JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 47, NO. 3, PP. 358–362 (2010)

REPLY

The Future of Natural Selection Knowledge Measurement: A Reply to Anderson et al. (2010)

Ross H. Nehm,[1] Irvin Sam Schonfeld[2]

[1]*College of Education and Human Ecology, Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio 43210*
[2]*Department of Psychology, The City College, The Graduate Center, City University of New York, New York*

Received 7 July 2009; Accepted 20 July 2009

The development of rich, reliable, and robust measures of the composition, structure, and stability of student thinking about core scientific ideas (such as natural selection) remains a complex challenge facing science educators. In a recent article (Nehm & Schonfeld 2008), we explored the strengths, weaknesses, and insights provided by a detailed exploration of three commonly used measures of student thinking about natural selection in a large sample (>100) of underrepresented minority students. One of our core findings was that all of the tools we studied—including the CINS—have strengths and weaknesses that must be carefully taken into consideration by those who employ, interpret, and act upon their outcomes.

The continuous reevaluation and improvement of measurement instruments is a fundamental component of test development because of the inherent limitations of *all* methods at our disposal for capturing and quantifying student knowledge. Exploring the efficacy and generalizability of measures requires the repeated study of students from different racial and ethnic groups, geographic regions, socioeconomic and language backgrounds, and content preparations. Additionally, new methods (such as Rasch analysis) allow more accurate and precise evaluations of instrument properties. Furthermore, many science assessment developers have ignored the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), which should be applied to all measures. We view Anderson, Fisher, and Smith's (AFS) (2010) defense of the CINS as sacrosanct to be antithetical to the spirit and reality of instrument development, evaluation, and improvement.

Looking Back: Criticizing Bishop and Anderson's Test, Defending the CINS

AFS's criticism of a test by Bishop and Anderson (1990) does nothing to support the quality of the CINS. Nevertheless, many of their criticisms are misleading and demand attention. AFS's criticism of Bishop and Anderson's ''salamander'' item, for example, is misinformed. The term ''blind cave salamander'' has in fact been used by the leading science journals of the U.S.A. and Europe (*Science* and *Nature*); in numerous published journal articles; and in the names of U.S. endangered species (e.g., Hendricks & Kezer, 1958; Hervant, Mathieu, & Durand, 2001; Kos, Bulog, Szél, & Röhlich, 2001; Sket 1997; Springer, 2007; Staff, 2006). If the term is misleading and promotes faulty conceptions, as AFS claim (without evidence), then we must ask why scientists use the term in their journals and in species names. Furthermore, some cave salamanders are not only blind but *lack eyes* (Fenolino, 2009): ''Salamanders such as *Eurycea rathbuni*,

Correspondence to: R.H. Nehm; E-mail: nehm.1@osu.edu
DOI 10.1002/tea.20330
Published online 1 December 2009 in Wiley InterScience (www.interscience.wiley.com).

E. waterlooensis, and *E. wallacei* all have only vestigial eyes or nothing at all . . . These characters are fixed in all known populations of these species.'' AFS also wrote, ''It is misleading to suggest that something as complex as sight could 'evolve away' entirely after a short time in a cave.'' They are also mistaken in their belief that complex features (such as eyes) cannot evolve rapidly in evolutionary terms (<10,000 years; Sket, 1997). Cave blindness is widespread in the natural world and *not* restricted to salamanders; according to Espinasa and Espinasa (2005) more than 100 *species* of fish living in caves ''. . . are blind or have some degree of eye degeneration.'' Finally, those familiar with Darwin's writings will be aware that blind cave animals were used in his popular *magnum opus* to *promote* evolutionary understanding (Darwin, 1859, notably pp. 137–138; Costa, 2009).[1]

AFS's criticism of Bishop and Anderson's (1990) ''duck'' item is confusing given that our study neither used it nor discussed it.

Many of AFS's points can only be interpreted as arising from an inattentive review of our article: First, the criticism of us that we noted a lack of *concurrent* validity for the CINS is remarkable; a careful reading of our article reveals that we never commented on the instrument's *concurrent* validity. Rather, we zeroed in on the CINS vis-à-vis *construct* validity, which is why we concerned ourselves with the convergent and discriminant validity of the CINS, the ORI, and the Oral Interview. If AFS do not accept our discriminant validity results for the CINS (which were in fact *supportive*) then their instrument currently fails to meet yet another quality control standard. We urge AFS to complete such a study. Regarding the nuances of our rock test, readers are welcome to explore a wide array of psychometric details in Duncan-Poitier (2009) and Pearson Educational Measurement (2006). AFS's conceptualization of discriminant validity runs counter to established perspectives on the subject (e.g., Anastasi & Urbina 1997, pp. 129–130).

Second, AFS claim that we presented *no* evidence that the CINS may overestimate key concepts. We encourage readers to examine Figure 1 of our 2008 article for one example. Using the CINS, nearly 70% of our sample purportedly understood the concept of ''population stability''; yet 0% ($n > 100$) of the *same* students employed this concept in their essay responses (or oral interviews). A 70% measurement difference should qualify as evidence of overestimation.

Finally, while AFS claim we ''ignored'' their PCA, we in fact directly compared Anderson et al.'s (2002) PCA results to ours (N&S, 2008, p. 1145): ''Thus, unlike Anderson et al.'s (2002) sample of non-majors, we did not find strong support for the different (PCA) components representing distinct evolutionary concepts in biology majors. Rather, we found one (PCA) factor that included a highly correlated suite of key concepts.'' We urge AFS to review our article more carefully.

AFS discuss a series of findings in support of the CINS that have neither been published nor peer reviewed. Surely the authors understand that the science education community holds itself to higher standards of evidence; we have no way of evaluating such claims.

Numerous CINS items display unacceptable levels of discriminability *and* difficulty values using *any* psychometric standard or methodology (CTT or IRT). Additionally, no items were matched to high performers in our sample (N&S, 2008, Table 9). It is worrisome that several CINS items examined in our 2009 study appeared to display DIF. No *ad hoc* explanations will mitigate these troublesome findings.

Even if all of the above problems with the CINS could be addressed, we would still be faced with the fact that the CINS does not inform an instructor as to whether a student understands *natural selection*; knowing all the ''pieces'' or elements of the theory of natural selection is not indicative of understanding of how these elements work together in Darwin's causal model (cf. Resnick & Resnick 1992). The CINS attempts—with some success—to assess the disarticulated fragments of this core idea; but it fails to provide any measure of students' abilities in regard to the degree to which they can assemble the pieces into a coherent and functional explanatory structure.

Looking Forward: The Future of Natural Selection Assessment

One of the most significant validity threats facing concept inventories like the CINS is overcoming what may be referred to as the ''either-or'' forced-choice (''misconception'' vs. scientific key concept[2]) item preference endemic to certain types of multiple-choice tests. Our work has shown that majorities of students harbor heterogeneous conceptions of natural selection comprised of *both* scientifically accurate *and*

contextually inaccurate cognitive elements in myriad models (Nehm, Haertig, & Ridgway, 2009; Nehm & Reilly 2007; Nehm, Ridgway, & Boone, 2009; Nehm & Schonfeld 2008). As our 2008 article demonstrated, the either-or constraints of the CINS channeled students into a ''false choice'' landscape, which subsequently produced an overestimation of key concepts and/or the selection of conceptions that did not appear to be held or prioritized by students (e.g., N&S, 2008, Fig. 1). Indeed, our use of open response instruments—coupled with oral interviews—clearly revealed ''mixed models'' or ''synthetic models'' of student thinking in regard to natural selection (see Ha & Cha 2009 for a cross-cultural example). Simply put, our study showed that the CINS misdiagnoses student thinking in some cases, which is a serious problem for a ''diagnostic test''. Using such results to plan curriculum, for example, could direct instructional attention away from spuriously prominent conceptions while ignoring more prevalent and problematic mental models. This is why we recommended that future research using the CINS also include an open-response test to compensate for these forced-choice constraints (*contra* AFS's interpretation of our results.

Forward-looking approaches that attempt to mitigate mental model misdiagnosis caused by forced-choice include: (1) Bao and Redish's (2006) ''Model Analysis'' of suites of carefully constructed closed-response items; (2) *synthetic model* multiple-choice instruments (as opposed to forced-choice *elements*); (3) Diagnostic Question Clusters (D'Avanzo et al., 2008); and (4) Computerized Lexical Analyses that mine open-response text in order to assemble valid representations of students' extant mental models while minimizing the human labor involved in grading (Haudek, Moscarella, Urban-Lurain, Merill, & Sweeder, 2009; Nehm, Haertig, & Ridgway, 2009). Collectively, these new models and methods offer exciting opportunities for science educators to more accurately understand student thinking about natural selection. Indeed, we are unsure whether ''misconceptions,'' ''alternative conceptions,'' or ''pieces'' theories (*sensu* DiSessa, 2008) best describe student thinking about natural selection because of a lack of rich and robust measurement instruments. For now, those terms serve as imprecise but convenient semantic placeholders.

Such new approaches are also of central importance to a significant finding in our article that was unappreciated by AFS: Item difficulty patterns and their significance for natural selection assessment design. Although AFS accept that their ''parallel'' items (testing for the *same* scientific concept) in fact have significantly *different* difficulties, they bypass the implications of this pattern; it is clear that what ''experts'' view as comparable items are not interpreted by students as such because superficial item features elicit different schemas—or activate different cognitive resources—producing the downstream effect of different item difficulty patterns (Chi, Feltovich, & Glaser, 1981). These difficulty patterns are fertile ground for unearthing the cognitive processing associated with evolutionary problem solving as well as redesigning test items that are consonant with such understanding (as urged by the NRC, 2001).

Conclusion

In summary, the CINS, like all assessment instruments, has limitations that must be carefully considered prior to use and constantly evaluated and revised in accordance with the AERA/APA/NCME *Standards*. The fact that the intrinsic constraints of *elemental* (vs. *synthetic*) forced-choice instruments (like the CINS) lead in some cases to faulty diagnoses of student mental models—and that ''parallel'' items display significantly different difficulties because of superficial item feature differences and contexts—is unsurprising but important. Furthermore, teachers should be less interested in tests that can only reveal isolated fragments of student thinking and be more interested in tests that can reveal how students choose to assemble and employ these elements in explanatory models. Fortunately, the new tools, technologies, and conceptual models that we discussed offer a wealth of exciting opportunities and hold great promise for freeing our community from the constraints of fragmented multiple choice assessment models. We invite AFS—and the community at large—to join us in our efforts to envision and build more sophisticated and valid models of science education assessment.

Notes

[1]Interestingly, much like the students in N&S (2008), Darwin appears to display a ''mixed'' or ''synthetic'' model of evolutionary loss comprised of both inaccurate elements (e.g., use and disuse) and ''key concepts'': ''. . . in the case of the cave-rat natural selection seems to have struggled with the loss of light and to have increased the size of the eyes; whereas with all the other inhabitants of the caves, disuse by itself seems to have done its work.'' (Darwin, 1859, pp. 137–138).

[2]Although Mayr indeed writes that ''population stability'' is central to how *Darwin* constructed natural selection, many other experts (e.g., Endler, 1992, p. 220; Lewontin, 1978; Patterson, 1978, p. 1; Pigliucci & Kaplan 2006, p. 14) do *not* mention population stability as one of the essential elements of natural selection. Our study of evolution experts ($n = 10$; Nehm, in preparation) likewise revealed that 0% employed this idea in their evolutionary responses to ORI items. We encourage more work on this issue, rather than exclusively appealing to the authority of Mayr (AFS, 2009). While some evolutionary biologists question the exclusive role of natural selection in *speciation* (e.g., Gould, 2002), AFS ignore such dissent and include it as a CINS ''key concept.'' It is unclear as to whether Mayr considered ''speciation'' as a key element of *natural selection*—or whether a consensus exists in regard to this issue (Gould, 2002). Regardless, AFS falsely attribute to Mayr the idea that population stability should be included in a diagnostic, criterion-referenced test assessing natural selection. He made no such claims.

References

AERA/APA/NCME. (1999). Standards for educational and psychological testing. Washington, DC: Author.

Anastasi, A., &Urbina, S. (1997). Psychological testing. 7th Ed. Prentice Hall: Upper Saddle River, NJ.

Anderson, D.L., Fisher, K.M., & Norman, G.J. (2002). Development and evaluation of the conceptual inventory of natural science. Journal of Research in Science Teaching, 39, 952–978.

Anderson, D.L., Fisher, K.M., & Smith, M. (2010). Support for the CINS as a Diagnostic Conceptual Inventory: Response to Nehm & Schonfeld (2008). Journal of Research in Science Teaching, 47, 354–357.

Bao, L., & Redish, E.F. (2006). Model analysis: Representing and assessing the dynamics of student learning. Physical Review Special Topics—Physics Education Research, 2, 010103. 10.1103/PhysRevSTPER.2.010103.

Bishop, B., & Anderson, C. (1990). Student conceptions of natural selection and its role in evolution. Journal of Research in Science Teaching, 27, 415–427.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121–152.

Costa, J.T. (2009). The annotated Darwin: A fascimile of the first edition of on the origin of species. Cambridge, MA & London, England: Harvard University Press.

Darwin, C. (1859). On the origin of species by means of natural selection. London: John Murray.

D'Avanzo, C., Morris, D., Anderson, A., Griffith, A., Williams, K., & Stamp, N. (2008). Diagnostic Question Clusters to Improve Student Reasoning and Understanding in General Biology Courses: Faculty Development Component. Proceedings of the CABS II conference. Available online at: http://bioliteracy. net/manuscripts08.pdf.

DiSessa, A. (2008). A bird's-eye view of the ''pieces'' vs. ''coherence'' controversy (from the ''pieces'' side of the fence). In S. Vosniadou (Ed.), International handbook of conceptual change (pp. 35–60). New York & London: Taylor and Francis.

Duncan-Poitier, J. (2009). New York State Testing Program science exam development. http://www.emsc.nysed.gov/ciai/mst/science/sciexamdev.htm.

Endler, J.A. (1992). Natural selection: Current uses. In E.F. Keller & E.A. Lloyd (Eds.), Keywords in evolutionary biology (pp. 220–224). Cambridge: Harvard University Press.

Espinasa, L., & Espinasa, M. (2005). Why do cave fish lose their eyes? A Darwinian mystery unfolds in the dark. Natural History, 114(5), 44–46.

Fenolino, D. (2009). E-mail letter to Nehm 5/28/2009.

Gould, S.J. (2002). The structure of evolutionary theory. Cambridge: Harvard University Press.

Ha, M., & Cha, H. (2009). Pre-service Teachers' Synthetic View on Darwinism and Lamarckism Paper presented at the *National Association for Research in Science Teaching* conference, Anaheim, CA.

Haudek, K., Moscarella, R., Urban-Lurain, M., Merill, J., & Sweeder, R. (2009). Using Lexical Analysis Software to Understand Student Knowledge Transfer between Chemistry and Biology. Paper presented at the *National Association for Research in Science Teaching* conference, Anaheim, CA.

Hendricks, L.J., & Kezer, J. (1958). An unusual population of a blind cave salamander and its fluctuation during one year. Herpetologica, 14, 41–43.

Hervant, F., Mathieu, J., & Durand, J. (2001). Behavioural, physiological and metabolic responses to long-term starvation and refeeding in a blind cave-dwelling (*Proteus anguinus*) and a surface-dwelling (*Euproctus asper*) salamander. Journal of Experimental Biology, 204(2), 269–282.

Kos, M., Bulog, B., Szél, A., & Röhlich, P. (2001). Immunocytochemical demonstration of visual pigments in the degenerate retinal and pineal photoreceptors of the blind cave salamander (*Proteus anguinus*). Journal Cell and Tissue Research, 303, 15–25.

Lewontin, R. (1978). Adaptation. Scientific American, 239, 212–228.

National Research Council. (2001). Knowing what students know. Washington, DC: National Academy Press.

Nehm, R.H. (manuscript in preparation). Problem solving performance in evolution experts and novices.

Nehm, R.H., Haertig, H., & Ridgway, J. (2009). Human vs. Computer Diagnosis of Mental Models of Natural Selection: Testing the Efficacy of Lexical Analyses of Open Response Text. *Transforming Undergraduate Biology Education: Mobilizing the Community for Change* conference, July 15–17, Washington, DC.

Nehm, R.H., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. Bioscience, 57(3), 263–272.

Nehm, R.H., Ridgway, J., & Boone, B. (2009). Exploring Differential Item Functioning (DIF) in the Measurement of Student Knowledge and Misconceptions of Natural Selection. Paper presented at the *National Association for Research in Science Teaching* conference, Anaheim, CA.

Nehm, R.H., & Schonfeld, I. (2008). Measuring knowledge of natural selection: A comparison of the CINS, and open-response instrument, and oral interview. Journal of Research in Science Teaching, 45(10), 1131–1160.

Patterson, C. (1978). Evolution. Ithaca: Cornell University Press.

Pearson Educational Measurement. (2006). *New York State Regents Examinations. Earth Science June 2006 Administration Technical Report*. 57 pp. New York State.

Pigliucci, M., & Kaplan, J. (2006). Making sense of evolution: The conceptual foundations of evolutionary biology. Chicago: University of Chicago Press.

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gilford & M.C. O'Conner (Eds.), Changing assessments: Alternative views of aptitude achievement and instruction (pp. 37–75). Boston: Kluwer.

Sket, B. (1997). Distribution of *Proteus* (Amphibia: Urodela: Proteidae) and its possible explanation. Journal of Biogeography, 24, 263–280.

Springer, C. (2007). The Texas blind salamander; San Marcos National Fish Hatchery and Technology Center. *Endangered Species Bulletin*, February 1, 2007, p. 16(2).

Staff. (2006). Troglobyte. *New Scientist* May 20, p. 56.