EDITORIAL

Evolution assessment: introduction to the special issue

Ross H. Nehm[1*] and Louise S. Mead[2]

Introduction

Since its founding a decade ago, *Evolution: Education and Outreach* has served as a unique forum for discussing and disseminating empirical and conceptual advances in our understanding of the many challenges facing evolution educators. The journal has also helped to raise awareness among scientists that many educational actions remain rooted in intuition and tradition rather than in evidence. Indeed, scientific research in education (National Research Council 2002) and "Scientific Teaching" (Handelsman et al. 2006) exemplify recent efforts by the community to harness the strengths of the scientific enterprise and apply them to the design, execution, and evaluation of instructional practices. Educational research is not, and never will be, epistemologically equivalent to scientific research. But this does not mean that some of the highly valued characteristics of scientific research cannot be successfully applied to education (National Research Council 2002). Crossing the disciplinary divide between education and science research is challenging, and fostering dialogue and interaction between scientists and educators focusing on evolution education remains one bridge that *Evolution: Education and Outreach* is committed to building.

As the scientific and educational communities continue to embrace more rigorous, evidence-based approaches to teaching and learning, they have become increasingly reliant on tools for gathering evidence suitable for building generalizable, evidence-based claims. Assessment tools or measurement instruments (tests, questionnaires) are commonly used to capture latent or unobservable attributes (e.g., knowledge, attitudes). Instruments are broadly defined as standardized tools for quantifying observations (National Research Council 2001; Liu 2010). Just as it is unlikely that biologists would value data generated by a new piece of laboratory equipment that lacks a certificate of precision, one that has never been used by another lab, or one that produces inconsistent results, evolution educators should be weary of educational instruments lacking analogous attributes.

Despite significant progress in recent years, research on the measurement and assessment of knowledge, skills, and dispositions central to evolution education remains comparatively limited relative to other foci (Nehm 2006; National Research Council 2012a). In order to draw greater attention to the topic of evolution assessment and further research efforts in this area, *Evolution: Education and Outreach* debuts its first special issue devoted to the topic of evolution assessment. The seven articles in this collection explore a range of topics: reviews of existing assessment instruments (e.g., concept inventories) and how they may be used by evolution educators; syntheses of the forms of evidence that have been gathered to support claims about the quality of evidence drawn from such tools; analyses of the psychometric properties and functioning of evolution education instruments in diverse samples; and studies of how different instructional contexts impact learning. Collectively, this body of work sheds light on many of the complex issues undergirding our efforts to generate evidence suitable for guiding practice. After reviewing some of the salient findings and perspectives outlined in this collection, we end with a brief discussion of possible next steps for future evolution assessment research.

Evolution assessment

The first article in the collection, by Furrow and Hsu (2019), introduces educators to 14 evolution education instruments and discusses the ways in which these tools

*Correspondence: ross.nehm@stonybrook.edu
[1] Department of Ecology and Evolution, Institute for STEM Education, Stony Brook University, Stony Brook, NY 11733, USA
Full list of author information is available at the end of the article

may be used to embark upon evidence-based evolution education. For example, Furrow and Hsu outline how these instruments may be used to: inform the development of course learning objectives; uncover student ideas and misconceptions prior to teaching; measure student learning in an activity, module, or course; or inspire the design of activities targeting challenging concepts. The article contains numerous examples from the authors' experiences that nicely illustrate how they have leveraged these tools to improve learning outcomes. The authors end by emphasizing that while there are many existing measurement instruments that should be more widely used, there are notable conceptual gaps (e.g., speciation, human evolution, sexual selection, quantitative genetics, evolutionary medicine, and biodiversity). Development of these tools will require collaborative efforts among evolutionary biologists and educators.

There are several different conceptual frameworks that may be used for evaluating the quality of measurement instruments (e.g. *Standards of Educational and Psychological Measurement* (AERA et al. 2014; Pellegrino et al. 2016). Mead et al. (2019) employ Messick's (1995) seminal framework to document the forms of evidence that have been used to support the inferences drawn from many commonly-used evolution education measurement instruments. As Mead et al. emphasize in their review, it is the *inferences* drawn from measurement tools that we seek to validate, not the instruments themselves. The *Standards* have served an essential role in the design of measurement instruments for a half century, but they remain largely unknown to evolution educators. Mead et al. review the diverse range of evidence types that are commonly used to support claims about the meanings of instrument scores. Yet as their study reveals, many forms of evidence are lacking for evolution education instruments, and such evidence has been drawn from a narrow range of participant populations (e.g., particular universities, student levels, and demographic groups). This situation is not unique to evolution education, however. A parallel study of instrument quality in genomics and bioinformatics education found that > 90% of published studies lacked *any* form of supporting evidence (Campbell and Nehm 2013). Clearly, much more work is needed to expand the forms of the evidence that are used in instrument development and validation. The analytical framework and supporting literature review that Mead et al. provide help to capture a valuable "snapshot" of the field as is stands today, as well as to guide further research on instruments and instrumentation. Most striking in their review was the number of times many of the instruments were administered in some novel way

without providing additional evidence that the inferences drawn from an instrument were valid and reliable.

Although genetic drift is an important evolutionary process, only one measurement instrument-the GeDI-has been developed to measure knowledge and misconceptions of it. Tornabene et al. (2018) explore the psychometric properties of the instrument using an updated theoretical framework (Rasch analysis) and a new participant sample (undergraduates from the northeastern United States). The study reviews the conceptual weaknesses of the most commonly-used analytical framework for developing and evaluating instruments (Classical Test Theory) and highlights the comparative strengths of Item Response Theory and Rasch Analysis. Tornabene et al. also explore whether the delivery of the GeDI instrument (i.e. the order of the item packets presented to students) impacts measures of performance. The study corroborates many claims advanced in the original validation study, bolstering confidence in the quality of the inferences drawn from GeDI scores. The study also highlights some improvements that would further differentiate high and low performing students. Overall, the study illustrates how validation is a never-ending process, and not one that ends with publication of a new instrument.

Three articles—Sbeglia and Nehm (2018), Romine et al. (2018), and Barnes et al. (2019)—focus their attention on three instruments (i.e., MATE, I-SEA, and GAENE) in order to better understand the measurement of evolution acceptance. The three articles provide an interesting study in contrasts given that they approach the challenge of evaluating these tools from different theoretical (instrument/construct vs. item focused) and psychometric (Classical Test Theory vs. Rasch) vantage points. As such, the three studies help to illustrate the diverse theoretical and methodological approaches that may be used to scrutinize the inferences that may be drawn from instrument scores.

Sbeglia and Nehm (2018) provide a detailed evaluation of the GAENE instrument. Their study advances work on the instrument in several ways. First, it uses a much larger and more diverse participant sample than the original study in a pre-post study design. Second, it examines whether the rating scale (e.g., agree, disagree) used to elicit student responses functions as anticipated. Third, it explores the impact of race, gender, and degree plan (biology majors and non-majors) on the acceptance measures produced by the instrument. Corroborating recent findings by Metzger et al. (2018), the study identifies several problematic aspects of item functioning that need to be addressed. In addition, it notes significant differences in Rasch measures between races and genders, but no significant differences between students pursuing biology vs. other degree plans. Finally, the study

illustrates that measuring pre-post evolution acceptance using the GAENE and MATE does not produce different inferences about evolution acceptance magnitudes–both indicate that completing an evolution-focused class was associated with a small effect.

Given that three instruments—MATE, I-SEA, and GAENE—have been developed to measure evolution acceptance, and varying perspectives exist regarding the benefits or drawbacks of each one, Romine et al. (2018) take a fresh approach and administer all 57 items from the three instruments as a single corpus to a large student sample. Perhaps the most intriguing finding of their study is that the 38 positively-worded items from across the instruments hang together as a dimension that Romine et al. refer to as the "acceptance of the truth of evolution", and the 19 negatively-worded items from across the instruments hang together as a dimension that they refer to as the "rejection of incredible ideas about evolution". Romine et al. argue that the empirical distinction between *acceptance of truth* and *rejection of incredible ideas* aligns with theoretical perspectives about evolution acceptance and conceptual change theories in science education (see Posner et al. 1982). Specifically, Romine et al. argue that their *acceptance of the truth of evolution* dimension aligns with the intelligibility and plausibility aspects of conceptual change theory, and the *rejection of incredible ideas about evolution* captures students' dissatisfaction with non-scientific ideas and the likelihood that they find the scientifically-acceptable idea fruitful. Overall, Romine et al. bring novel insights to the debate concerning the functioning of these three measurement instruments while advancing theoretical perspectives on the construct of evolution acceptance.

Taking a similar approach as Romine et al. (2018), Barnes et al. (2019) administered items from several evolution acceptance instruments as a single corpus. In contrast to the approach taken by Romine et al. (see above), Barnes et al. analyzed the findings from each instrument separately. By employing the same participant sample, Barnes et al. were able to explore the similarities and differences in scores generated by the instruments and thereby examine how instrument choice would impact results and conclusions about evolution acceptance. An impressive aspect of the study is the diversity of participant background variables collected (e.g., parental education, political affiliation, religiosity, religious affiliation). The study findings corroborate several themes from prior work: religiosity was a statistically significant negative predictor across all instruments, the nature of science was a statistically significant positive predictor of evolution acceptance across all evolution acceptance instruments, and the strength and statistical significance of the relationship between evolution understanding

and evolution acceptance was variable depending on the instrument chosen. The study provides many insights into contradictory findings in the evolution education literature.

Beggrow and Sbeglia (2019) provide an interesting point of departure from the other articles in the collection by exploring the assessment of evolutionary knowledge and understanding across two *different* degree programs (anthropology and biology) within the *same* university. In addition to analyzing evolutionary reasoning across these educational contexts, Beggrow and Sbeglia develop new, human-focused versions of the ACORNS instrument in order to examine anthropology and biology students' reasoning about human evolutionary change. The investigation explores an important cognitive question: it attempts to tease apart the roles that familiarity with taxa and traits play when measuring reasoning about evolutionary gain and loss. Their study of evolutionary reasoning about humans builds upon a growing body of work showing that organismal contexts used to uncover student reasoning have significant impacts (Nehm 2018; Sbeglia and Nehm 2019). It also highlights the complexity of comparing evolutionary reasoning in the two student groups given that they differ dramatically in background characteristics. The differences between populations in the study complicate the question of whether learning about evolution using human examples is an effective educational strategy.

Next steps in evolution assessment

Although the collection of articles in this special issue encompasses a diverse array of educational contexts, methodologies, and theoretical vantage points, many more topics remain in urgent need of attention. For example, given the growing adoption of the Next Generation Science Standards (National Research Council 2012b) by U.S. states, the most conspicuous absence in this special issue is articles exploring how to assess scientific practices such as explanation, argumentation, and model building within the context of evolutionary biology (e.g., Passmore et al. 2017, pp. 127–30). Some studies have focused on the assessment of evolutionary explanations (e.g., Kampourakis and Nehm 2014; Nehm 2018) but this work has yet to fully engage in the assessment of the epistemic underpinnings of such tasks (e.g., What "counts" as an evolutionary explanation? What epistemic features are most worthy of emphasis?). Evolutionary biologists and philosophers of biology must work with educators to craft guidelines for introducing these often tacit disciplinary practices.

Enactment of scientific practices (e.g., argumentation, explanation) in educational settings is necessarily reliant on student-generated language and discourse, and

yet nearly all of the assessments in evolution education remain multiple-choice or Likert-scale formats. Next-generation methods like machine learning offer promising solutions for the automated analysis of text (e.g. Nehm et al. 2012; Moharreri et al. 2014) but they remain restricted to simple tasks grounded in commonly-assessed domains (e.g., natural selection). Clearly, the assessment of evolutionary practices will require innovative tools, methods, and disciplinary frameworks. The development of language-rich, technology-based, next-generation assessments remains a crucial next step for the evolution education community.

In addition to expanding the *types* of domain-specific skills being assessed (e.g., argumentation) and corresponding measurement *tools* (e.g., text analysis), many *concepts* within evolutionary biology lack rigorously-developed instruments. A recent review of the literature by Ziadie and Andrews (2018) provides a helpful overview of these conceptual gaps (e.g., speciation, population genetics, macroevolution). Coupled with Mead et al.'s (2019) quality control framework for evolution education instruments, Zaidie and Andrews have paved a path for future and emerging scholars in evolution education interested in measurement and assessment.

Finally, just as evolutionary biologists need to update their awareness and understanding of new tools and techniques, the evolution education community would benefit from professional development opportunities for enhancing instructor understanding of the theories and methods undergirding evolution assessments. For example, finding and using new phylogenetic algorithms or genomic analysis tools may be relatively straightforward, but understanding how they work and their limitations is often less apparent. The same is true of educational measurement instruments. While it is easy to obtain and use a measurement tool, understanding how it works and the limitations inherent to it are complex. Evolution educators would benefit from opportunities to learn about the theories of validity used to conceptualize educational measurement (e.g., construct validity, validity as argument), the methodologies available for analyzing and interpreting scores (e.g., Rasch, IRT), and the psychosocial factors at play in administering them (e.g., stereotype threat).

Scientific societies focusing on evolution should offer opportunities at national conferences for scientists and educators to gain awareness of the benefits and drawbacks of using the tools that the research community has generated, and foster collaborations among scientists and educators to tackle many of the challenges outlined above. Evidence-based evolution education is an essential target, but it requires an understanding of the strengths and weaknesses of the evidence used to forge claims, guide reform, and evaluate actions. Hopefully, the articles in this special issue will help to generate momentum in this important area of evolution education.

Author details
[1] Department of Ecology and Evolution, Institute for STEM Education, Stony Brook University, Stony Brook, NY 11733, USA. [2] BEACON Center for the Study of Evolution in Action, Michigan State University, 567 Wilson Rd., East Lansing, MI 48824, USA.

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

Published online: 21 February 2019

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

Barnes ME, Dunlop HM, Holt EA, Zheng Y, Brownell SE. Different evolution acceptance instruments lead to different research findings. Evo Ed Outreach. 2019. **(in press)**.

Beggrow EP, Sbeglia GC. Do disciplinary contexts impact the learning of evolution? Assessing knowledge and misconceptions in anthropology and biology students. Evol Ed Outreach. 2019;12:1.

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.

Furrow RE, Hsu JL. Concept inventories as a resource for teaching evolution. Evo Ed Outreach. 2019. **(in press)**.

Handelsman J, Miller S, Pfund P. Scientific teaching. New York: W.H. Freeman; 2006.

Kampourakis K, Nehm RH. History and philosophy of science and student explanations and conceptions. In: Matthews M, editor. Handbook of the history and philosophy of science in science and mathematics teaching, vol. I. New York: Springer; 2014. p. 377–400.

Liu X. Using and developing measurement instruments in science education: a rasch modeling approach. Charlotte: Information Age; 2010.

Mead L, Kohn C, Warwick A, Schwartz K. Applying measurement standards to evolution education assessment instruments. Evo Edu Outreach. 2019;. https://doi.org/10.1186/s12052-019-0097-y.

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.

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

Moharreri K, Ha M, Nehm RH. EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. Evo Edu Outreach. 2014;7(1):15. https://doi.org/10.1186/s12052-014-0015-2.

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

National Research Council. Scientific research in education. Washington, D.C: National Academies Press; 2002.

National Research Council. Thinking evolutionarily. Washington, D.C: National Academies Press; 2012a.

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

Nehm RH. Faith-based evolution education? Bioscience. 2006;56(8):638–9.

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

Nehm RH, Ha M, Mayfield E. Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. J Sci Educ Technol. 2012;21(1):183–96.

Passmore C, Schwarz CV, Mankowski J. Developing and using models. In: Schwarz CV, Passmore C, Reiser BJ, editors. Helping students make sense of the world using the next generation science and engineering practices. Arlington: NSTA Press; 2017.

Pellegrino JW, DiBello LV, Goldman SR. A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. Educ

Psychol. 2016;51(1):59–81. https://doi.org/10.1080/00461520.2016.1145550.

Posner GJ, Strike KA, Hewson PW, Gertzog WA. Accommodation of a scientific conception: towards a theory of conceptual change. Sci Educ. 1982;67(4):489–508.

Romine WL, Todd AN, Walter EM. 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. Evo Edu Outreach. 2018;11:17.

Sbeglia G, Nehm RH. Measuring evolution acceptance using the GAENE: influences of gender, race, degree-plan, and instruction. Evo Edu Outreach. 2018. https://doi.org/10.1186/s12052-018-0091-9.

Sbeglia G, Nehm RH. Do you see what I-SEA? A Rasch analysis of the psychometric properties of the inventory of student evolution acceptance. Sci Educ. 2019. https://doi.org/10.1002/sce.21494.

Tornabene RE, Lavington E, Nehm RH. Testing validity inferences for genetic drift inventory scores using rasch modeling and item order analyses. Evo Edu Outreach. 2018. https://doi.org/10.1186/s12052-018-0082-x.

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. https://doi.org/10.1187/cbe.17-08-0190.