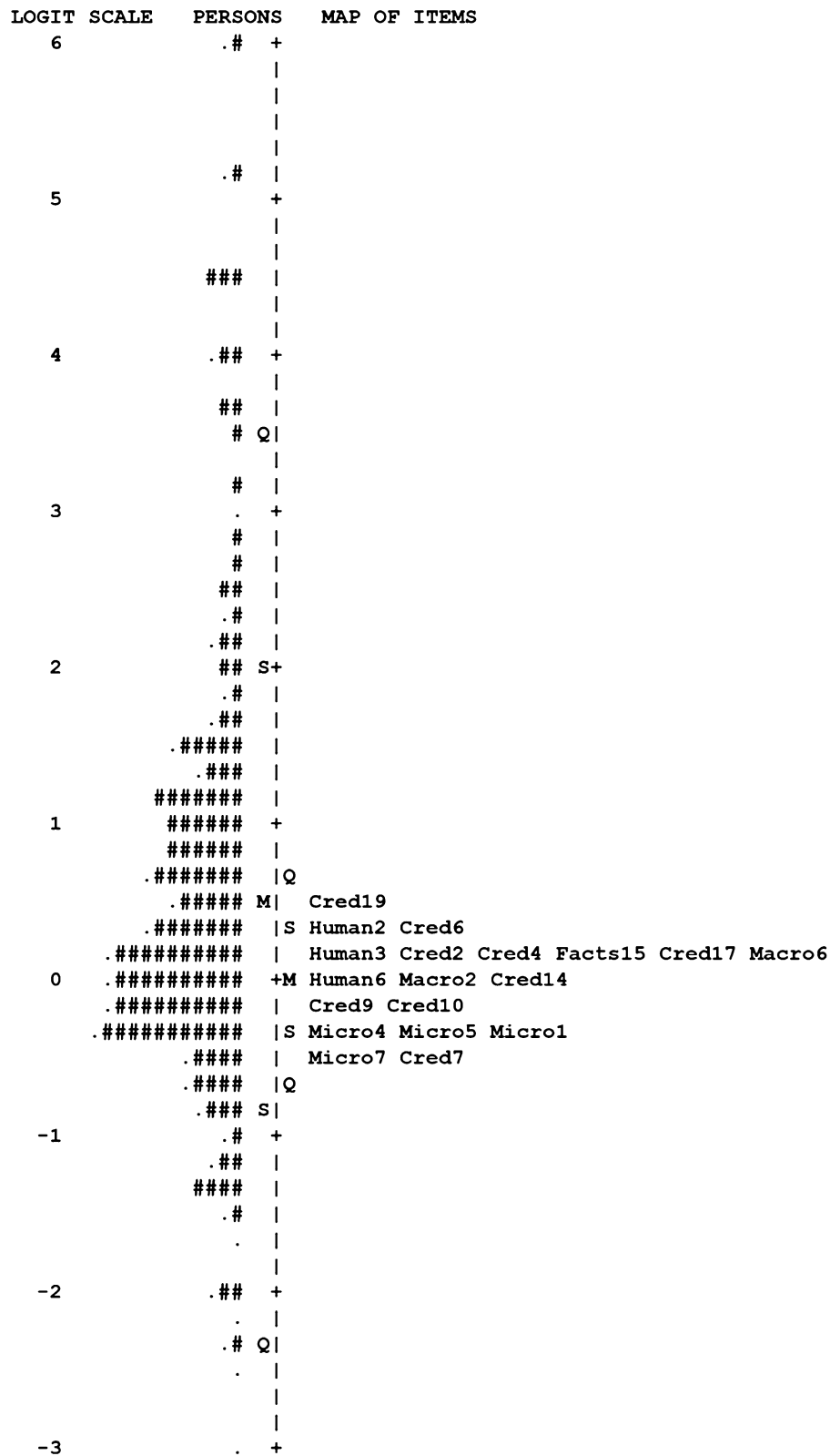
```
        LOGIT SCALE    PERSONS    MAP OF ITEMS
            6             .#    +
                                |
                                |
                                |
                                |
                          .#    |
            5                   +
                                |
                                |
                         ###    |
                                |
                                |
            4            .##    +
                                |
                          ##    |
                           #   Q|
                                |
                           #    |
            3              .    +
                           #    |
                           #    |
                          ##    |
                          .#    |
                         .##    |
            2             ## S+
                          .#    |
                         .##    |
                       .#####   |
                        .###    |
                       #######  |
            1           ######  +
                        ######  |
                      .#######  |Q
                       .##### M|  Cred19
                      .####### |S Human2 Cred6
                    .########## |  Human3 Cred2 Cred4 Facts15 Cred17 Macro6
            0     .##########  +M Human6 Macro2 Cred14
                  .##########  |  Cred9 Cred10
                 .##########  |S Micro4 Micro5 Micro1
                       .####  |  Micro7 Cred7
                       .####  |Q
                       .### S|
           -1            .#   +
                         .##   |
                       ####    |
                         .#    |
                          .    |
                                |
           -2           .##    +
                          .    |
                         .#  Q|
                          .    |
                                |
                                |
           -3             .    +
```

Fig. 4 Person-item map of the items from the MATE and I-SEA measuring *rejection of incredible ideas about evolution*. Item reliability = 0.96 (separation = 4.65)

scale, we see a mix of macroevolution and microevolution items. Much of this muddiness clears up when we model the positively and negatively worded items as separate sub-constructs (Figs. 3, 4). A two-dimensional parametrization gives the expected result that the items addressing evolution over short times scales are easiest for students to accept, those addressing evolution over long time scales are more difficult for students to accept, and those addressing human evolution are most difficult for students to accept. We see this as evidence that treating these instruments, and evolution acceptance in general, two-dimensionally gives a more realistic account of how college students think as they decide whether or not to accept evolution.

Previous work with the MATE also supports the idea that evolution acceptance should be treated two-dimensionally. Despite the fact that the MATE has mostly been used as a unidimensional instrument, an analysis by Wagler and Wagler (2013) suggests that the MATE behaves multidimensionally, and that it can be difficult to predict which items load on which dimensions as participants and contexts change. Romine et al. (2017), on the other hand, suggest that the behavior of the MATE is quite predictable in undergraduate students, and put forth a two-dimensional parametrization for the MATE that is similar to that discovered independently in this study. The only item result that was not replicated is Item 5 on the MATE (Cred5). In our previous work, this was the only positively-worded item which loaded more strongly onto the *acceptance of credibility* than the *acceptance of facts* dimension. In this study, it loaded onto the *acceptance of truth* dimension, which is more in line with what we would expect given that it is a positively-worded item.

This multidimensionality of evolution acceptance as measured by the MATE and I-SEA, and lack of basis upon a predefined framework for evolution acceptance, were driving factors for construction of the GAENE, which its authors claim is a generalizable unidimensional measure of evolution acceptance (Smith et al. 2016). Our analysis suggests that the main factor making the GAENE unidimensional is that all of its items are positively-worded. This means that the GAENE measures *acceptance of truth* but not *rejection of incredible ideas.* With respect to its contribution to measurement of evolution acceptance, both Figs. 2 and 3 show that the spread of the difficulty of the items is greater than those of the other instruments, meaning that these items provide information about students of a greater ability range than the items from the MATE and I-SEA. While this is attractive from a Rasch modeling perspective (Boone 2016), we also observe that some of the more difficult items from the GAENE do not fit well with the Rasch model, meaning that even students

with high levels of acceptance tend to express low levels of acceptance on these items. To follow, we discuss reasons for the misfit of these and other items, and what this may tell us about how to improve our measures of evolution acceptance.

Misfitting items and implications for improving current measures

When the items from these three instruments are treated unidimensionally, we observe the pattern that it is the negatively-worded items which tend to misfit with the Rasch model, and also tend to load positively onto the residual factor (Tables 1, 3). Nine of the 19 negatively-worded items exhibit some misfit (a mean squares index over 1.3) with the Rasch model in this case. These results collectively illustrate the multidimensionality of the negatively-worded items under the assumption of a unidimensional model for evolution acceptance. It makes sense that the Rasch model would tend to model the positively-worded items more faithfully, and thus identify negatively-worded items as anomalous, since the 38 positively-worded items double the 19 items which are negatively-worded.

This apparent bias disappears when positively- and negatively-worded items are treated as separate dimensions as suggested by the EFA, making it straightforward to explore more precisely potential issues with the wording of particular items. On the *acceptance of truth* dimension, three items (GAENE1, GAENE6, and GAENE8) exhibited a mean squares fit index above 1.50, which is indicative that these items may be harmful to the validity of the scale as it is defined by the corpus of 38 items in this construct (Wright et al. 1994). It happens that these items are also of moderate-to-high difficulty. This suggests a tendency for higher accepting students to express lower levels of acceptance on these items despite expressing high levels of acceptance on the other items measuring this construct. GAENE8, the most difficult item (Fig. 3), states: *nothing in biology makes sense without evolution.* From an expert biologist's perspective, evolution may serve as the centerpiece for understanding biology, but a college student who accepts evolution may feel that topics like anatomy and physiology, microbiology, and cell biology can be understood adequately without first understanding evolution. The context is not there since evolution is often not covered in college courses addressing these subjects.

GAENE1, *everyone should understand evolution*, sits at the middle of the scale (Fig. 3). Along the logic of GAENE8, the misfit of GAENE1 is likely caused by responses from students who accept evolution, but nonetheless do not view it as a necessity for engaging in other courses of study or for advancing one's quality of life.

GAENE6 sits at the top of the Rasch difficulty scale along with GAENE8 (Fig. 3). This item 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.* Fear of public speaking is quite common (Harris et al. 2002), and it is straightforward to argue that one can display acceptance of evolution without extraversion. In addition to evolution acceptance this item measures willingness to engage in public discourse; the latter having little to do with evolution acceptance.

Two items within the *rejection of incredible ideas* construct misfit with the Rasch model: MATEcred17 and MATEcred19. These items were also of relatively high difficulty, indicating that students who tended to reject the other misconception statements nonetheless tended to accept these. Students may have accepted MATEcred17, *much of the scientific community doubts if evolution occurs,* since the term "scientific community" is quite broad. A student who has taken a variety of science classes has likely seen that evolution is not discussed in most of these classes—especially the physical sciences, but even in biology (Padian 2010). It would therefore make sense for students to accept this statement regardless of their tendency to reject other ideas which are not credible. The misfit of MATEcred19, *with few exceptions, organisms came into existence about the same time,* is more difficult to explain. We argue that the *rejection of incredible ideas* construct is comparatively aligned with one's understanding of evolution (cognition), so the tendency for high ability students to accept this idea may be reflective of the persistent misunderstanding of deep time and the evolutionary timeline—one of the most difficult concepts for students to comprehend (Rosengren et al. 2012). Common practice may be to simply eliminate these items from the corpus in future research. However, doing this would carry the consequence of eliminating a potentially important part of the construct. We caution against making assessment decisions based on a single number. Instead the nature of the specific construct of interest should also be considered when deciding which items to use from these three instruments.

Suggestions for using current evolution acceptance instrumentation

A key question that arises from the above discussion is, how should the MATE, I-SEA, and GAENE be used in research? Much of the previous discourse around these instruments has centered on the question: which instrument works best? The I-SEA was developed to make our measures of evolution acceptance more fine-grained (Nadelson and Southerland 2012). A key motivation for developing the GAENE was to overcome the epistemological inadequacies of the MATE and I-SEA (Smith

et al. 2016). Although we do not find anything inherently wrong with using a particular instrument, our data indicate that loyalty to a particular instrument is unnecessary. We instead suggest a reframing of the dialogue toward a discussion of the merits and limitations of particular items instead of the instruments themselves. For example, if an evaluation project calls for a focus on human evolution or macroevolution, then we suggest that relevant items from both the MATE and the I-SEA could be combined into a single scale. On a different project, a researcher may wish to obtain a measure that is less context dependent and more based on an individual student's understanding of what constitutes the theory. In this case, the pool of items may be drawn primarily from the GAENE along with the more context-independent items from the I-SEA and MATE. We would, however, recommend that researchers using the GAENE consider removing the items asking students to "…argue in favor of evolution…" unless improving a student's extraversion is an important part of the evaluation plan.

Limitations of this study and current instrumentation

We would like to conclude by acknowledging some limitations of this study and our current measures of evolution acceptance toward understanding evolution acceptance conceptually. When making conclusions about the dimensionality of current evolution acceptance measures, it is important to consider the population being measured. The focus of this study was science non-majors at the college level. This population is unique due to the large variation in the types of non-major students in college who take general education science courses. We would expect the two dimensional structure found in this study to be retained in other populations with high internal variation like middle or high school, or the general public. However, if these items were administered to science majors or science teachers, participants would likely give consistently high ratings on all of the items, and the apparent two-dimensionality observed in this study may become negligible. This pattern has been observed in a recent study (Metzger et al. 2018). At the other end of the spectrum, if these items were administered to a population which held tenacious anti-evolutionary views, the ratings may be consistently low, and the item corpus may also become more unidimensional. We leave exploration of these phenomena to future research. In light of future research, it is useful to keep in mind that we took an a posteori approach, meaning that we factorized the items from these three instruments using variable-centered clustering (Collins and Lanza 2013) without establishing a hypothesized measurement structure from the outset. Any measurement model will

Romine *et al. Evo Edu Outreach*   (2018) 11:17

Page 18 of 20

introduce a particular bias to validity analyses, and we recognize that conclusions regarding the usefulness of particular items may have been different if we had used these instruments as they had been used previously.

An additional limitation to the conclusions of this study is that while item valence generates measures which are two-dimensional, it is not clear whether or not this variation is useful for helping us measure evolution acceptance. Some researchers suggest that cognitive fatigue may be responsible for the pattern of dimensionality based on item valence (Smith et al. 2016). This hypothesis suggests that there is construct-irrelevant variation in the negatively-worded items which results from their being difficult to interpret. This is an alternative to our interpretation that positively- and negatively-worded items are measuring two distinct constructs which comprise unique and useful ways to look at evolution acceptance. These competing hypotheses have not been explored empirically in evolution education. Concern with negative wording is not new, and has surfaced in other fields. For example, negative wording in the Job Diagnostic Survey (Hackman and Oldham 1975) was found to create deviations from unidimensionality (Kulik et al. 1988), leading Schmitt and Stuits (1985) to suggest that a respondent's education level may play a role in this. This claim was later not supported using a structural equation modeling methodology (Cordery and Sevastos 1993). We similarly find the cognitive fatigue hypothesis to be an unlikely explanation for the multidimensionality of current evolution acceptance items for several reasons. First, although Smith et al. (2016) states: "…disagreeing with a negative item constitutes a double negative" (p. 1312), a majority of the negatively-worded items on the MATE and I-SEA are not written in double negatives. Secondly, there is little reason to believe that a negatively-worded item like *Species were created to be perfectly suited to their environment, so they do not change* (from the I-SEA) would be more difficult for an undergraduate student to interpret than a positively worded item like *I think new species evolved from ancestral species* (from the I-SEA). Just as there are currently no data to support the cognitive fatigue hypothesis, we also do not have data to support our favored hypothesis that these are unique and useful dimensions. Our models show that these dimensions are highly separable psychometrically, but we cannot necessarily conclude that this separateness is useful as opposed to a systematic measurement artifact. This said, we find the *acceptance of truth* and *rejection of incredible ideas* derived from our data to be intellectually satisfying given their alignment with the unique elements of conceptual change theory and the subsequent additional insight it adds to

Deniz et al's (2008) and Ha et al's (2012) frameworks discussed previously.

Upon undertaking this research, we did not know what type of solution to expect, which is why we started with an exploratory factor approach. Would the items be unidimensional? Would sub-constructs be delineated by type of evolution as in the I-SEA? Would positive or negative wording play a role? We expected that it might be a combination of these things. The clean two-dimensional factor structure delineated solely by item valence came as a surprise to us. Since this pattern has also come up in previous work on college students taking general education science courses (Romine et al. 2017), and to a lesser degree in more advanced students (Metzger et al. 2018), we feel it is no coincidence and deserves further exploration and discussion.

A more significant limitation of our current evolution acceptance instrumentation comes out of the way the information is solicited; namely, items comprise a student's evaluation of specific statements related to evolution, which Scherer (2005) defines as an action of appraisal. While it is interesting to understand how students appraise a statement related to the theory of evolution, we have not yet developed ways of going beyond appraisal toward analysis of students' emotional experiences, which are critical for understanding the actual feelings students harbor towards evolution and how these relate to the short-term decisions students make around learning evolution. Emotion goes beyond cognitive appraisal in that it manifests as bodily symptoms, motivation towards specific actions, and communication of one's actual behavioral intentions (Scherer 2005), which are related directly to the perceived relevance of a particular object or event to one's well-being (Frijda 1986). Although a good amount of work has been done towards constructing useful measures of evolution acceptance, students' emotions about learning in general, and particularly those related to learning the theory of evolution, remain largely unmeasured and ill-understood. Given that a student's feelings constitute a key component of contemporary theoretical models of evolution acceptance (Deniz et al. 2008; Ha et al. 2012), measurement of emotion is an important next step in understanding students' affect towards evolution.

Author details

[1] Department of Biological Sciences, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45431, USA. [2] Boonshoft School of Medicine, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45431, USA. [3] Department of Biology, California State University—Fresno, M/S SB 73, Fresno, CA 93740, USA.

Publisher's Note

References

Alters BJ, Alters S. Defending evolution in the classroom: a guide to the creation/evolution controversy. Burlington: Jones & Bartlett Learning; 2001.

Andrich D. An extension of the Rasch model for ratings providing both location and dispersion parameters. Psychometrika. 1982;47(1):105–13.

Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care. 2004;42(1):1–7.

Beard J, Nelson C, Nickels M. Evolution and the nature of science institutes (ENSI). 2014. Retrieved from http://www.indiana.edu/~ensiweb/home.html. Retrieved 21 Oct 2018.

Bentler PM. Comparative fit indexes in structural models. Psychol Bull. 1990;107(2):238–46.

Bentler PM, Bonnet DC. Significance tests and goodness of fit in the analysis of covariance structures. Psychol Bull. 1980;88(3):588–606.

Bond T, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. New York: Routledge; 2015.

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

Boone WJ, Scantlebury K. The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. Sci Educ. 2006;90(2):253–69.

Boone WJ, Staver JR, Yale MS. Rasch analysis in the human sciences. Berlin: Springer; 2013.

Brewer C, Smith D. Vision and change in undergraduate biology education: a call to action. Washington, DC: AAAS; 2011.

Bryant FB, Yarnold PR. Principal-components analysis and confirmatory factor analysis. In: Grimm LG, Yarnold PR, editors. Reading and understanding multivariate statistics. Washington, DC: American Psychological Association; 1995. p. 99–136.

Cattell RB. The scree test for the number of factors. Multivar Behav Res. 1966;1(2):245–76.

Collins LM, Lanza ST. Latent class and latent transition analysis: with applications in the social, behavioral, and health sciences, vol. 718. New York: Wiley; 2013.

Cordery JL, Sevastos PP. Responses to the original and revised job diagnostic survey: is education a factor in responses to negatively worded items? J Appl Psychol. 1993;78(1):141.

Corwin LA, Runyon C, Robinson A, Dolan EL. The laboratory course assessment survey: a tool to measure three dimensions of research-course design. CBE Life Sci Educ. 2015;14(4):37.

Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract Assess Res Eval. 2005;10(7):1–9.

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.

Donnelly LA, Kazempour M, Amirshokoohi A. High school students' perceptions of evolution instruction: acceptance and evolution learning experiences. Res Sci Educ. 2009;39:643–60.

Gallup G. Evolution, creationism and intelligent design. 2010. Retrieved from http://www.gallup.com/poll/21814/Evolution-Creationism-Intelligent-Design.aspx. Retrieved 18 Dec 2018.

Fabrigar LR, Wegener DT. Exploratory factor analysis. Oxford: Oxford University Press; 2011.

Frijda NH. The emotions. Cambridge: Cambridge University Press; 1986.

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.

Hackman JR, Oldham GR. Development of the job diagnostic survey. J Appl Psychol. 1975;60(2):159.

Harris S, Kemmerling RL, North MM. Brief virtual reality therapy for public speaking anxiety. Cyberpsychol Behav. 2002;5(6):543–50.

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). Evolution. 2011;4:117–32. https://doi.org/10.1007/s12052-010-0294-1.

Hendrickson AE, White PO. Promax: a quick method for rotation to oblique simple structure. Br J Math Stat Psychol. 1964;17(1):65–70.

Henson RK, Roberts JK. Use of exploratory factor analysis in published research: common errors and some comment on improved practice. Educ Psychol Measur. 2006;66(3):393–416.

Hewson PW. A conceptual change approach to learning science. Eur J Sci Educ. 1981;3(4):383–96.

Hogan K. Exploring a process view of students' knowledge about the nature of science. Sci Educ. 2000;84(1):51–70.

Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model. 1999;6(1):1–55.

Johnson RL, Peeples EE. The role of scientific understanding in college: student acceptance of evolution. Am Biol Teacher. 1987;49:93–6.

Kline P. An easy guide to factor analysis. New York: Routledge; 2014.

Kulik CT, Oldham GR, Langner PH. Measurement of job characteristics: comparison of the original and the revised job diagnostic survey. J Appl Psychol. 1988;73(3):462.

Lakatos I. Falsification and the methodology of scientific research programmes. Can theories be refuted?. Berlin: Springer; 1976. p. 205–59.

Lerner LS. Good science, bad science: teaching evolution in the states. Washington, D.C.: Thomas B. Fordham Foundation; 2000.

Linacre JM. WINSTEPS Rasch measurement computer program. Chicago: WINSTEPS.com; 2006.

Linacre JM, Tennant A. More about critical eigenvalue sizes in standardized-residual principal components analysis (PCA). Rasch Measu Trans. 2009;23(3):1228.

Lovely EC, Kondrick LC. Teaching evolution: challenging religious preconceptions. Integrat Comparat Biol. 2008;48:164–74.

Masci D. For Darwin Day, 6 facts about the evolution debate. 2017. Retrieved from http://www.pewresearch.org/fact-tank/2017/02/10/darwin-day/. Retrieved 3 Sept 2018.

Masters GN. Item discrimination: when more is worse. J Educ Meas. 1988;25(1):15–29.

Metzger KJ, Montplaisir D, Haines D, Nickodem K. Investigating undergraduate health sciences students' acceptance of evolution using MATE and GAENE. Evolution. 2018;11(1):10.

Muthén BO, Muthén LK. Mplus 7 base program. Los Angeles: Muthén & Muthén; 2012.

Nadelson LS, Hardy KK. Trust in science and scientists and the acceptance of evolution. Evolution. 2015;8(1):9.

Nadelson LS, Southerland SA. Examining the interaction of acceptance and understanding: how does the relationship change with a focus on macroevolution. Evolution. 2010;4:82–8.

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(11):1637–66.

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

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

Nehm RH, Schonfeld IS. Does increasing biology teacher knowledge of evolution and the nature of science lead to greater preference for the teaching of evolution in schools? J Sci Teacher Educ. 2007;18:699–723.

NGSS Lead States. Next generation science standards: for states, by states. Washington, DC: The National Academies Press; 2013.

Padian K. How to win the evolution war: teach macroevolution! Evolution. 2010;3(2):206–14.

Popper K. Philosophy of science. In: Mace CA, editor. British Philosophy in the Mid-Century. London: George Allen and Unwin; 1957.

Posner GJ, Strike KA, Hewson PW, Gertzog WA. Accommodation of a scientific conception: toward a theory of conceptual change. Sci Educ. 1982;66(2):211–27.

Robbins JR, Roy P. The natural selection: identifying & correcting non-science student preconceptions through an inquiry-based, critical approach to evolution. Am Biol Teacher. 2007;69:460–6.

Romine WL, Todd AN. Valuing evidence over authority: the impact of a short course for middle-level students exploring the evidence for evolution. Am Biol Teacher. 2017;79(2):112–9.

Romine WL, Barrow LH, Folk WR. Exploring secondary students' knowledge and misconceptions about influenza: development, validation, and implementation of a multiple-choice influenza knowledge scale. Int J Sci Educ. 2013;35(11):1874–901.

Romine WL, Todd AN, Clark TB. How do undergraduate students conceptualize acid–base chemistry? Measurement of a concept progression. Sci Educ. 2016;100(6):1150–83.

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. 2017;54(5):642–71.

Rosengren KS, Brem SK, Evans EM, Sinatra GM, editors. Evolution challenges: Integrating research and practice in teaching and learning about evolution. Oxford: Oxford University Press; 2012.

Rutledge ML, Sadler KC. Reliability of the measure of acceptance of the theory of evolution (MATE) instrument with university students. Am Biol Teacher. 2007;69(6):332–5.

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

Scherer KR. What are emotions? And how can they be measured? Soc Sci Inform. 2005;44(4):695–729.

Schmitt N, Stuits DM. Factors defined by negatively keyed items: the result of careless respondents? Appl Psychol Meas. 1985;9(4):367–73.

Scott EC. Evolution vs creationism: an introduction. Berkeley: University of California Press; 2005.

Short SD, Hawley PH. Evolutionary attitudes and literacy survey (EALS): development and validation of a short form. Evolution. 2012;5(3):419–28.

Smith MU. Current status of research in teaching and learning evolution: I. Philosophical/epistemological issues. Sci Educ. 2010;19(6–8):523–38.

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

Steiger JH. Understanding the limitations of global fit assessment in structural equation modeling. Person Individ Differ. 2007;42(5):893–8.

Strike KA, Posner GJ. A revisionist theory of conceptual change. In: Duschl R, Hamilton R, editors. Philosophy of science, cognitive psychology, and educational theory and practice. Albany: Suny; 1992. p. 147–76.

Todd A, Romine WL. Validation of the learning progression-based assessment of modern genetics in a college context. Int J Sci Educ. 2016;38(10):1673–98.

Todd A, Romine WL, Cook Whitt K. Development and validation of the learning progression-based assessment of modern genetics in a high school context. Sci Educ. 2017;101(1):32–65.

Treagust DF, Duit R. Conceptual change: a discussion of theoretical, methodological and practical challenges for science education. Cult Sci Educ. 2008;3(2):297–328.

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.

Wright BD. Comparing Rasch measurement and factor analysis. Struct Equ Model. 1996;3(1):3–24.

Wright BD, Linacre JM, Gustafson JE, Martin-Loff P. Reasonable mean square fit values. Rasch Meas Trans. 1994;8(3):370.