Evolution: Education and Outreach

REVIEW

Open Access

Applying measurement standards to evolution education assessment instruments

Louise S. Mead[1,2*], Cory Kohn[1,3], Alexa Warwick[1,4] and Kathryn Schwartz[1,2]

Abstract

Over the past 25 years a number of instruments have been published that attempt to measure understanding and acceptance of evolution. Science educators have been administering these instruments and reporting results, however, it is not clear these instruments are being used appropriately. The goal of this paper is to review these instruments, noting the original criteria and population for which evidence of validity and reliability was assessed, and to survey other publications that report their use, examining each for evidence of validity and reliability with subsequent populations. Our hope is that such a comprehensive review will engage researchers and practitioners in a careful examination of how they intend to use a particular instrument and whether it can provide an accurate and meaningful assessment of the desired outcomes. We encourage the community to administer evolution education assessments with the consideration of an instrument's measurement support and past use with similar populations. We also encourage researchers to add additional evidence of validity and reliability for these instruments, especially if modifications have been made to the instrument or if its use has been extended to new populations.

Keywords: Evolution education, Assessment, Natural selection, CINS, ACORNS, I-SEA, GAENE, CANS, MATE

Background

Evolution is both a foundational concept and organizing principle in biology and as such has secured a central place in biology education as evidenced by science education reforms (National Research Council 2012; Brownell et al. 2014). Yet, a disconnect still exists between the central role of evolution in biology, student understanding of evolutionary mechanisms, and the general level of public acceptance as measured by polling questions administered by organizations such as Gallop (Swift 2017) and Pew Research Center (Funk and Rainie 2015). To further complicate its teaching and learning, the various relationships between acceptance and understanding of evolution and the nature of science (Smith 2010a; Smith and Siegel 2004), along with religiosity and the use of teleological reasoning (Allmon 2011; Shtulman 2006), impact student understanding and potentially their ability to successfully integrate evolutionary concepts into their understanding of the biological world (Sinatra et al. 2003; Smith 2010b). In a recent study of the general public, Weisberg et al. (2018) found that knowledge of evolution predicted level of acceptance, possibly suggesting student views may be amenable to change. However, a different study suggests teleological reasoning and not acceptance of evolution influences understanding of natural selection (Barnes et al. 2017). The relationship between understanding and acceptance is complex, and while not addressed directly in this paper, it is important to be aware of this complexity when assessing students and evaluating instruments. The wording and content of an assessment can impact student responses if their acceptance hinders their ability to answer questions addressing understanding. There are a number of papers that provide extensive discussion of this particular challenge to teaching and learning evolution (Smith 2010a, b), however, we have not addressed this directly in our review of instruments aside from potential issues associated with a particular instrument based on our review criteria.

Educational research has also found that how a student responds to questions on the topic of evolution is context dependent, e.g. taxa, or the direction of change via trait

*Correspondence: lsmead@msu.edu
[1] BEACON Center for the Study of Evolution in Action, Michigan State University, 567 Wilson Rd, East Lansing, MI 48824, USA
Full list of author information is available at the end of the article

gain vs. loss (Nehm et al. 2012; Nehm and Ha 2011), and many students retain naive or non-scientific concepts even after instruction (Ha et al. 2015; Nehm and Reilly 2007). Given these findings, and the various challenges to student understanding of evolution (Branch and Mead 2008; Mead and Scott 2010a, b; Petto and Mead 2008), many science educators are now interested in assessing how well students understand, and in some cases, accept, the basic premise and mechanisms underlying evolutionary change, in either formative or summative ways. In addition, instructors seek to assess the effectiveness of curricular interventions designed to improve student understanding.

Perhaps as a result of recent interest in the teaching and assessment of evolution, or the growing field of discipline-based education research, a number of instruments designed to assess student understanding and acceptance of evolution have been created over the last 25 years (see Table 1 for examples). At the undergraduate biology level, these include, but are not limited to, assessments designed to measure student understanding of natural selection (e.g. concept inventory of natural selection—CINS, Bishop and Anderson 1990; concept assessment of natural selection—CANS, Kalinowski et al. 2016), macroevolution (e.g. measure of understanding of macroevolution—MUM, Nadelson and Southerland 2009); genetic drift (e.g. genetic drift inventory—GeDI, Price

Table 1 List of published instruments that measure understanding and/or acceptance of evolution reviewed in current paper

Instrument	Full name and brief description	Citation
ECT	*Evolution concept test* Six items, combination of open ended and Likert-scale type ***Natural selection***	Bishop and Anderson (1990)
CINS	*Concept inventory of natural selection* Twenty multiple choice questions ***Natural selection***	Anderson et al. (2002)
MATE	*Measure of acceptance of the theory of evolution* Twenty-five-point Likert questions ***Acceptance of evolution***	Rutledge and Warden (1999); Rutledge and Sadler (2007)
MUM	*Measure of understanding of macroevolution* Twenty-seven multiple choice & 1 free response ***Deep time, phylogenetics, speciation, fossils, nature of science***	Nadelson and Southerland (2009)
KEE	*Knowledge of evolution exam* Ten questions	Moore and Cotner (2009)
EALS-long and short forms	*Evolutionary Attitudes and Literacy Survey* Long form: 104 Likert-scaled questions Short form: 64 Likert-scaled questions ***Religiosity, science understanding and attitudes as relates to evolution***	Hawley et al. (2011) Short and Hawley (2012)
ACORNS	*Assessing contextual reasoning about nature selection* Unlimited number of open-ended questions ***Natural selection, non-adaptive change***	Nehm et al. (2012)
I-SEA	*Inventory of student acceptance of evolution* Twenty-four items ***Microevolution, macroevolution, human evolution***	Nadelson and Southerland (2012)
EvoDevoCI	No full title Eleven multiple choice questions ***Evolutionary developmental biology***	Perez et al. (2013)
ATEEK	*Assessment tool for evaluating evolution knowledge* Four open-ended questions ***Genotype, phenotype, change in allele frequencies***	White et al. (2013)
GeDI	*Genetic drift inventory* Twenty-two agree/disagree statements ***Genetic drift***	Price et al. (2014)
GAENE	*Generalized acceptance of evolution evaluation* Thirteen Likert items ***Evolution acceptance***	Smith et al. (2016)
CANS	*Concept assessment of natural selection* Twenty-four multiple choice ***Natural selection***	Kalinowski et al. (2016)

The italicized words correspond to the full name of the instrument, the bold italics refer to the content topic addressed by the instrument et al. 2014); and acceptance of evolution (e.g. Measure of the Acceptance of the Theory of Evolution—MATE, Rutledge and Warden 1999; Evolutionary Attitudes and Literacy Survey—EALS, Hawley et al. 2011; generalized acceptance of evolution evaluation—GAENE, Smith et al. 2016). These instruments can provide an opportunity for instructors to measure gains in student understanding; however, the conclusions drawn from them are dependent on the quality, accuracy, and relevancy of the instrument. For example, in a review of assessments addressing student understanding of bioinformatics concepts, Campbell and Nehm (2013) found many of the instruments they reviewed provided only minimal evidence of reliability or validity.

The decision to use any instrument should include an examination of the instrument and its development to ascertain if it meets the accepted measurement standards, specifically whether there is strong evidence that the instrument provides valid and reliable results. Evidence that an instrument provides valid results suggests the variable being measured by the instrument accurately represents the construct or item of interest. Evidence that an instrument provides reliable results suggests the instrument gives consistent results when implemented under similar circumstances. There are multiple forms of evidence for reliability (e.g. stability, internal consistency, interrater reliability) and validity (e.g. content, internal and external structure, generalization). Box 1 provides examples of the different sources of evidence that can be used to evaluate validity and reliability (Messick 1995; Campbell and Nehm 2013; AERA 2014).

Box 1. Methods and descriptions for various sources of validity and reliability (modified from Messick 1995; Campbell and Nehm 2013; AERA 2014)

Source	Description	Methodology (examples)
Validity—do scores represent the variable(s) intended?		
Content	Assessment represents knowledge domain	Expert survey, textbook analysis, Delphi Study
Substantive	Thinking processes used to answer are as intended	"Think aloud" interviews, cognitive task analysis
Internal structure	Items capture intended construct structure	Factor analysis, Rasch analysis
External structure	Construct aligns with expected external patterns	Correlational analysis
Generalization	Scores meaningful across populations and contexts	Comparisons across contextual diversity, Differential item functioning
Consequences	Scores lead to positive or negative consequences	Studying social consequences resulting from use of test score
Reliability—refers to the consistency of the measure		
Stability	Scores consistent from one administration to another	Stability coefficient
Alternate forms	Scores comparable when using similar items	Spearman-Brown double length formula: split half
Internal consistency	Items correlate with one another	Coefficient alpha (Cronbach's), Kuder-Richardson 20
Inter-rater agreement	Assessment scored consistently by different raters	Cohen's or Fleiss's kappa

Assessment of student understanding in educational settings should include systematic evaluation of instruments in order to meet the quality control benchmarks established by, for example, the American Educational Research Association (AERA et al. 2014). Not doing so is "at odds with the principles of scientific research in education" (Campbell and Nehm 2013) and since a reliance on faulty or misleading information for the purposes of evaluation and reform is misguided, it is therefore necessary to establish an assurance of such information's positive utility. Campbell and Nehm (2013) are careful to point out that validity and reliability are not properties of the instrument itself, but rather relate to the inferences derived from the scores it produces. It is therefore incorrect to describe an assessment instrument itself as being valid and reliable. Instead, our interpretation of validity and reliability needs to shift such that an assessments' scores and implementation contexts are foremost. For example, a correct statement is that the instrument produces valid and reliable inferences under the particular circumstances it was administered. One cannot assume that an instrument developed using a population of undergraduate non-majors in their 1st year of college necessarily has the same evidence of reliability and validity for a population of students in an upper level evolution course.

In our own efforts to identify ways of assessing understanding of evolutionary concepts, we found many studies simply reported using a published instrument, often modified from an earlier published instrument, and often lacking any additional information about the implementation or adherence to measurement standards. To address these issues, we (1) reviewed the various published instruments designed to measure understanding and acceptance of evolution, (2) examined the types of evidence of validity and reliability provided in the original publication(s), and (3) characterized the use of these instruments in subsequent publications, specifically noting any additional evidences of reliability and validity.

Methods

In 2016 and 2017 we (LM, CK, AW, KS) carried out searches of Google Scholar, ERIC, and Web of Science using the following keyword searches: "student understanding of evolution"; "student understanding of natural selection"; "student acceptance of evolution". We compiled a list of papers that referenced these key phrases, focusing on ones that were aimed at college undergraduates. We reviewed abstracts to identify papers that specifically mentioned measuring student understanding or acceptance of evolution using the following criteria: population—undergraduates; level/course—any; content assessed—evolution understanding, evolution acceptance, natural selection, genetic drift. If the information could not be readily assessed from the abstract, we examined the methods section of the paper in more detail. In this initial review of the published literature it became clear that many of the papers we reviewed referenced using some portion of an earlier published instrument or set of questions. For example, many studies reported using portions of the original assessment developed by Bishop and Anderson (1990). We used this information to identify a set of 13 instruments that would become the focus of the remainder of our research, and that appeared to form the basis of many studies.

The criteria for our more in-depth analysis of assessment instruments included instruments created with the intention of being used by others to assess understanding and acceptance of evolution. We made three exceptions to these criteria: the ECT referenced in Bishop and Anderson (1990), the KEE (knowledge of evolution exam) referenced in Moore and Cotner (2009), and the ATEEK (assessment tool for evaluating evolution knowledge) referenced in White et al. (2013). We chose to include these because they were subsequently treated as instruments by other researchers who used them as the basis of assessing student understanding. Two of these, the KEE and ATEEK, were given a specific name for use and referenced by others. We did not include instruments measuring genetics only or combinations of other biological sub-disciplines (e.g. EcoEvo-MAPS in Summers et al. 2018) because we wanted to evaluate only instruments reported to measure student understanding and/or acceptance of evolution. We also chose to exclude the topic of phylogenetics for a number of reasons. First, phylogenetic trees are visual representations of both patterns and processes, and therefore it can be difficult to isolate specific elements from a cognitive perspective (Novick and Catley 2012). Second, at the time of our review, the only published instruments included one provided in Baum et al. (2005), the Basic Tree Thinking Assessment, which was developed as a formative quiz and not meant to be used as an assessment instrument (pers. com.), and the PhAT (Phylogeny Assessment Tool) comprised only three questions (Smith et al. 2013), all related to a single phylogenetic tree.

Our final list included 13 focal instruments (Table 1). We first reviewed the original publication and characterized the instrument (i.e., content and population assessed, type and number of questions, how it was developed) and the evidence of reliability and validity described in the population. These original instruments were reviewed and discussed by all co-authors so as to ensure consistency.

Next, we performed a citation search for each of the focal instruments to generate a list of publications that cited the instrument, suggesting possible use. We performed these searches using Google Scholar, first performing a search of the original paper (e.g. Bishop and Anderson 1990) and then examining all of the papers listed as "cited by" (e.g. at the time of our search Google Scholar reported 703 papers had cited Bishop and Anderson 1990). Our data represent publications that appeared in Google Scholar through March 2018. Our review of these secondary publications involved an initial read of the abstract, followed by a search for the original reference. These methods allowed us to ascertain if the secondary publication used the original instrument. If the paper did use the focal instrument, the paper was marked for later review. Once we identified papers that reported use of the focal instruments, all authors reviewed a subset in entirety, checking for consistency in identifying new populations and new uses. Each author then took one or more of the focal instruments and reviewed all secondary uses, further characterizing these citations and recording the use of the focal instrument. For each publication (secondary usage) we recorded the population, a description of the portion of instrument used (e.g. Andrews et al. (2011) reported using an abbreviated CINS comprised of 10 of the original 20 questions), additional evidence for reliability/validity (e.g. Rissler et al. (2014) reported Cronbach's alpha associated with administration of the MATE to undergraduates at the University of Alabama). To determine whether the study used the instrument on a new population we considered: (1) geographic area; (2) grade level; (3) field of study; and (4) academic level—introductory courses, advanced courses, or graduating seniors. We categorized the population based on the geographic region of the United States (midwestern, southwestern, southeastern, western, northwestern, northeastern) or the country. In the case of papers that were in languages other than English we relied on Google translator to evaluate if and how an instrument was used. In some cases, the description of the population in the new implementation was less specific than that of the original population in which case we did not consider it a new population because we could not tell whether the new implementation was potentially inclusive of the original population. For grade, field of study, and academic level we identified the following categories: undergraduates not enrolled in a specific course, undergraduates enrolled in a non-majors introductory biology course, undergraduates enrolled in a majors-level introductory biology course, undergraduates enrolled in an advanced biology course, undergraduates enrolled in a psychology course, undergraduate preservice teachers, high school teachers, high school students. When questions arose regarding how to characterize a particular use, we discussed it as a group that included at least three of the authors at any given point. For studies suggesting new implementations we were especially interested to know whether new uses of the instrument also included new measures of reliability/validity, as applicable. We evaluated these based on the criteria and examples outlined in Box 1. We recorded these data for each study we encountered.

Results

Initial review of focal instruments

Our initial review of the 13 focal instruments published between 1990 and 2016 found that two instruments included multiple versions (MATE, EALS). For the MATE we considered two of the versions unique enough to evaluate separately. The EALS Short-form was created directly from the Long-form and we therefore combined results for this instrument. Two of the assessments included only open ended, constructed response questions (ACORNS—assessing contextual reasoning about natural selection, ATEEK). Two included both constructed response and multiple-choice questions (ECT, MUM), and the remainder were some form of multiple choice, including Likert, agree/disagree, etc. (CINS, MATE, I-SEA, EALS, KEE, GAENE, GeDI, EvoDevoCI, CANS). We recorded information on instrument design, concepts covered, initial population, and evidence of validity and reliability. One (KEE) reported neither evidence of validity nor reliability, one reported some form of evidence of reliability only (ATEEK) and one reported evidence of validity only (ECT). Given the limitations of the KEE and ATEEK we do not discuss them further in this section, but results of our analysis can be found in Table 2. The remainder of the instruments had at least one type of evidence of both validity and reliability reported in the original publication. All assessments included undergraduates, either majors or non-majors, at some point during development. The early version of the MATE assessed high school biology teachers, but a later version was used with undergraduates. The I-SEA and GAENE included high school students in addition to undergraduates during development.

Assessments measuring natural selection

The ECT developed by Bishop and Anderson (1990) clearly served as the foundation for a number of subsequent studies, and the ORI in particular noted questions coming directly from the ECT. The original instrument developed by Bishop and Anderson consisted of six questions and claimed to measure understanding of natural selection among non-major undergraduates at a large midwestern university. The authors indicated that interrater reliability (IRR) was evaluated, stating that reliability was checked "by comparing the codes assigned to randomly selected student responses by two different coders" and that if disagreements occurred "coding was modified to produce better agreement". When disagreement between coders occurred, the coding procedure was modified to produce better agreement. However, no statistic for IRR was provided. The authors also report a number of sources of evidence of validity—review of textbook material as content, and student interviews as substantive.

The ACORNS instrument, developed following the ORI (open response instrument) which was based on the ECT, evaluates student "ability to use natural selection to explain evolutionary change" across a range of conditions (trait gain, trait loss, etc.). The instrument does focus on assessing elements of natural selection and non-scientific explanations (misconceptions) but also provides the option of scoring student responses for non-adaptive explanations for change as well (e.g. random changes in response to sampling error and drift). Nehm et al. (2012) report evidence of internal consistency by measuring Cronbach's alpha for key concepts and misconceptions (0.77 and 0.67 respectively) and report that IRR was greater than 80%. Content validity was assumed because the questions represent a number of possible biological scenarios. Evidence of internal consistency was provided by student interviews, and external structure was evaluated by comparing student responses on ACORNS questions to scores on the CINS. Using the ACORNS does require training in how to score student responses, alternatively, instructors can use EvoGrader (Moharreri et al. 2014) a machine-learning program that has been trained to score ACORNS questions.

Table 2 Summary of review of citations reporting new implementations of each instrument

Instrument	Abbr.	Original population	Reliability (R) of original instrument on original population	Validity (V) of original instrument on original population	# Publications that used instrument (#DiffVers.Only, #DiffPopOnly, #BothDiff)	New evidence of R for new version and/or new pop	New evidence of V for new version and/or new pop
Evolution concept test (natural selection)	ECT	UG-NM, MW	None	Content Substantive	27 (3, 2, 20)	Internal consistency (1, 3%); Inter-rater reliability (1, 3%)	Content (1, 3%) Substantive (1, 3%)
Concept inventory of natural selection	CINS	UG-NM, SW	Internal consistency (3)	Content Substantive Internal structure	31(1, 19, 10)	Internal consistency (2, 6%)	Substantive (2, 6%)
Measure of the acceptance of the theory of evolution	MATE	HS-T, MW (1999) UG-NM, S (2007)	Internal consistency (HS-T, MW; 2) Internal consistency (UG-NM, S; 4) Stability (UG-NM, S)	Content (HS-T, MW) Internal structure (HS-T, MW) Internal structure (UG-NM, S; 2) External structure (UG-NM, S)	88 (0, 41, 38)	Internal consistency (48, 54%)	Content (9, 10%) Internal structure (8, 9%) External structure (1, 1%)
Measure of understanding of macroevolution	MUM	UG-M, UG-MA, SE	Internal consistency (2)	Content	6 (0, 2, 1)	Internal consistency (1, 16%)	Content (1, 16%) Internal structure (1, 16%)
Knowledge of evolution exam	KEE	UG, MW	None	None	7 (0, 2, 1)	None	None
Evolution attitudes and literacy survey	EALS	UG-P, MW (LF) UG-M, MW (SF)	Internal consistency	Content Internal structure	8 (3, 0, 4)	Internal consistency (1, 12%)	Internal structure (1, 12%)
Accessing contextual reasoning about natural selection	ACORNS	UG, MW	Internal consistency Inter-rater agreement	Content Substantive External structure	9 (0, 0, 0)	N/A	N/A
Inventory of student evolution acceptance	I-SEA	HS-S, UG, W	Internal consistency	Content Internal structure	3 (0, 1, 0)	None	None
Evo-Devo concept inventory	EvoDevo CI	UG-M, UG-MA, MW, NE, S	Internal consistency Stability Alternate forms	Content Substantive External structure	1(0, 0, 0)	N/A	N/A
Assessment tool for evaluating evolution knowledge	ATEEK	UG-M, UG-MA, MW	Inter-rater agreement	None	2 (0, 0, 0)	None	None
Genetic drift inventory	GEDI	UG-MA, NW, SE, MW	Stability Internal consistency	Content Substantive Generalization	1 (0, 0, 0)	N/A	N/A
Generalized acceptance of evolution evaluation	GAENE	HS-S, UG, USA	Internal consistency	Content Internal structure	1 (0, 0, 0)	None	N/A
Concept assessment of natural selection	CANS	UG-M, W	Internal consistency Stability	Content Substantive	0 (0, 0, 0)	N/A	N/A

The CINS was originally developed as a 20-question instrument with evidence of validity and reliability provided for undergraduate non-majors in the southwestern region of the United States. The authors used Kuder-Richardson 20 to examine reliability, obtaining measurements of 0.58 and 0.64 on initial sections of the instrument. A good classroom instrument should have a reliability coefficient of 0.60 or higher. Expert reviewers provided evidence of content validity, interviews were used to evaluate if student responses on the multiple-choice questions reflected their thinking and principle component analysis (PCA) was used to examine internal structure. The authors also claimed that the instrument was generalizable because the original population used during development came from "large, ethnically diverse, community colleges". However, specific information about the demographics of the population was not provided and this claim has not been directly tested.

The CANS is composed of 24 multiple choice questions designed to measure five concepts related to natural selection: variation, selection, inheritance, mutation, and how these elements work together to cause evolution. Initial development was iterative, relying on student interviews and expert review to asses evidence of substantive and content validity, respectively. Kalinowski et al. (2016) also applied Item Response Theory to assess how well sets of questions assessed the same concept and if student responses fit a priori expectations. The authors also compared scores before and after instruction to evaluate reliability, reporting Cronbach's alpha before and after instruction (0.87 and 0.86, respectively), providing good evidence of reliability. The authors estimated that 88% of the variance in test scores in the experimental classroom was due to differences in student understanding of natural selection.

Assessments measuring additional evolutionary concepts

We found a single instrument purported to measure student understanding of macroevolution. The MUM was developed to measure student understanding of five essential concepts related to macroevolution: deep time, phylogenetics, fossils, speciation, and nature of science. Development of the instrument relied on responses generated by undergraduates taking courses in either introductory biology or upper-level evolution at a large southeastern university. Textbook analysis and expert reviews were used as evidence of content validity. The authors used Cronbach's alpha as a measure of internal consistency and report a value for the entire sample that is considered acceptable (0.86). However, Cronbach's alpha varied across their samples, ranging from values considered questionable to values considered acceptable, possibly suggesting the instrument provides better evidence for some populations than others. No additional evidence was provided.

The GeDI was developed to measure upper-level biology majors understanding of genetic drift as a process of evolutionary change. The authors used an iterative development process that included open-ended questions, student interviews, multiple expert reviews, and item analysis. The final instrument was also evaluated for evidence of reliability. A coefficient of stability of 0.82 was reported in a test–retest administration. Cronbach's alpha varied across populations (0.58–0.88), and the authors note that the concepts covered in the instrument align best with upper-level evolution courses.

The EvoDevo CI is a concept inventory developed specifically to measure student understanding of six core concepts related to evolutionary changes caused by development. The authors relied on iterative development that included expert review, student interviews, testing and item revision. They reported Cronbach's alpha, calculated for different groups, as a measure of whether the instrument assessed the intended construct among biology majors. In addition, tests for evidence of reliability reported good stability as measured by Pearson correlation of 0.960, $P < 0.01$.

Assessments reporting to measure acceptance of evolution

The MATE was designed to measure overall acceptance of evolutionary theory by assessing perceptions of concepts considered fundamental to evolution. Originally developed using a population of high school biology teachers (Rutledge and Warden 1999), it was then updated using undergraduate non-majors (Rutledge and Sadler 2007). Both versions include 20 items assessed using a five-point Likert scale. The original version published by Rutledge and Warden (1999) reported internal consistency using Cronbach's alpha (0.98) as evidence of reliability, expert review by a panel of five experts as evidence of content validity, and a principle factor analysis as evidence of internal structure validity. The second version of the MATE examined reliability of the instrument for a population of non-major undergraduate students and reported Cronbach's alpha reliability coefficient of 0.94 as evidence of internal consistency. No additional evidence was reported.

The EALS Long-Form was developed to assess predominant regional belief systems and their roles in science understanding and attitudes, particularly as pertain to evolution, drawing from previous literature and published instruments to generate Likert scale items. The EALS Short-Form was then tested on undergraduates in an introductory biology course. Both forms included items for the 16 lower order constructs and then used confirmatory analysis to determine the six higher order constructs. We suspect the EALS Short-Form is more likely to be used, and therefore provide a summary here. Additional information on the long form can be found in Table 2. The authors reported a range of alpha coefficients for the 16 lower-order constructs as evidence of internal consistency and suggested loadings from a confirmatory factor analysis provided evidence of internal structure validity.

The I-SEA was also designed to measure student acceptance of evolution, based on three subscales: microevolution, macroevolution, and human evolution. Development included using open-ended questions and student interviews. An initial 49 item Likert scale instrument was developed and tested, and then modified to the current 24 item instrument. The overall Cronbach's alpha was 0.95, providing evidence of internal consistency. Experienced biology teachers, science teacher educators, and college biology faculty served as expert reviewers, providing evidence of content validity. Evidence of internal structure was measured using an exploratory factor analysis, however, there were some issues here because only loadings for the first four items for each subscale were reported, making it difficult to fully evaluate these measures. The populations used in development included high school students and undergraduates, predominantly at institutions in the western United States.

The most recently published instrument developed that measures acceptance of evolution is the GAENE, specifically designed to measure only acceptance of evolution, defined as "the mental act or policy of deeming, positing, or postulating that the current theory of evolution is the best current available scientific explanation of the origin of new species from preexisting species". The GAENE was also developed based on other instruments, relying on extensive interviews and testing, followed by multiple rounds of revision, and expert feedback. Smith et al. (2016) reported Cronbach's alpha of 0.956 for later versions, providing excellent evidence of internal consistency. Evidence of validity was provided by Rasch analysis, demonstrating discrimination between respondents with low and high levels of acceptance, and PCA that supported a unidimensional structure accounting for 60% of the variance. A range of populations were used in developing the instrument, including high school students and undergraduates at a range of institutions.

Secondary uses of focal instruments

Using the "cited by" link provided in Google scholar for each of the publications associated with the 13 focal instruments, we examined over 2000 peer-reviewed citations that made reference to one or more of the 13 focal instruments. Many of the citations simply referenced the publication but did not use any portion of the instrument. We did identify 182 studies that used at least one of the 13 instruments we reviewed. Figure 1 shows the relative frequency of re-use of each of the instruments ranging from 0 (CANS) to 88 (MATE). We defined a new use of the instrument as either using a different version (altered measurement scale or item set and item rewording or language translation) and/or administering the instrument to a new population. Our review found that most new uses of the instruments did involve either administration to a new population and/or the use of a revised version, particularly if the instrument was published more than 5 years ago (Fig. 2, Table 2). Figure 2a shows the proportion of studies that indicated a new use of the instrument for six of the 13 instruments. Figure 2b shows the proportion of these new uses that reported

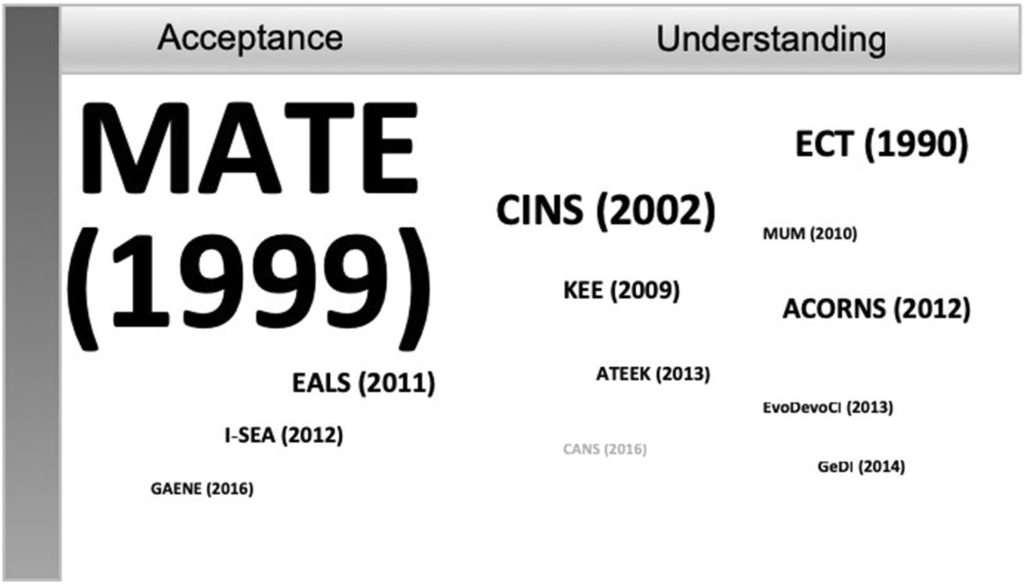

Fig. 1 Proportional re-use of all instruments. For example, the MATE was used in 70 subsequent studies, the I-SEA in only three. Gray text indicates the assessment has yet to be used in a new study. Instruments are organized according to construct (content and psychology dimension)

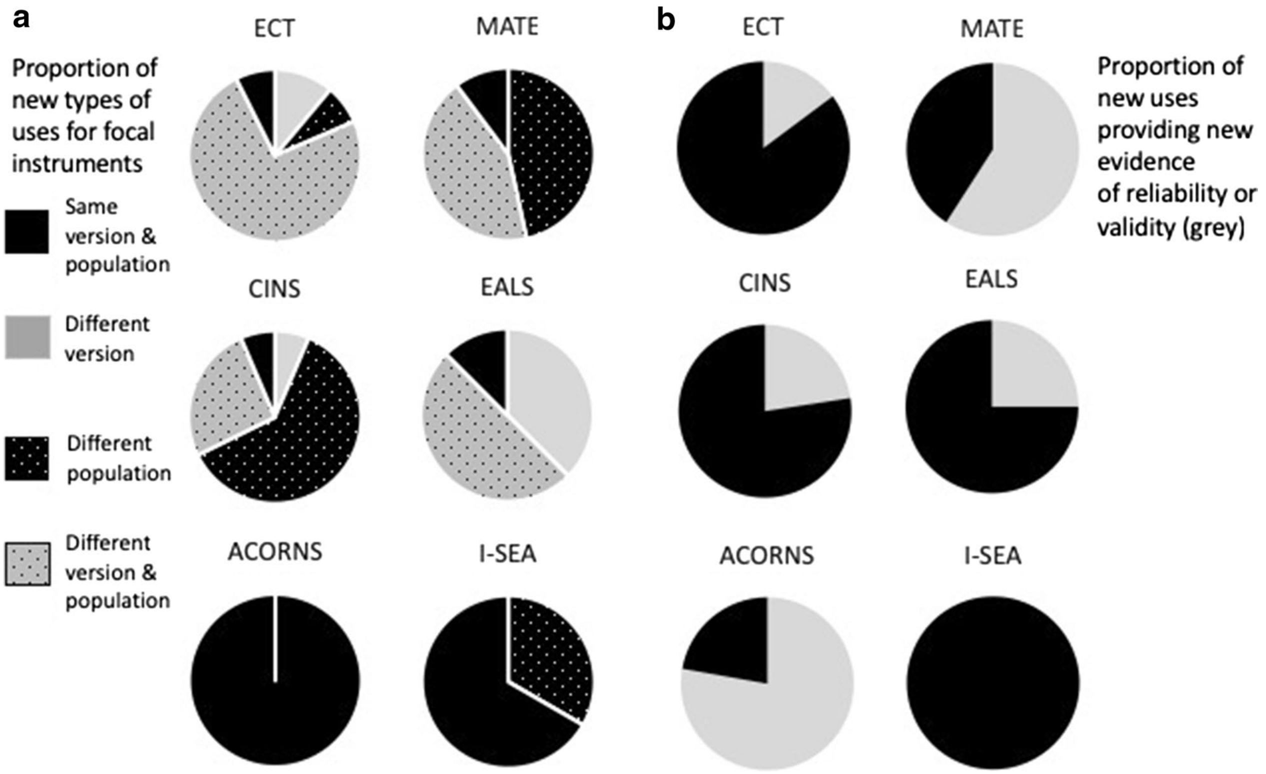

Fig. 2 **a** The proportion use of instruments categorized by type of use, e.g. proportion of secondary uses for the ECT that altered the original version. **b** The proportional of secondary uses that reported additional or new evidence of reliability or validity, whether for a new population or new implementation of the instrument

new evidence of reliability or validity. Figure 2 shows only a subset of the instruments as a number of instruments were so recently published that there have been few secondary uses. Table 2 summarizes all data, indicating the specific types of reliability and validity evidence provided. Additional file 1: Table S1 is a searchable database with additional details for each of the secondary uses of the instruments.

The ECT, first published by Bishop and Anderson (1990), was initially used with undergraduate non-majors. Our analysis suggests the instrument (or some approximation of the instrument) has been used in 27 subsequent studies. Two studies (Nehm and Reilly 2007; Andrews et al. 2011) altered the ECT, three studies administered the complete instrument to a new population (Settlage 1994; Demastes et al. 1995), and 20 of the re-administrations of the ECT involved a new population and used only a subset of the original questions presented in Bishop and Anderson (1990). Included in this category were studies that report using the ORI (open response instrument) because Nehm and Reilly (2007) report modifying questions from Bishop and Anderson (1990) in creating the ORI. We also found reference to the ACORNS questions as being derived from the ECT as well; however, we evaluated the ACORNS separately. In many cases, reuse of the ECT did not include any new evidence of reliability and validity (Fig. 2b). The exceptions involved uses of the ORI, new implementations often included new measures (Ha et al. 2012, Nehm and Schonfeld 2007). For example, Nehm and Schonfeld (2007) provided additional evidence of both reliability (i.e., internal consistency and IRR) and validity (e.g. content and substantive) for students in a graduate teacher education program.

We identified 31 publications that referenced using the Concept Inventory for Natural Selection (CINS), one used some version of the instrument (Pope et al. 2017), most likely administering a portion of the full instrument, 19 administered the instrument to a new population, and ten studies reported using the instrument with a new population and changing the question structure. A few of these studies reported additional evidence of reliability and validity. Athanasiou and Mavrikaki (2013) reported evidence of reliability (Cronbach's alpha) and validity (construct validity using PCA) for biology and non-biology majors in Greece. Nehm and Schonfeld (2008) report additional evidence of convergent validity (between the CINS and ORI) and discriminant validity for undergraduate biology majors in northeast. Ha et al. (2012) also looked at the correlation between scores on the ORI and the CINS, and report Cronbach's alpha for undergraduates in preservice biology. Weisberg et al. (2018) administered the CINS to a sample from the general public and reported Cronbach's alpha. Finally, Pope et al. (2017) also report Cronbach's alpha and interrater reliability for biology majors in the northeast.

The ACORNS instrument has been used in nine subsequent studies. The ability to vary the open-ended questions allows researchers to create new versions without altering the general framework of the instrument, therefore none of the subsequent uses were considered new versions. The original population reported in Nehm et al. (2012) stated the population used to assess reliability and validity were undergraduates at a midwestern university. The instrument was then used in subsequent studies, most commonly listing the population as undergraduate biology majors. It was therefore not possible to determine if the re-uses of the instrument qualified as new populations. However, all of these studies did report IRR as evidence of reliability.

The MUM has been used infrequently, perhaps because of issues identified by Novick and Catley (2012) or because instructors are often more interested in students understanding of natural selection. However, Romine and Walter (2014) administered the MUM to undergraduates enrolled in non-majors' biology and found internal construct validity to be strongly supported using Rasch analysis but did find a single construct as opposed to five in the original study. Of the studies that do report using the instrument, two report using slightly modified versions and one modified the version and administered it to a new population.

At the time of our analysis, the concept assessment of natural selection (CANS), the knowledge of evolution exam (KEE), the Assessment Tool for Evaluating Evolutionary Knowledge (ATEEK), the genetic drift inventory (GeDI), and the EvoDevo Concept Inventory (EvoDevo CI) had not been used very often and currently no additional evidence of reliability or validity has been provided for these instruments.

For the MATE, of the total 88 new uses of the instrument, 48 of the implementations provided new evidence of reliability while 18 provided new evidence of validity, although with wildly different rigor (Fig. 2b). Having been one of the original and seemingly most versatile instruments, the MATE has been implemented in quite diverse contexts and forms, including being used in fourteen countries, and translated to five other languages, often with multiple independent translations. The primary non-USA and non-English use of the MATE is in Turkey and Turkish, and with likely six independent translations. Many populations unique from the original in terms of educational background have been assessed, including early childhood or primary school teachers, university faculty, and museum visitors. The number of items administered have fluctuated between 4 and 27 through item reduction, splitting, and/or combination with other items (not including other identified instruments). Finally, the measurement scale has varied between four-, six-, and seven-point Likert scales. Notable implementations that introduce validity and reliability evidence are largely limited to Turkish populations (Akyol et al. 2010, 2012a, b; Irez and Özyeral Bakanay 2011; Tekkaya et al. 2012; Yüce and Önel 2015) with two notable studies (Manwaring et al. 2015 and Romine et al. 2017) providing the strongest evidence of internal structure validity with populations similar to the original American undergraduate implementations. The dearth of evidence regarding validity for the MATE pales in comparison to its diversity of implementations—an undesirable state indeed for measurements standards.

We found eight additional uses of the Evolution Attitudes and Literacy Survey (EALS), either the short or long form. Three studies reported using the EALS in the original format and administered it to similar populations as those used in the initial studies. One altered the format and another four changed both the version and the population. Of these only one reported new evidence of reliability or validity (Mead et al. 2015).

The Inventory of Student Evolution Acceptance (I-SEA) and the Generalized Acceptance of Evolution Evaluation (GAENE) have also not been used very often. In the case of the I-SEA only one publication reported using the instrument and it was not possible to determine if it was a new population or new version. However, no additional evidence of reliability or validity were provided. We suspect the GAENE has not been used because it was so recently published. However, the strong evidence offered in the initial description of the instrument suggest it may be used more often in the future.

Discussion

The ability of any instrument to measure student understanding is dependent on a number of factors—for example, the development process, initial population assessed, evidence of validity and reliability, evaluation of what we think it measures, and consistency in measurement (Campbell and Nehm 2013). We found new uses of the original instruments overall provided sparse new evidence of validity or reliability and encountered various issues while evaluating the instruments and their subsequent reuse. These included the narrow character of the original population (e.g. MATE) and the failure of adhering to measurement standards by entirely lacking validity and reliability evidence (e.g. KEE). In reviewing subsequent uses it was often difficult to ascertain what portion and/or version of the original instrument was used, for example, some studies simply referenced using questions from Bishop and Anderson (1990) but did not indicate which questions were used (Gregory and Ellis 2009). Further, the authors of the MATE have published four distinct versions (Rutledge and Sadler 2007, 2011; Rutledge and Warden 1999, 2000) that differ with respect to item wording and/or ordering, and this fact has remained unremarked upon in the literature.

Use of the MATE is further complicated by the fact that, although there is evidence of validity, it is not clear what is meant by "acceptance" (Smith 2010a). More recently, the internal structure of the MATE in terms of the number and identity of measurable constructs (i.e., named sets of items measuring the same concept) has been found to be unclear. Wagler and Wagler challenged the content and internal structure validity for the MATE, and studies report the MATE represents one (Rutledge and Warden 1999; Rissler et al. 2014; Deniz et al. 2008), two (Romine et al. 2017), four (Manwaring et al. 2015), six (untested: Rutledge and Sadler 2007), or an unidentifiable number of constructs (e.g. Wagler and Wagler 2013; Hermann 2012, 2016; Rowe et al. 2015). However, more recently, Romine et al. (2017) has suggested the MATE is psychometrically sound.

We also encountered published debates regarding validity, including content and substantive validity, for the MUM (i.e., Novick and Catley 2012; Nehm and Kampourakis 2014). Novick and Catley (2012) found significant issues with respect to validity evidence for the MUM, suggesting it does not adequately measure student understanding of macroevolution. However, Romine and Walter (2014) challenged the findings of Novick and Catley (2012) suggesting that their analysis provided evidence that the MUM is a psychometrically sound instrument. These debates emphasize again the importance of testing any instrument for evidence of reliability and validity when using it in a new implementation.

Instruments developed more recently (GeDI, EvoDevCI, CANS, GAENE) have not yet been used widely. However, we note that these studies included relatively broad initial populations in their development and provided multiple lines of evidence for both reliability and validity, suggesting these may be useful across a wide range of future implementations.

Conclusions and recommendations

The focus on evaluating teaching and learning in undergraduate biology has led to the creation of a number of different instruments that can be used to assess student understanding and acceptance of evolution. However, it is clear that examining each instrument for evidence of reliability and validity for a particular intended use is important for being able to make accurate and valid inferences. Our analysis of published instruments provides useful information to consider. We strongly recommend that research on student understanding and acceptance of evolution include continued evaluation. For example, owing to its popularity in the literature, we have specific recommendations for readers if they intend to administer the MATE. The authors' most recent version (Rutledge and Sadler 2011) is the soundest grammatically and, although further study on this is warranted. Therefore, this English version is most highly recommended, if modifications are desired due to cultural incongruence, ESL (English Second Language) interpretation, non-English usability, neutrality avoidance, etc. Doing so would maintain adherence to measurement standards and aid comparison within the literature by reducing the increasing diversity of versions lacking any—let alone adequate—evidence of validity and reliability. However, unease regarding the content and internal structure validity for the MATE (see above) was a driving factor in the creation of alternative instruments to measure acceptance (i.e., EALS, I-SEA, GAENE). The GAENE in particular went through multiple iterations, included a broad population in its testing, and meets criteria for measuring "acceptance of evolution" (Smith et al. 2016), in addition to evidence of reliability and validity.

In addition to concerns about evidence of validity and reliability, many studies reported using only portions of a particular instrument. In some cases, however, it may be important to use the instrument as developed—administering all of the items and using their original wording and measurement scale—if one wishes to draw comparisons or rely on previous evidence of validity and reliability for similar populations. While some forms of validity (for example substantive or content) may not be affected, instruments are developed to measure a particular construct, or set of related constructs, and changing the structure of the assessment may influence how well it measures the constructs of interest.

We strongly support extending measurement criteria to all the instruments reviewed here and recommend against using instruments for which the original publication did not report evidence of reliability and validity, or for which this evidence is weak. Researchers should review the literature, paying particular attention to alignment between learning goals and choice of instrument. Furthermore, as instruments are modified and/or used on new populations, measurement standards should be adhered to, and reported in the literature. Such reports will further extend the uses of these instruments and strengthen the ability of researchers to draw meaningful conclusions from studies.

In addition, we want to recognize that many of the instruments developed more recently (e.g. CANS, GeDI, EvoDevoCI, GAENE) include multiple lines of evidence referencing strong reliability and validity, and these should be used as models for continued development of new instruments. Developers of scientific instruments need to clearly lay out under what conditions their assessment is to be used and to encourage those using the assessment outside of those parameters to gather more evidence. Ziadie and Andrews (2018) point out that any assessment should include the dimensions of the topic that are important to assess and include consistent methodology and interpretation of results.

Our review highlights the importance of applying measurement standards to instruments, hopefully helping researchers to assess student understanding and acceptance of evolution. We have provided a supplemental database that allows researchers to easily examine a particular instrument, and any subsequent uses that may help determine if it is an appropriate instrument for a given population. We cannot emphasize enough, however, that it is imperative that any new implementation of these instruments be tested according to accepted measurement criteria and that researchers publish any new evidence of reliability and validity.

Additional file

Additional file 1. Searchable database of an overview of each instrument reviewed and characterization of any published studies that report using the instrument, specifying additional evidence of reliability and validity for new implementation.

Abbreviations

ACORNS: assessing contextual reasoning about natural selection; ATEEK: assessment tool for evaluating evolution knowledge; CANS: concept assessment of natural selection; ECT: evolution concept test; CINS: concept inventory of natural selection; EALS: Evolutionary Attitudes and Literacy Survey; ESL: english second language; EvoDevoCI: evolutionary developmental concept inventory; GAENE: generalized acceptance of evolution evaluation; GeDI: genetic drift inventory; IRR: inter-rater reliability; I-SEA: inventory of student acceptance of evolution; KEE: knowledge of evolution exam; MATE: measure of acceptance of the theory of evolution; MUM: measure of understanding of macroevolution; ORI: open response instrument; PCA: principle component analysis.

Authors' contributions

LSM, CK, AW and KS all contributed equally to reviewing the literature. LSM took the lead on writing. CK and AW contributed specific sections associated with instruments they reviewed. KS presented original results at Evolution 2016. All authors read and approved the final manuscript.

Author details

[1] BEACON Center for the Study of Evolution in Action, Michigan State University, 567 Wilson Rd, East Lansing, MI 48824, USA. [2] Department of Integrative Biology, Michigan State University, 288 Farm Lane, East Lansing, MI 48824, USA. [3] The W. M. Keck Science Department of Claremont McKenna, Scripps, and Pitzer Colleges, 925 N. Mills Ave, Claremont, CA 91711, USA. [4] Lyman Briggs College, Michigan State University, 919 East Shaw Lane, East Lansing, MI 48825, USA.

Acknowledgements

The BEACON Evo Ed Curriculum group provided helpful feedback on initial discussions of the manuscript. Initial ideas for project were also discussed with Dr. Ross Nehm whom we thank for emphasizing the importance of measurement standards.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

We have created a searchable Google spreadsheet for the Additional file 1: Table S1. Please see file at the following link: https://drive.google.com/open?id=16998xXA620Uf-Aan5MZH87TPztAxhNuM.

Funding

Howard Hughes Medical Institute; National Science Foundation, Grant Numbers: OIA-0939454; DBI-144683.

Publisher's Note

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

Received: 19 November 2018 Accepted: 8 January 2019
Published online: 08 February 2019

References

Akyol G, Tekkaya C, Sungur S. The contribution of understandings of evolutionary theory and nature of science to pre-service science teachers' acceptance of evolutionary theory. Procedia Soc Behav Sci. 2010;9:1889–93. https://doi.org/10.1016/j.sbspro.2010.12.419.

Akyol G, Tekkaya C, Sungur S. Examination of pre-service science teachers' perceptions and understanding of evolution in relation to socio-demographic variables. Procedia Soc Behav Sci. 2012a;31:167–72. https://doi.org/10.1016/j.sbspro.2011.12.036.

Akyol G, Tekkaya C, Sungur S, Traynor A. Modeling the interrelationships among pre-service science teachers' understanding and acceptance of evolution, their views on nature of science and self-efficacy beliefs regarding teaching evolution. J Sci Teacher Educ. 2012b;23(8):937–57. https://doi.org/10.1007/s10972-012-9296-x.

Allmon WD. Why don't people think evolution is true? Implications for teaching, in and out of the classroom. Evol Educ Outreach. 2011;4(4):648–65. https://doi.org/10.1007/s12052-011-0371-0.

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

Anderson DL, Fisher KM, Norman GJ. Development and evaluation of the conceptual inventory of natural selection. J Res Sci Teach. 2002;39(10):952–78.

Andrews TM, Kalinowski ST, Leonard MJ. "Are humans evolving?" A classroom discussion to change student misconceptions regarding natural selection. Evo Educ Outreach. 2011;4:456–66.

Athanasiou K, Mavrikaki E. Conceptual inventory of natural selection as a tool for measuring Greek University Students' evolution knowledge: differences between novice and advanced students. Int J Sci Educ. 2013;36(8):1262–85.

Barnes ME, Evans EM, Hazel A, Brownell SE, Nesse RM. Teleological reasoning, not acceptance of evolution, impacts students' ability to learn natural selection. Evol Educ Outreach. 2017;10(1):7. https://doi.org/10.1186/s12052-017-0070-6.

Baum D, Dewitt Smith S, Donovan SS. The tree thinking challenge. Science. 2005;310:979–80.

Bishop BA, Anderson CW. Student conceptions of natural selection and its role in evolution. J Res Sci Teach. 1990;27(5):415–27.

Branch G, Mead LS. "Theory" in theory and practice. Evol Educ Outreach. 2008;1(3):287–9. https://doi.org/10.1007/s12052-008-0056-5.

Brownell SE, Freeman S, Wenderoth MP, Crowe AJ. BioCore guide: a tool for interpreting the core concepts of vision and change for biology majors. CBE Life Sci Educ. 2014;13(2):200–11. https://doi.org/10.1187/cbe.13-12-0233.

Campbell CE, Nehm RH. A critical analysis of assessment quality in genomics and bioinformatics education research. CBE Life Sci Educ. 2013;12(3):530–41. https://doi.org/10.1187/cbe.12-06-0073.

Demastes SS, Settlage J Jr, Good R. Students' conceptions of natural selection and its role in evolution: cases of replication and comparison. J Res Sci Teach. 1995;32(5):535–50.

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(4):420–43. https://doi.org/10.1002/tea.20223.

Funk C, Rainie L. Public and scientists' views on science and society. Washington: Pew Research Center; 2015.

Gregory TR, Ellis CA. Conceptions of evolution among science graduate students. Bioscience. 2009;59(9):792–9.

Ha M, Baldwin BC, Nehm RH. The long-term impacts of short-term professional development: science teachers and evolution. Evol Educ Outreach. 2015;8(1):11. https://doi.org/10.1186/s12052-015-0040-9.

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(1):95–121. https://doi.org/10.1002/tea.20449.

Hawley PH, Short SD, McCune LA, Osman MR, Little TD. What's the matter with Kansas?: the development and confirmation of the evolutionary attitudes and literacy survey (EALS). Evol Educ Outreach. 2011;4(1):117–32. https://doi.org/10.1007/s12052-010-0294-1.

Hermann RS. High school biology teachers' views on teaching evolution: implications for science teacher educators. J Sci Teacher Educ. 2012;24(4):597–616. https://doi.org/10.1007/s10972-012-9328-6.

Hermann RS. Elementary education majors' views on evolution: a comparison of undergraduate majors understanding of natural selection and acceptance of evolution. Electr J Sci Educ. 2016; 20(6). http://ejse.southwestern.edu/article/view/16305.

Irez S, Özyeral Bakanay ÇD. An assessment into pre-service biology teachers' approaches to the theory of evolution and nature of science. Egitim ve Bilim. 2011;36(162):39.

Kalinowski ST, Leonard MJ, Taper ML. Development and validation of the conceptual assessment of natural selection (CANS). CBE Life Sci Educ. 2016;15(4):ar64. https://doi.org/10.1187/cbe.15-06-0134.

Manwaring KF, Jensen JL, Gill RA, Bybee SM. Influencing highly religious undergraduate perceptions of evolution: Mormons as a case study. Evol Educ Outreach. 2015;8:23. https://doi.org/10.1186/s12052-015-0051-6.

Mead LS, Clarke JB, Forcino F, Graves JL. Factors influencing minority student decisions to consider a career in evolutionary biology. Evol Educ Outreach. 2015;8(1):6. https://doi.org/10.1186/s12052-015-0034-7.

Mead LS, Scott EC. Problem concepts in evolution part I: purpose and design. Evol Educ Outreach. 2010a;3(1):78–81. https://doi.org/10.1007/s12052-010-0210-8.

Mead LS, Scott EC. Problem concepts in evolution part II: cause and chance. Evol Educ Outreach. 2010b;3(2):261–4. https://doi.org/10.1007/s12052-010-0231-3.

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

Moharreri K, Minsu H, Nehm RH. EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. Evol Educ Outreach. 2014;7:15.

Moore R, Cotner S. Rejecting Darwin: the occurrence & impact of creationism in high school biology classrooms. Am Biol Teacher. 2009;71(2):e1–4. https://doi.org/10.1662/005.071.0204.

Nadelson LS, Southerland SA. Development and preliminary evaluation of the measure of understanding of macroevolution: introducing the MUM. J Exp Educ. 2009;78(2):151–90. https://doi.org/10.1080/00220970903292983.

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

National Research Council. A framework for K-12 science education: practices, crosscutting concepts, and core ideas. Washington, DC: The National Academies Press; 2012.

Nehm RH, Beggrow EP, Opfer JE, Ha M. Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. Am Biol Teacher. 2012;74(2):92–8. https://doi.org/10.1525/abt.2012.74.2.6.

Nehm RH, Ha M. Item feature effects in evolution assessment. J Res Sci Teach. 2011;48(3):237–56. https://doi.org/10.1002/tea.20400.

Nehm RH, Kampourakis K. History and philosophy of science and the teaching of macroevolution. In: Matthews MR, editor. International handbook of research in history, philosophy and science teaching. Dordrecht: Springer; 2014. p. 401–21.

Nehm RH, Reilly L. Biology majors' knowledge and misconceptions of natural selection. Bioscience. 2007;57(3):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(5):699–723. https://doi.org/10.1007/s10972-007-9062-7.

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

Novick LR, Catley KM. Assessing students' understanding of macroevolution: concerns regarding the validity of the MUM. Int J Sci Educ. 2012;34(17):2679–703. https://doi.org/10.1080/09500693.2012.727496.

Perez KE, Hiatt A, Davis GK, Trujillo C, French DP, Terry M, Price RM. The EvoDevoCI: a concept inventory for gauging students' understanding of evolutionary developmental biology. CBE Life Sci Educ. 2013;12(4):665–75.

Petto AJ, Mead LS. Misconceptions about the evolution of complexity. Evol Educ Outreach. 2008;1(4):505–8. https://doi.org/10.1007/s12052-008-0082-3.

Pope DS, Rounds CM, Clarke-Midura J. Testing the effectiveness of two natural selection simulations in the context of a large-enrollment undergraduate laboratory class. Evol Educ Outreach. 2017;10(1):3. https://doi.org/10.1186/s12052-017-0067-1.

Price RM, Andrews TC, McElhinny TL, Mead LS, Abraham JK, Thanukos A, Perez KE. The genetic drift inventory: a tool for measuring what advanced undergraduates have mastered about genetic drift. Cell Biol Educ. 2014;13(1):65–75. https://doi.org/10.1187/cbe.13-08-0159.

Rissler LJ, Duncan SI, Caruso NM. The relative importance of religion and education on university students' views of evolution in the Deep South and state science standards across the United States. Evol Educ Outreach. 2014;7(1):1–17. https://doi.org/10.1186/s12052-014-0024-1.

Romine William L, 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. https://doi.org/10.1002/tea.21380.

Romine William Lee, Walter EM. Assessing the efficacy of the measure of understanding of macroevolution as a valid tool for undergraduate non-science majors. Int J Sci Educ. 2014;36(17):2872–91. https://doi.org/10.1080/09500693.2014.938376.

Rowe MP, Gillespie BM, Harris KR, Koether SD, Shannon L-JY, Rose LA. Redesigning a general education science course to promote critical thinking. CBE Life Sci Educ. 2015;14(3):ar30. https://doi.org/10.1187/cbe.15-02-0032.

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. https://doi.org/10.1662/0002-7685(2007)69%5b332:ROTMOA%5d2.0.CO;2.

Rutledge ML, Sadler KC. University students' acceptance of biological theories—is evolution really different? J Coll Sci Teach. 2011;41(2):38–43.

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

Rutledge ML, Warden MA. Evolutionary theory, the nature of science & high school biology teachers: critical relationships. Am Biol Teacher. 2000;62(1):23–31. https://doi.org/10.1662/0002-7685(2000)062%5b0023:ETTNOS%5d2.0.CO;2.

Settlage J Jr. Conceptions of natural selection: a snapshot of the sense-making process. J Res Sci Teach. 1994;31(5):449–57.

Short SD, Hawley PH. Evolutionary Attitudes and Literacy Survey (EALS): development and validation of a short form. Evol Educ Outreach. 2012;5(3):419–28.

Shtulman A. Qualitative differences between naïve and scientific theories of evolution. Cogn Psychol. 2006;52:170–94.

Sinatra GM, Southerland SA, McConaughy F, Demastes JW. Intentions and beliefs in students' understanding and acceptance of biological evolution. J Res Sci Teach. 2003;40(5):510–28. https://doi.org/10.1002/tea.10087.

Smith JJ, Cheruvelil KS, Auvenshine S. Assessment of student learning associated with tree thinking in an undergraduate introductory organismal biology course. CBE Life Sci Educ. 2013;12(3):542–52. https://doi.org/10.1187/cbe.11-08-0066.

Smith MU. Current status of research in teaching and learning evolution: I. Philosophical/epistemological issues. Sci Educ. 2010a;19(6–8):523–38. https://doi.org/10.1007/s11191-009-9215-5.

Smith MU. Current status of research in teaching and learning evolution: II. Pedagogical issues. Sci Educ. 2010b;19(6–8):539–71. https://doi.org/10.1007/s11191-009-9216-4.

Smith MU, Siegel H. Knowing, believing, and understanding: what goals for science education? Sci Educ. 2004;13(6):553–82. https://doi.org/10.1023/B:SCED.0000042848.14208.bf.

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.

Summers MM, Couch BA, Knight JK, Brownell SE, Crowe AJ, Semsar K, Wright CD, Smith MK. EcoEvo-MAPS: an ecology and evolution assessment for introductory through advanced undergraduates. CBE Life Sci Educ. 2018;17(2):18. https://doi.org/10.1187/cbe.17-02-0037.

Swift A. In U.S., belief in creationist view of humans at New Low. Gallop. 2017.

Tekkaya C, Akyol G, Sungur S. Relationships among teachers' knowledge and beliefs regarding the teaching of evolution: a case for Turkey. Evol Educ Outreach. 2012;5(3):477–93. https://doi.org/10.1007/s12052-012-0433-y.

Wagler A, Wagler R. Addressing the lack of measurement invariance for the measure of acceptance of the theory of evolution. Int J Sci Educ. 2013;35(13):2278–98. https://doi.org/10.1080/09500693.2013.808779.

Weisberg DS, Landrum AR, Metz SE, Weisberg M. No missing link: knowledge predicts acceptance of evolution in the United States. Bioscience. 2018;68(3):212–22. https://doi.org/10.1093/biosci/bix161.

White PJT, Heidemann MK, Smith JJ. A new integrative approach to evolution education. Bioscience. 2013;63(7):586–94.

Yüce Z, Önel A. The comprehension of nature of science by science teacher candidates and the determination of their acceptation levels of evolutionary theory. Turk Stud Int Period Lang Lit Hist Turk Turkic. 2015;10:857–72.

Ziadie MA, Andrews TC. Moving evolution education forward: a systematic analysis of literature to identify gaps in collective knowledge for teaching. CBE Life Sci Educ. 2018;17(1):ar11. https://doi.org/10.1187/cbe.17-08-0190.