International Journal of Science Education, 2014
http://dx.doi.org/10.1080/09500693.2014.938376

Assessing the efficacy of the Measure of Understanding of Macroevolution as a valid tool for undergraduate non-science majors

William Lee Romine[a] and Emily Marie Walter[b*]
[a]*Biological Sciences, Wright State University, Dayton, OH, USA;* [b]*Physics, Western Michigan University, Kalamazoo, MI, USA*

Efficacy of the Measure of Understanding of Macroevolution (MUM) as a measurement tool has been a point of contention among scholars needing a valid measure for knowledge of macroevolution. We explored the structure and construct validity of the MUM using Rasch methodologies in the context of a general education biology course designed with an emphasis on macroevolution content. The Rasch model was utilized to quantify item- and test-level characteristics, including dimensionality, reliability, and fit with the Rasch model. Contrary to previous work, we found that the MUM provides a valid, reliable, and unidimensional scale for measuring knowledge of macroevolution in introductory non-science majors, and that its psychometric behavior does not exhibit large changes across time. While we found that all items provide productive measurement information, several depart substantially from ideal behavior, warranting a collective effort to improve these items. Suggestions for improving the measurement characteristics of the MUM at the item and test levels are put forward and discussed.

Keywords: *Macroevolution; Biology; College science; Assessment; Rasch measurement*

Evolution is a central organizing principle of the biological sciences (American Association for the Advancement of Science [AAAS], 1993; Dobzhansky, 1973; Kagan, 1992; National Research Council [NRC], 1996). Despite this, only 4 of 10 Americans agree that humans and higher apes share a common ancestor (Newport, 2009). Among Americans with some college education, 49% accept the theory of

*Corresponding author. Physics, Western Michigan University, Kalamazoo, MI, USA. Email: emily.walter@wmich.edu

evolution for plants and non-human animals, and only 22% agree that humans evolved from earlier species (Lovely & Kondrick, 2008). Since evolution is a critical and foundational component of scientific literacy (AAAS, 2011; Bybee, 1997; National Association of Biology Teachers, 2010), and the university setting is one of the final opportunities to influence scientific literacy of college-educated individuals, it is vital to understand what undergraduates know about evolution.

It is critical to have valid and reliable instruments to accurately measure undergraduates' knowledge as instructors and researchers examine their students' evolution thinking and learning. Instrument development in the past has focused on students' understanding of natural selection, including the Conceptual Inventory of Natural Selection (Anderson, Fisher, & Norman, 2002) and an open response instrument by Nehm and Schonfeld (2008). These instruments measure students' knowledge of microevolutionary concepts, i.e. the genetic variation within a population (i.e. allele frequencies) due to natural selection (Campbell & Reece, 2005). Macroevolutionary concepts, in contrast, include the formation, radiation, and extinction of species and higher order natural groups comprising a most recent common ancestor and all its descendants over long periods of time (Campbell & Reece). Until recently, instruments to measure knowledge of macroevolution concepts were limited to the Tree Thinking Concept Inventory, an unpublished instrument assessing knowledge of phylogenetic trees (Naegle, 2009).

Measure of Understanding of Macroevolution

Nadelson and Southerland (2010) filled a gap in the literature through development of the Measure of Understanding of Macroevolution (MUM) to comprehensively measure students' knowledge of macroevolution. This 27-item dichotomous multiple choice instrument was designed to measure five dimensions related to the understanding of macroevolution: deep time (5 items), phylogenetics (4 items), speciation (8 items), fossils (5 items), and the nature of science (5 items). These qualitatively defined dimensions were derived through considerable background research, including an overview of the National Science Education Standards for high school biology (NRC, 1996), the learning progression work of Catley, Lehrer, and Reiser (2005), and the contents of three undergraduate-level textbooks on evolution (Barton, Briggs, Eisen, Goldsten, & Nipam, 2007; Freeman & Herron, 2007; Futuyma, 2005).

Items underwent a thorough two-round content review. In the first round, a draft of the instrument was sent to 20 evolution education experts. A revised draft was sent to five college biology faculty members for further evaluation of content accuracy and consistency with the college biology curriculum, and four undergraduate education students for suggestions on improving readability for college students. The authors then gave the assessment to three cohorts of college students, with 667, 74, and 54 participants, respectively.

Construct validity of the MUM was evaluated through the perspective of classical test theory (CTT). This included the use of the internal consistency definition of reliability measured by Cronbach's alpha and the point-biserial correlation (the correlation between a correct response and total score) as a measure of item discrimination. The instrument measured the students in their study with a reliability of 0.82, indicating satisfactory measurement precision. All items had positive point-biserial correlations, indicating that students scoring higher on the assessment have a greater tendency to get the item correct than low scorers, a fundamental criterion for construct validity.

Criticism of the MUM

Despite the documented need for an instrument measuring knowledge of macroevolution (Catley, 2006), the MUM has undergone a significant amount of criticism. Much of this is summarized in Novick and Catley (2012), who 'recommend that this test not be used by researchers and educators interested in evolution education until a suitable revision is available that addresses the issues raised here' (p. 2685).

Novick and Catley raise a number of criticisms against the MUM. While Nadelson and Southerland claim satisfactory reliability and validity, Novick and Catley disagree, stating that the MUM should not be used in its current form due to the lack of validity evidence. Although they acknowledge that the MUM has a strong theoretical foundation and content validity, they criticize the MUM for its lack of criterion validity and construct validity. Novick and Catley proceed to raise issues regarding the wording of test distracters. Given the wording of distracters, they hypothesized that students could get the answers correct without even looking at the question stem and proceeded to test this hypothesis following the precedent of Katz, Lautenschlager, Blackburn, and Harris (1990). They gave the test to a sample of 69 college students and ran item-level *t*-tests to test the null hypothesis that the probability of a student getting the correct answer is no better than chance, which they initially defined as 0.25. Taking into account the proportion of students selecting each distracter, they calculated item availability and reran the *t*-tests against a corrected chance estimate for each item. The idea here is that if a student eliminates a distracter, then the null probability of guessing the correct answer increases. Assuming the 0.25 null probability, they rejected the null hypothesis for 22 of the 27 items. The null hypothesis was rejected for 19 of the 27 items when corrected chance estimates were used. These analyses are indicative that college students do not need the stem to get items on the MUM correct, which the authors use as evidence for the lack of construct validity.

Problems Remain

It is intriguing that 69 students in the Novick and Catley study were able to select the correct answer without the stem for many items with a probability greater than chance. However, there are important problems with their study design and underlying assumptions that need to be addressed before concluding that the MUM is invalid, poorly written, and useless as a research tool. The primary issue involves the definition of guessing that was used to generate null hypothesis criteria for the item-level *t*-tests. The use of the uncorrected chance null probability of 0.25 for a four-choice item inherently assumes that students choose an answer by spinning a roulette wheel. The 'roulette wheel' model for guessing also served as a foundation for the corrected chance null probability, only acknowledging that one or more poorly worded distracters could have been eliminated before students randomly chose a response. If the MUM were given to a randomly guessing population, the roulette wheel model for guessing might hold credence. However, it is reasonable to expect that undergraduate students would critically evaluate each response, even without the question stem, and attempt to select the most reasonable answer based on their prior knowledge of macroevolution.

Second, the use of these results as evidence for the lack of construct validity would imply that selecting answers without reading the stem had little relationship with an individual's knowledge of macroevolution. After discussing their findings as evidence for the lack of construct validity, they acknowledge: 'these correct and incorrect answer alternatives could still be good indicators of students' understanding of macroevolutionary concepts if examinees were simply asked to evaluate each one on its own merits' (p. 2699).

The final potential problem muddying the interpretation of the Novick and Catley (2012) results is that the MUM without the item stems was administered after giving two previous test booklets containing items related to macroevolution with their stems. The authors acknowledge the possibility that these previous tests could have affected responses on the stemless MUM. The effect of previous tests could have been quantified by giving the stemless MUM first to a portion of the participants, or avoided completely by doing this for all participants.

Purpose of the Research

Novick and Catley (2012) call for a study validating the MUM through item response theory (IRT) or CTT methods, tackling the issue of construct validity in particular. The purpose of this paper is to quantify the construct validity of the MUM through Rasch methodologies while simultaneously testing the claims against the construct validity of the MUM presented in Novick and Catley (2012). In addition, we take the opportunity to evaluate the stability of the MUM in the context of a general education biology course. We address four questions:

(1) What is the dimensionality of the MUM?
(2) Does the MUM show evidence for construct validity from the perspective of the Rasch model?
(3) Does guessing damage the MUM's construct validity?
(4) Do the measurement characteristics of the MUM undergo significant changes over the course of a general education biology course?

Methods

Subjects and Test Administration

The MUM in its original form was administered to a convenience sample of undergraduate students in a general education biology course at a public research-intensive university in the Midwest USA. Participants had to be 18 years of age and not majoring in biological sciences to be eligible to participate. The test was administered over an electronic learning management system (BlackBoard) during the second week of the semester and again after the evolution unit (week 15). Completing the MUM was a required component of the class; students were given points for completion of the test, but not for the correctness of their responses. Since students were not graded for the correctness of their responses, we are not concerned about students using outside resources (e.g. the internet, class notes) to answer correctly over the electronic learning management system. Three hundred and fifteen students completed the pre-test and 291 completed the post-test. Two hundred and seventy of these students completed both pre- and post-tests.

Our subjects were similar to Novick and Catley (2012) and Nadelson and Southerland (2010) with respect to year in school and prior biology coursework (Table 1). Our study differs from these studies in that we examined students within the context of a general education biology course, not individuals sampled from a

Table 1. Comparisons among the participant groups of this study, Nadelson and Southerland (2010) and Novick and Catley (2012)

	This study	Nadelson and Southerland (2010) Cohort 1	Novick and Catley (2012)
N	270 students	667 students	69 students
Major(s)	All majors except biology	Biology majors	All majors
Year	2nd semester freshmen and sophomores	1st semester freshmen	All years
Past biology instruction	Completed high school biology	Not provided; likely completed high school biology	Completed high school biology
Proportion without prior macroevolution instruction (%)	80.7	84	87
Course setting	General education biology course	Introductory biology course	Conducted outside of a course context
Course audience	Non-science majors	Biology majors	Not applicable
University setting	Large, public, and research intensive	Large, research intensive	Private, research intensive
Geographic setting	Midwest USA	Southeastern USA	Southeastern USA

general population of willing undergraduates. Our subjects could have been more actively engaged with biological concepts (evolution or otherwise) than Novick and Catley's subjects not enrolled in a biology course. However, we did not expect that course engagement to lead to a lesser likelihood of guessing, since our subjects received credit for completion regardless of performance.

Based on reported completion of various semester-long college biology courses, Novick and Catley estimated that 87% of their subjects had no previous macroevolution instruction. This study and Walter (2013) took a slightly different approach in determining previous macroevolution instruction for this sample of participants. Analyses included one-way ANOVA tests comparing MUM Rasch measures among groups from different settings of most recent evolution instruction (no instruction, elementary/middle school, high school, and college) and among different durations of past evolution instruction (no instruction, about 1 day, about 1 week, and greater than 1 week). Independent *t*-tests were also run to compare groups that had and had not completed college courses in geology, anthropology, or an additional biology course. There were no significant differences among or between any of these groups for the pre- or post-test MUM measures, with the exception of a significant difference of pre-test measures ($p < .026$) between individuals who had completed a different college-level biology course ($n = 56$) and those who had not ($n = 212$). This significant difference between groups did not carry over to the post-test measures ($p = .058$). Since four of the 56 participants reported that this college biology course did not include evolution, we estimated that 52 participants may have had prior macroevolution instruction. This left 218 of 270 participants (80.7%) who likely had no previous macroevolution instruction, a proportion similar to Novick and Catley (2012) and Cohort 1 of Nadelson and Southerland (2010; Table 1).

The participants completed an evolution unit of thirteen 50-minute class periods, nine of which focused on macroevolution-related content (Walter, 2013). Although the unit was not specifically designed to increase MUM scores, the content presented was relevant to four of five MUM dimensions (fossils, speciation, nature of science, and phylogenetics). The instructor presented six examples of transitional fossils, four examples of speciation, and reviewed the nature of science as it pertained to evolution. Additionally, students completed a 50-minute paper and pencil activity on how to interpret phylogenetic trees. Few of the examples, organisms, or phylogenetic trees presented in class were seen in items on the MUM, with the exception of items 19–23 (referring to MUM Figure 5), a phylogenetic tree depicting the evolutionary relationships of gibbons, orangutans, gorillas, chimpanzees, and humans. The subject matter of these items was similar to ideas discussed in class, but MUM Figure 5 was not used in class.

Deep time was the only topic assessed on the MUM that was not covered in detail by the intervention. The deep time questions on the MUM use specific dates for particular events. The instructor did not give such specificity for the geologic and evolutionary events she discussed. However, she did spend two 50-minute class periods discussing the history of life, including the magnitude of deep time scales and the age of the Earth.

Justification for the Rasch model

Mathematically, dichotomous IRT models, including the Rasch model, can be written as a subset of the three-parameter model, a logistic model which expresses the probability of a student of a given ability (θ) getting an item correct as a function of the item's discrimination parameter (a), the item's difficulty (b), and the probability of guessing (c).

The three-parameter logistic (3PL) model can be simplified in a number of ways (de Ayala, 2009). When guessing is not expected to be an issue, the guessing parameter can be constrained to zero. This leaves the two-parameter logistic (2PL) model, in which difficulty and discrimination parameters are allowed to vary for each item. The 2PL model can be further simplified to a one-parameter logistic (1PL) model by forcing the discrimination parameter to be the same for each item and adjusted to the value which achieves the best average fit over all items. A 1PL or 2PL model can be further simplified by constraining the discrimination parameter to be 1 for all items, thus defining the simple logistic model, also called the Rasch model (de Ayala, 2009).

Analyzing the mathematical forms of these various models begs the question of which model to choose. This question is not simply a matter of mathematical abstraction, but largely dependent on one's perspective toward test validation and has undergone considerable debate (Andrich, 2004). Proponents of multi-parameter IRT models propose an empirical approach to test validation. From a scientific perspective, theories should be discovered from data (Birnbaum, 1968; Lord & Novick, 1968). While additional scrutiny of data not conforming to one's expectations is encouraged, no reasonable scientist would throw out carefully collected data simply due to lack of conformity with a model. Rather, a revision of the previous model may be considered. From a statistical perspective, systematic error due to underfitting or overfitting can lead to biased item parameter estimates, making proper model specification a potentially important problem to address.

However, the purpose of this study was validation. In a validation context, the purpose is not to fit the data well, but rather to evaluate the quality of the data produced by the measurement tool, which is a reflection of the quality of the measurement tool itself (Wright, 1992). In light of this goal, instead of the IRT approach which is aimed at choosing a model that best fits the data, we used the Rasch approach – the Rasch model was used as a criterion upon which to evaluate the quality of the data produced by the MUM. The Rasch criterion can be expressed as follows: the probability of a student answering an item correctly should be proportional only to the difference between the student's ability and the item's difficulty (Wright & Stone, 1979). An item conforming to this fundamental criterion will fit well with the Rasch model, which equates the log-odds of a student answering an item correctly with the difference between the student's ability and the item's difficulty. Another attractive attribute of the Rasch model is that it provides a set of non-crossing curves (Andrich, 2004), allowing comparison of person and item measures on the same scale (Wright & Stone, 1979) and allows construction of a scale that is invariant across the continuum of student ability (Wright, 1992). Consequently, the lack of fit to the Rasch model can be used to identify items that provide misleading measurement information, do not discriminate well, and/or miscategorize students (Linacre, 2010).

Evaluation of Reliability, Dimensionality, and Construct Validity

Rasch analysis. We began by investigating reliability, dimensionality, and construct validity of the MUM from the Rasch perspective using BIGSTEPS software. Satisfactory fit with the Rasch model provides important evidence for construct validity of items. The idea is that if an item on the MUM is truly effective in giving a reliable score for knowledge of macroevolution, then lower achieving students should have a lower probability of getting the item correct than higher achieving students. Mean squares infit and outfit indices, which have an expected value of 1.00, were used to quantify fit with the Rasch model. Infit is an information weighted form of outfit and is thus comparatively insensitive to outliers (Bond & Fox, 2007). Items with mean squares fit indices between 0.5 and 1.5 are generally considered productive for measurement of the underlying construct from the Rasch perspective (Wright & Linacre, 1996). If the MUM is lacking in construct validity due to misleading wording and/or guessing, then we would expect to see mean squares fit indices outside of these boundaries.

Mean squares fit indices can give important insights into the behavior of an item, including discrimination and guessing (Wright & Stone, 1979). An item with moderate discrimination (close to 1) and small probability for guessing would have a mean squares fit index close to 1, indicating that the item conforms to what the Rasch model would expect. Mean squares fit indices above 1 imply error in the data with respect to the model, which would occur if (a) low-achieving students got an item correct as a result of guessing or (b) high-achieving students missed an item due to misleading wording. If guessing were a significant problem on the MUM, then we would expect mean squares fit values well above 1 for many items. We may further expect these misfitting items to have lower difficulty measures since lower ability students could answer them correctly. Conversely, a mean squares fit index below 1 indicates that the item discriminates between high and low achievers better than the Rasch model would expect. An item with a mean squares fit below 0.5 behaves suspiciously well, indicating possible influence of additional underlying dimensions positively correlated with knowledge of macroevolution (Masters, 1988). For example, if an item is complexly worded, it may measure reading ability in addition to knowledge of macroevolution. On such an item, the probability of a correct response would be inflated for a high achiever but attenuated for a low achiever.

Dimensionality of the MUM can also be evaluated using the Rasch model. If underlying dimensions are unaccounted for by the Rasch model, they should manifest through a nonrandom trend in the error (Linacre, 1998). We used principal components analysis (PCA) on standardized Rasch residuals to evaluate whether the error with respect to the Rasch model was random. A first-factor eigenvalue below

2 indicates randomness in the Rasch residuals, providing evidence for the assumption of unidimensionality (Raiche, 2005). If the MUM has five dimensions as Nadelson and Southerland (2010) state, then PCA on Rasch residuals should give a first-factor eigenvalue well above 2.

Test-level validity of the MUM across two time points was evaluated using person-item maps for pre- and post-tests, where person and item measure distributions are compared on the same scale. Since an item will provide the greatest amount of information for students with ability at the same level as the item's difficulty, significant overlap of person and item measure distributions provides evidence that the MUM provides useful measurement information for the sample. Changes in overlap between these distributions were noted on the pre- and post-test. Item-level changes were evaluated using differential item functioning (DIF) tests of Rasch difficulty measures. Under the assumption of item invariance, a two-tailed Z-test ($\alpha = 0.05$) was used to test the null hypothesis that item measures were similar at both time points.

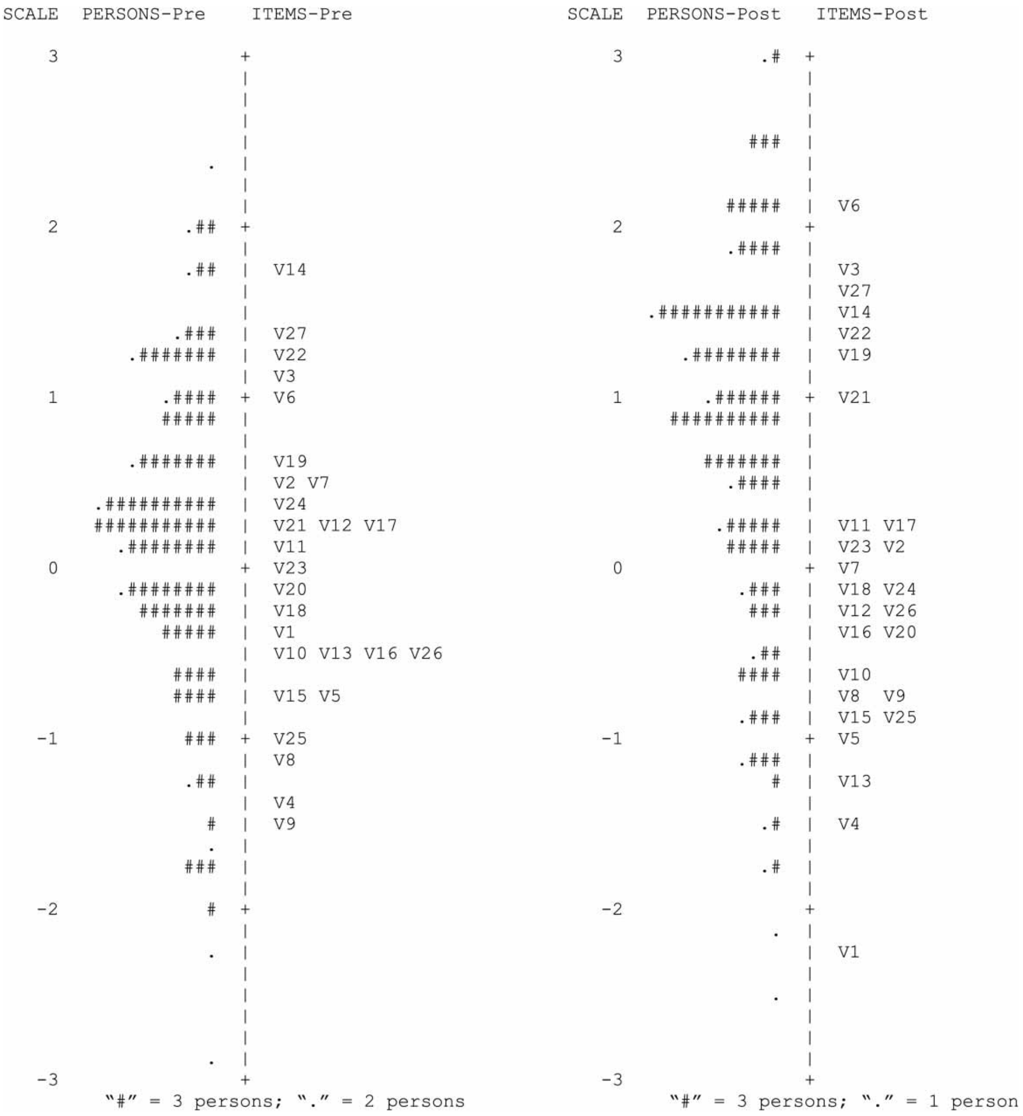
Results

Rasch Analysis

Rasch analysis indicates that the MUM is an approximately unidimensional assessment, not five-dimensional as the original authors indicate. PCA on Rasch residuals gives a first-factor eigenvalue of 1.77 items of variance for the pre-test and 2.20 for the post-test. We consider these pre- and post-test values to collectively support the assertion that the MUM is unidimensional. A value of 2.20 indicates small presence of additional factors after the intervention. It makes sense that measures for individual content topics on the MUM would be more separable in students who know more about macroevolution. However, the proximity of this value to 2 indicates that variance associated with any additional factors overlaps considerably with knowledge of macroevolution, the main factor of interest, and thus little systematic variance is left in the residuals.

Satisfactory person reliability measures suggest that the MUM precisely measures knowledge of macroevolution. We find that the Rasch model provides reliable measures for person and item parameters on both the pre- and post-tests. Reliability of person measures was 0.74 and 0.78 for the pre- and post-tests, respectively. Item measurement reliability was 0.98 for both tests. High item reliability indicates that item measures are clearly distinguishable from each other, suggesting that items provide unique tick marks along the Rasch scale.

The map of pre- and post-test item and person measures on the same scale (Figure 1) gives a visual picture of how these distributions compare. Both strengths and weaknesses of the MUM as a measurement tool become apparent from these person-item maps. Since an item provides the most information for a participant with matching ability level, test information will be maximized when item and person measure distributions match each other. A considerable strength of the



```
SCALE  PERSONS-Pre    ITEMS-Pre              SCALE  PERSONS-Post   ITEMS-Post

    3                +                           3            .# +
                     |                                           |
                     |                                           |
                     |                                           |
                     |                                       ### |
                   . |                                           |
                     |                                           |
                     |                                     ##### | V6
    2            .## +                           2               +
                     |                                     .#### |
                 .## | V14                                       | V3
                     |                                           | V27
                     |                                .########### | V14
                .### | V27                                       | V22
            .####### | V22                                .######## | V19
                     | V3                                        |
    1          .#### + V6                        1       .###### + V21
               ##### |                                ########## |
                     |                                           |
            .####### | V19                                ####### |
                     | V2 V7                                .#### |
         .########## | V24                                       |
         ########### | V21 V12 V17                         .##### | V11 V17
           .######## | V11                                 ##### | V23 V2
    0                + V23                       0               + V7
           .######## | V20                                  .### | V18 V24
             ####### | V18                                   ### | V12 V26
               ##### | V1                                        | V16 V20
                     | V10 V13 V16 V26                       .## |
                #### |                                      #### | V10
                #### | V15 V5                                    | V8  V9
                     |                                      .### | V15 V25
   -1            ### + V25                      -1               + V5
                     | V8                                   .### |
                 .## |                                         # | V13
                     | V4                                        |
                   # | V9                                     .# | V4
                   . |                                           |
                 ### |                                        .# |
                     |                                           |
   -2              # +                          -2               +
                     |                                         . |
                   . |                                           | V1
                     |                                           |
                     |                                         . |
                     |                                           |
                     |                                           |
                   . |                                           |
   -3                +                          -3               +
        "#" = 3 persons; "." = 2 persons             "#" = 3 persons; "." = 1 person
```

Figure 1. Wright maps of pre- and post-test person and item difficulty measures.

MUM is that item difficulties spread across the scale between −2 and 2, illustrating the efficacy of items in providing information on students at a range of cognitive levels. However items V10, V13, V16, and V26 on the pre-test provide maximum information about participants with ability levels around −0.5. Since there are not many participants with ability in this range, two or three of these items could likely be removed without significantly attenuating measurement reliability for college students possessing knowledge levels similar to the pre-test group.

On the post-test, which represents how the MUM may measure students with greater knowledge of macroevolution, a gap between 0 and 1 exists, and a majority of the items catch the bottom half of the ability distribution. Adding items with greater difficulty would serve as a way to raise measurement reliability for a more knowledgeable group of students. At the bottom extreme, item V1 contributes little information on the participants and therefore adds negligibly to measurement of students on the post-test; if the MUM were used for measuring biology majors, this item could potentially be removed altogether without significant attenuation of reliability.

While the person-item maps indicate that the measurement efficiency of the MUM could be improved, satisfactory reliability measures suggest that items on the MUM provide sufficient information about non-biology majors for group comparisons. A greater number of moderate-to-difficult items would serve to raise measurement reliability to a level sufficient for individual comparisons (Tennant & Connaghan, 2007).

In addition to showing unidimensionality and reliability, we find that most of the items on the MUM fit well with the Rasch model at both time points (Table 2). On the pre-test, infit indices range from 0.81 to 1.19 and outfit indices range from 0.70 to 1.57; on the post-test, infit indices range from 0.75 to 1.22 and outfit indices range from 0.58 to 1.79. All fit indices fall between 0.5 and 1.5 with the exception of the outfit indices of item V3 (pre = 1.57, post = 1.79) on both tests, and items V6 (pre = 1.27, post = 1.63) and V27 (pre = 1.31, post = 1.64) on the post-test. This indicates that these items may not discriminate well between high and low achievers due to misleading wording and/or guessing. While it is difficult to precisely quantify which phenomenon is most prevalent through the Rasch model, we observe that V3, V6, and V27 are among the most difficult items on the MUM. Since high Rasch difficulty measures result from a low proportion of students answering the item correctly, it is unlikely that the misfit of these items with the Rasch model is caused by guessing. It is more likely that ambiguous or misleading wording caused the higher ability students to misinterpret and miss these items, which would serve to inflate their difficulty. While the outfit values for these items are high, especially on the post-test, the less outlier-sensitive infit values fall within the bounds of satisfactory fit. Pre-test infit values sit at 1.17, 1.16, and 1.19 and post-test values at 1.17, 1.13, and 1.21 for V3, V6, and V27, respectively.

Rasch item analyses indicate that the MUM provides effective measures for knowledge of macroevolution. A strong case for construct validity is built through the MUM's unidimensionality, satisfactory reliability, and individual items that fit the Rasch criterion. Infit statistics between 0.5 and 1.5 for all items and outfit statistics in this range for a majority of the items also provide evidence that guessing and misleading wording are not significant issues, as Novick and Catley (2012) propose.

Stability of the MUM

DIF tests (Table 2) show that a majority of items underwent no significant changes in difficulty across the intervention. Items V1, V7, V12, V13, and V24 significantly

Table 2. Rasch difficulty measures and fit statistics for the MUM

	Rasch Item difficulty parameters and DIF tests							Rasch mean squares fit indices			
Item	b_{pre}	$\mathrm{SE_b}$	b_{post}	$\mathrm{SE_b}$	$b_{\mathrm{post}} - b_{\mathrm{pre}}$	Z_{diff}	sig	$\mathrm{Infit_{pre}}$	$\mathrm{Outfit_{pre}}$	$\mathrm{Infit_{post}}$	$\mathrm{Outfit_{post}}$
Q1	−0.43	0.12	−2.21	0.22	−1.78	−7.10	0.00	1.03	1.03	0.93	1.02
Q2	0.48	0.12	0.18	0.13	−0.30	−1.70	0.09	1.00	1.02	1.07	1.07
Q3	1.14	0.14	1.70	0.15	0.56	2.73	0.01	1.17	1.57	1.17	1.79
Q4	−1.28	0.14	−1.53	0.18	−0.25	−1.10	0.27	0.89	0.79	0.91	0.68
Q5	−0.76	0.13	−0.99	0.15	−0.23	−1.16	0.25	0.83	0.78	0.80	0.68
Q6	1.03	0.13	2.15	0.16	1.12	5.43	0.00	1.16	1.27	1.13	1.63
Q7	0.54	0.12	−0.05	0.13	−0.59	−3.33	0.00	1.07	1.17	1.00	0.97
Q8	−1.12	0.14	−0.76	0.15	0.36	1.75	0.08	0.95	1.00	0.84	0.74
Q9	−1.48	0.15	−0.78	0.15	0.70	3.30	0.00	0.81	0.70	0.75	0.62
Q10	−0.49	0.12	−0.58	0.14	−0.09	−0.49	0.63	0.95	0.92	0.96	0.93
Q11	0.12	0.12	0.22	0.13	0.10	0.57	0.57	0.98	1.02	1.08	1.12
Q12	0.16	0.12	−0.19	0.13	−0.35	−1.98	0.05	1.05	1.04	0.94	0.94
Q13	−0.55	0.12	−1.26	0.16	−0.71	−3.55	0.00	0.89	0.87	0.80	0.58
Q14	1.70	0.15	1.49	0.14	−0.21	−1.02	0.31	1.04	1.29	1.19	1.34
Q15	−0.76	0.13	−0.90	0.15	−0.14	−0.71	0.48	0.83	0.80	0.85	0.75
Q16	−0.49	0.12	−0.32	0.14	0.17	0.92	0.36	1.00	1.03	0.96	0.92
Q17	0.18	0.12	0.28	0.13	0.10	0.57	0.57	0.95	0.93	1.04	1.04
Q18	−0.29	0.12	−0.14	0.13	0.15	0.85	0.40	1.01	0.99	0.85	0.82
Q19	0.58	0.13	1.20	0.14	0.62	3.25	0.00	1.13	1.24	1.13	1.28
Q20	−0.11	0.12	−0.42	0.14	−0.31	−1.68	0.09	0.97	0.97	0.99	0.98
Q21	0.29	0.12	0.95	0.13	0.66	3.73	0.00	1.08	1.11	1.14	1.19
Q22	1.24	0.13	1.42	0.14	0.18	0.94	0.35	1.06	1.09	1.22	1.50
Q23	0.05	0.12	0.17	0.13	0.12	0.68	0.50	0.91	0.89	0.94	0.96
Q24	0.38	0.12	−0.14	0.13	−0.52	−2.94	0.00	0.97	0.99	0.86	0.81
Q25	−0.99	0.13	−0.83	0.15	0.16	0.81	0.42	0.92	0.80	0.79	0.62
Q26	−0.52	0.12	−0.29	0.14	0.23	1.25	0.21	0.94	0.88	0.94	0.85
Q27	1.35	0.14	1.60	0.15	0.25	1.22	0.22	1.19	1.31	1.21	1.64

decreased in difficulty on the post-test. This is a likely result of the topics addressed on these items receiving greater emphasis during the intervention than other topics assessed on the MUM. This is supported by the person-item maps (Figure 1), which document an increase in the student ability distribution relative to the item difficulty distribution on the post-test. Given this trend, it is somewhat surprising that items V3, V6, V9, V19, and V21 had significantly higher difficulty on the post-test than the pre-test. Possible reasons for this are discussed below.

Discussion

Construct Validity of the MUM and Suggestions for Improving Items

A number of strong criticisms against the MUM are put forth by Novick and Catley (2012), including that the MUM should not be used in its current form due to insufficient construct validity evidence. This claim runs contrary to our data. In general, we find that items on the MUM provide productive measurement information on a single underlying construct (knowledge of macroevolution). This conclusion is supported through satisfactory fit of items with the Rasch model, unidimensionality, and satisfactory reliability.

Construct validity. A major piece of evidence brought forth against the construct validity of the MUM is that students are able to guess the correct answers without reading the stem. We argue that the underlying assumptions in the Novick and Catley (2012) analyses are unrealistic, including the assumptions that students guess randomly and that choosing the correct answer is unrelated to knowledge of macroevolution. Our data indicate that, while students may be able to choose the correct response without reading the stem on some questions, much of this comes from their previous understanding of macroevolution.

Our findings also run contrary to the claim that the construct validity of the MUM is challenged by the ease of guessing the correct response. Items which showed the poorest fit with the Rasch model tended to be the more difficult items. Since prevalence of guessing would bring down the difficulty of an item, it is likely that ambiguous wording, not guessing, is the cause for the misfit of certain items.

The item-level *t*-tests and item availability analyses in Novick and Catley (2012) suggest that students are able to guess the correct answer for many items, even when students were not given a stem. Furthermore, qualitative evaluation of each item reveals a general trend that the chance of a student getting an item correct without the stem increases with the amount of information given in the answer options. For example, item V3 asks students to select when modern whales diverged from land animals. Due to the specificity of the stem and lack of information given in the answer options, most students could not answer this question correctly at a frequency higher than chance.

Item V3

According to evolutionary theory, whales have evolved from land animal ancestors over time. How much time do you think the evolution process might have taken?

*A. Fifty million years.	C. Five hundred thousand years
B. Five million years.	D. Five hundred million years

Note. *Correct answer

Conversely, Novick and Catley calculated the highest correct-without-the-stem result for item V8. In this case, we posit that the answer options in this question have enough information to determine the answer without the stem, should the student have enough knowledge of macroevolution.

Item V8

Some speculate that the eye is too complex to have resulted from evolution. Yet, evidence suggests organisms may have had eyes for nearly 500 million years. What might scientists infer about the eyes of ancient organisms?

A. Only animals living in the bright sunlight develop eyes because they need them and use them
B. Eyes would bear no resemblance to how eyes are structured today, and would not be recognized as eyes
*C. The eyes of ancient organisms would have some characteristics that are similar to eyes found in organisms alive today
D. Only animals with bones would really be trying to develop useful eyes

Note. *Correct answer

Independently of the V8 stem, students with an understanding of macroevolution could identify option C as true: The eyes of ancient organisms would have some characteristics that are similar to eye found in organisms alive today. Using similar knowledge, a student could eliminate option B as false. A student could eliminate option A using knowledge of vestigial traits, the fallacy of Lamarckian evolution, or familiarity with animals adapted to see infrared frequencies. Option D could likewise be eliminated using knowledge of Lamarckian evolution.

Since putting key information in answer options instead of the stem does not compromise construct validity, we are led to consider other potential problems with this test writing approach. One issue to consider is efficiency. Researchers often seek to get a reliable score and to administer the assessment quickly. Adding more items is an easy way to increase reliability (Cortina, 1993), but adds to the required time to complete the test. In the time it takes for a student to read and interpret the distracters for an item like V8, two or three questions with focused stems could potentially be completed (Haladyna, Downing, & Rodriguez, 2002). Revising items in the MUM to include more information in the stems could shorten test time while still preserving or improving desirable psychometric properties, such as reliability.

Novick and Catley (2012) also note that certain distracters in the MUM are implausible. Researchers could consider eliminating the least effective distracter for each item to address this issue. Few items have more than one or two effective distracters (Haladyna & Downing, 1993), and the item availability analysis indicates that this also applies to the MUM (Novick & Catley, 2012). Furthermore, items with three response options tend to offer the highest reliability in the shortest time without compromising the integrity of the item (Haladyna et al., 2002; Rodriguez, 2005). For example, option D in V8 could likely be eliminated without affecting the integrity of the item. Option A already elicits Lamarckian misconceptions, and college students would likely know that some invertebrates have eyes.

Discussion of Problematic Items

While our analyses show that the MUM in its current form has sufficient construct validity and reliability, Rasch analyses reveal potential problems with individual items that may not be noticeable through CTT methodologies. Although none of the items displayed gross misfit with the Rasch model, three items (V3, V6, and V27) had Rasch outfit statistics higher than 1.5. Due to the high difficulty of these items, poor discrimination due to ambiguous wording, not guessing, is likely to blame. This warrants a closer look at the wording of these items.

Item V27

The supercontinent depicted in Figure 6 is known as Gondwana. This supercontinent existed roughly

A. Five million years
B. One and a half billion years
*C. One hundred fifty million years
D. Three hundred and fifty thousand years

Note. *Correct answer

Items V3 (displayed previously) and V27 assess knowledge of specific time scales: V3 regarding the date of the divergence of modern whales from land animals, and V27 the time span that the Gondwana landmass existed. We suspect that these two items deviate from knowledge of macroevolution, possibly accounting for the slight departure from unidimensionality shown with respect to the Rasch model on the post-test. Finding the correct answer depends on the use of reference dates, such as the age of the Earth, the origin of life, and other events in Earth's history, as well as an understanding of geologic time scales (Murck, Skinner, & Mackenzie, 2008). In other words, a student answering correctly could understand time scales and predominant life forms during major eras, but not understand the process of macroevolution itself. This also accounts for why students with a good understanding of macroevolution (high MUM scores) may have missed this item.

Novick and Catley (2012) recommend rewriting the deep time questions (V3, V8, V15, V21, V27) to assess understanding of relative time. This would assess deep time directly in the context of macroevolution as opposed to geology (Dodick & Orion, 2003). While inclusion of these items in the MUM in their current form does not render the instrument unusable, excluding or revising these items may enhance reliability and efficiency when student memorization of the geologic column is not of direct interest.

The other problematic item was V6, a question about a hypothetical coastal sheep population swimming hundreds of meters for seaweed. Options A, B, and C give three plausible outcomes regarding the fate of this population. The correct option, D, states that there are so many possible outcomes that it is impossible to predict which one will occur. This set of options is problematic as a student who understands and accepts macroevolution might not choose D. They may understand that genetic mutation is random, but educated predictions about a population could be made based on a defined environmental pressure (which is specified in the stem). Conversely, option D is similar to an 'all of the above/none of the above' response, a popular default option chosen by students who do not know the correct answer, thereby decreasing the difficulty of the item and its contribution to test information (Haladyna et al., 2002). We recommend that researchers using the MUM consider revising this item to include a single best option and developing distracters which reflect specific misconceptions.

Item V6

The origins of the transformation from land animal to sea creature may be observed among some wild sheep that have lived on the coast for hundreds of years. These sheep like to eat seaweed and kelp so much that they are often observed swimming into the water to eat it. If we returned millions of years later to observe these animals, what might you see?

A. Sheep who wanted to be better swimmers and so developed the ability to swim great distances to eat kelp
B. Two distinct but related sheep-like organisms, one that lives in the water and eats kelp, the other lives on land and eats plants
C. These sheep will become extinct because they will not be able to find other food and only their fossils will remain
*D. There are so many possible outcomes that there is really no way to predict what will be seen

Note. *Correct answer

Item V6 was originally intended to target knowledge of speciation, the idea that new species arise from a combination of mutations, natural selection, chance and reproductive, geographic, and/or behavioral isolation (Futuyma, 2005). Option A targets Lamarckian misconceptions about natural selection (i.e. a species can direct its own evolution based on its needs). Options B and C also assume a degree of predictability to the scenario, leaving option D. We would prefer to see only one of the three options proposing a predictable (non-chance) outcome. The second incorrect option could target a different misconception about speciation, such as the role of mutations: 'Mutations would be more likely in the swimming sheep population, eventually causing them to become a new species.' The correct option should be true to the necessary conditions for speciation without assuming a given outcome: 'It is possible that two different sheep species will result due to the isolation of one sheep group from the other.'

Stability of the MUM

From the person-item maps (Figure 1), we witness that the distribution of person measures sits higher on the Rasch scale on the post-test than the pre-test. This is an expected response to a 15-week class covering topics on macroevolution. Along the same token, it may be expected that items may either show invariance or become easier for students by the end of the intervention. However, the significant increase in the difficulty of items V3, V6, V9, V19, and V21 on the post-test is more difficult to explain through the effect of the intervention, especially since improvement of knowledge of macroevolution is evident between the pre- and post-tests. It is interesting that two of these items, V3 and V6, were previously identified as potentially ambiguous items due to their high difficulties and outfit indices. Since ambiguous wording tends to work against higher achieving students who read the question carefully and consider all of the options, it makes sense that these items may become more difficult for a group with more knowledge about macroevolution. The increase in difficulty of items V9, V19, and V21 is more difficult to explain since mean squares fit indices indicate satisfactory fit with the Rasch model. Items V19 and V21 have above-average difficulty and experienced slightly poorer fit with the Rasch model on the post-test, which possibly indicates the presence of wording which may have confused some of the higher ability students. However, these small effects should be interpreted with caution; additional study may be needed before informed revision of items V19 and V21 can be undertaken.

Item V9

Most vertebrate fossils are the bones of these ancient organisms, and it is unlikely that we will find fossils of their eyes. This is because

A. Animals close their eyes when they die, and the eyes are buried under layers of fossils
B. Primitive eyes are so small that they are easily overlooked as fossils
C. Primitive eyes are so different that scientists are not looking for the right structures
*D. Eye tissue typically decays before it can form fossils

Note. *Correct answer

V9 is a comparatively easy item that displays better than expected fit with the Rasch model, inviting the possibility that it is worded in a way that either tips all but the lowest ability students off to the correct response or systematically draws the less literate students toward a particular incorrect response. It is interesting that options B and C refer to the possibility that fossilized eyes exist, but that scientists are unable to detect them. As a result, item V9 measures lack of trust in scientists and the scientific enterprise in addition to knowledge of the process of fossilization. Previous data show that knowledge and acceptance of evolution (Nadelson & Southerland, 2010) are positively correlated. Furthermore, it is reasonable to expect that students with lower acceptance of evolution will tend to have lower levels of trust in scientific explanations, especially those drawing upon evidence in support of evolution. It is therefore likely that students with the least amount of knowledge of macroevolution will be drawn to these options while students with higher levels of knowledge could eliminate these immediately. Given these arguments, it is possible that the increase in difficulty of item V9 on the post-test is reflective of the fact that certain students with lower literacy were less trusting of evolutionary biologists and their data after the intervention and were therefore more drawn to options B and C on the post-test.

Conclusion

We agree with the assertions of Novick and Catley (2012) regarding the semantic issues on particular MUM items, and with their suggestion that the MUM should assess relative and not absolute geologic time. Despite these issues, our data indicate that the MUM is a psychometrically sound instrument. The Rasch analysis indicates that students do not randomly guess answers without considering the question stem. Rather, students are likely answering based on their prior knowledge of key information in the distracters.

Additional revisions should focus on reducing the number of distracters for each item, eliminating those deemed implausible or those which elicit the same misconception as other distracters. We also suggest that some of the stems (e.g. V8) be revised to include key information in the stem and not the distracters, which will reduce test taking time and may improve reliability.

While the MUM provides a unidimensional, valid, and stable scale for measuring students' understanding of macroevolution, we feel that consideration and implementation of some of the suggestions put forward by Novick and Catley (2012) and this study will only improve the efficacy of the MUM as an important tool for measuring students' understandings of macroevolutionary concepts.

References

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy: A project 2061 report*. New York, NY: Oxford University Press.

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC: Author.

Anderson, D. L., Fisher, K. M., & Norman, G. L. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, *39*, 952–978. doi:10.1002/tea.10053

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*(1), 1–7.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

Barton, N. H., Briggs, D. E. G., Eisen, J. A., Goldsten, D. B., & Nipam, H. P. (2007). *Evolution*. New York, NY: Cold Spring Harbor Laboratory Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–545). Reading, MA: Addison-Wesley.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence-Erlbaum Associates.

Bybee, R. W. (1997). *Achieving scientific literacy: From purpose to practices*. Portsmouth, NH: Heinemann.

Campbell, N. A., & Reece, J. B. (2005). *Biology* (7th ed.). San Francisco, CA: Pearson Benjamin Cummings.

Catley, K. M. (2006). Darwin's missing link—a novel paradigm for evolution education. *Science Education*, *90*, 767–783. doi:10.1002/sce.20152

Catley, K. M., Lehrer, R., & Reiser, B. (2005). *Tracing a prospective learning progression for developing understanding of evolution*. Paper commissioned by the National Academies Committee on Test Design for K-12 Science Achievement. Retrieved from www7.nationalacademies.org/bota/Evolution.pdf

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, *35*, 125–129.

Dodick, J., & Orion, N. (2003). Cognitive factors affecting student understanding of geologic time. *Journal of Research in Science Teaching*, *40*, 415–442.

Freeman, S., & Herron, J. C. (2007). *Evolutionary analysis* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

Futuyma, D. J. (2005). *Evolution*. Sunderland, MA: Sinauer.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*, 999–1010.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 309–333.

Kagan, D. M. (1992). Implications of research on teacher belief. *Educational Psychologist*, *27*, 65–90.

Katz, S., Lautenschlager, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science*, *1*, 122–127.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, *2*, 266–283.

Linacre, J. M. (2010). Winsteps (Version 3.70.0). Retrieved from www.winsteps.com

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lovely, E. C., & Kondrick, L. C. (2008). Teaching evolution: Challenging religious preconceptions. *Integrative and Comparative Biology*, *48*, 164–174. doi:10.1093/icb/icn026

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*, 15–29.

Murck, B. W., Skinner, B. J., & Mackenzie, D. (2008). *Visualizing geology* (p. 524). New York, NY: Wiley.

Nadelson, L. S., & Southerland, S. A. (2010). Development and evaluation for a measuring understanding of macroevolutionary concepts: Introducing the MUM. *Journal of Experimental Education*, *78*, 151–190. doi:10.1080/00220970903292983

Naegle, E. (2009). *Patterns of thinking about phylogenetic trees: A study of student learning and the potential of tree-thinking to improve comprehension of biological concepts* (Ph.D. Dissertation). Idaho State University, Pocatello, ID.

National Association of Biology Teachers. (2010). *Mission statement*. Retrieved from http://www.nabt.org/websites/institution/index.php?p=1

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

Nehm, R. H., & Schonfeld, I. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*, *45*, 1131–1160. doi:10.1002/tea.20251

Newport, F. (2009). *On Darwin's birthday, only 4 in 10 believe in evolution*. Retrieved from http://www.gallup.com/poll/114544/Darwin-Birthday-Believe-Evolution.aspx

Novick, L. R., & Catley, K. M. (2012). Assessing students' understanding of macroevolution: Concerns regarding the validity of the MUM. *International Journal of Science Education*, *34*, 2679–2703. doi:10.1080/09500693.2012.727496

Raiche, G. (2005). Critical eigenvalue sizes in standardized residual principle components analysis. *Rasch Measurement Transactions*, *19*, 1012.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13.

Tennant, A., & Connaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care and Research*, *57*, 1358–1362.

Walter, E. M. (2013). *The influence of pedagogical content knowledge (PCK) for teaching macroevolution on student outcomes in a general education biology course* (Doctoral dissertation). ProQuest Publication Number 3576041.

Wright, B. D. (1992). IRT in the 1990s: Which models work best? 3PL or Rasch? *Rasch Measurement Transactions*, *6*(1), 196–200.

Wright, B. D., & Linacre, J. M. (1996). Reasonable mean-square fit values, Part 2. In J. M. Linacre (Ed.), *Rasch measurement transactions* (p. 370). Chicago, IL: Mesa Press.

Wright, B. D., & Stone, M. A. (1979). *Best test design*. Chicago, IL: Mesa Press.