International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

Development and Application of a Novel Rasch-based Methodology for Evaluating Multi-Tiered Assessment Instruments: Validation and utilization of an undergraduate diagnostic test of the water cycle

William L. Romine, Dane L. Schaffer & Lloyd Barrow

To cite this article: William L. Romine, Dane L. Schaffer & Lloyd Barrow (2015) Development and Application of a Novel Rasch-based Methodology for Evaluating Multi-Tiered Assessment Instruments: Validation and utilization of an undergraduate diagnostic test of the water cycle, International Journal of Science Education, 37:16, 2740-2768, DOI: 10.1080/09500693.2015.1105398

To link to this article: http://dx.doi.org/10.1080/09500693.2015.1105398

International Journal of Science Education, 2015
Vol. 37, No. 16, 2740–2768, http://dx.doi.org/10.1080/09500693.2015.1105398

Development and Application of a Novel Rasch-based Methodology for Evaluating Multi-Tiered Assessment Instruments: Validation and utilization of an undergraduate diagnostic test of the water cycle

William L. Romine[a*], Dane L. Schaffer[b] and Lloyd Barrow[c]
[a]*Department of Biological Sciences, Wright State University, Dayton, OH, USA;* [b]*Division of Science, Minot State University, Minot, ND, USA;* [c]*University of Missouri Science Education Center, University of Missouri, Columbia, MO, USA*

We describe the development and validation of a three-tiered diagnostic test of the water cycle (DTWC) and use it to evaluate the impact of prior learning experiences on undergraduates' misconceptions. While most approaches to instrument validation take a positivist perspective using singular criteria such as reliability and fit with a measurement model, we extend this to a multi-tiered approach which supports multiple interpretations. Using a sample of 130 undergraduate students from two colleges, we utilize the Rasch model to place students and items along traditional one-, two-, and three-tiered scales as well as a misconceptions scale. In the three-tiered and misconceptions scales, high confidence was indicative of mastery. In the latter scale, a 'misconception' was defined as mastery of an incorrect concept. We found that integrating confidence into mastery did little to change item functioning; however, three-tiered usage resulted in higher reliability and lower student ability estimates than two-tiered usage. The misconceptions scale showed high efficacy in predicting items on which particular students were likely to express misconceptions, and revealed several tenacious misconceptions that all students were likely to express regardless of ability. Previous coursework on the water cycle did little to change the prevalence of undergraduates' misconceptions.

Keywords: *Multi-tiered diagnostic assessment; Water cycle; Misconceptions; Rasch model*

*Corresponding author. Department of Biological Sciences, Wright State University, Dayton, OH, USA. Email: romine.william@gmail.com

Given the importance of understanding water, it is no surprise that from *Science for All Americans* (American Association for the Advancement of Science [AAAS], 1990) to the present-day *Next Generation Science Standards: For States, By States* (*NGSS*) (NGSS Lead States, 2013), science policy developers have specified water as an integral part of science literacy for K-12 students and our citizens. Water is a necessary ingredient for life. With the global human population approaching eight billion in the next decade, the ability to obtain fresh water will be a high priority for all nations. Despite its importance, most individuals have an inadequate understanding about water (Henriques, 2000; Phillips, 1991). Related topics such as weather and climate cannot be adequately explained without a basic scientific understanding of the cycling of water in and out of the atmosphere and its ability to transmit heat from the equator to the Polar Regions. Moreover, policies involving global climate change and its effect on Earth's water resources cannot be fully comprehended without a fundamental understanding of the water cycle.

Why do we need a multi-tiered diagnostic assessment (MTDA) on the water cycle? There are a limited number of scientific studies which have examined students' misunderstandings of the water cycle and its components. These studies have focused upon K-12 students (e.g. Covitt, Gunckel, & Anderson, 2009; Gunckel, Covitt, Salinas, & Anderson, 2012). However, misconceptions about water are of particular concern for undergraduate students who are expected to become responsible citizens and those who are going to become teachers. It is unrealistic for teachers to address students' misconceptions about water if they hold these misconceptions themselves.

To the end of addressing student's misconceptions, we need assessments and models for analysis that can uncover students' scientific misconceptions accurately and efficiently. This study presents a comprehensive and valid three-tiered instrument for measuring undergraduates' knowledge and misconceptions about the water cycle. We will demonstrate the efficacy of the confidence tier in increasing measurement validity and reliability, and utilize the third tier with respect to a novel framework to illuminate specific misconceptions undergraduates possess about the water cycle on a hierarchical scale. We will show that using Rasch ability and misconception domains together provides a holistic picture of the scientific misconceptions students are likely to hold, and how these may change with students' underlying knowledge and prior experience.

Purpose of the Research

We present the development and validation of a three-tiered diagnostic test of the water cycle (DTWC). While much preliminary work has been completed to establish content validity of items on the DTWC and to make a strong case for construct validity based on classical test theory (CTT) (Schaffer, 2013), a holistic effort to establish construct validity of these items based on the Rasch measurement framework has not been undertaken. Furthermore, there is currently no model-based framework for exploring the prevalence of students' misconceptions about the water cycle. We therefore proceed to validate the DTWC in the context of a Rasch measurement framework,

and then use this framework to formulate a general probabilistic model allowing prediction of undergraduate students' misconceptions related to hydrologic processes on the individual level. Through this process of development, validation, and utilization of the DTWC, we addressed four questions:

(1) What is the construct validity and reliability of the DTWC, and how does adding the reason and confidence tiers affect construct validity and reliability?
(2) How effective is the DTWC in revealing misconceptions about the water cycle in undergraduate students?
(3) What are the most persistent misconceptions related to water cycle processes that undergraduate students possess?
(4) What is the relationship between students' prior experience with the water cycle and their tendency to display misconceptions about the water cycle?

Review of Literature

Framework for MTDA Design

The theoretical framework used by Schaffer (2013) to develop and validate the DTWC was recommended by Treagust (1986, 1988, 1995) and based originally upon a constructivist learning theory (David Treagust, personal communication, 5 September 2012). Many researchers have advocated the use of this particular strategy to develop and validate MTDAs (e.g. Lin, 2004; Odom, 1992; Pesman & Eryilmaz, 2010; Tsai & Chou, 2002). Treagust (1995) states that there are three main stages in developing a MTDA:

(1) defining the content,
(2) obtaining information about students' conceptions, and
(3) developing a diagnostic instrument (p. 330).

In the first stage, defining the content, a researcher needs to identify the propositional knowledge statements (PKSs). These PKSs establish the knowledge an individual needs to have for a complete conceptual understanding or for obtaining scientific literacy about a topic (e.g. transpiration is a living process by which plants give water vapor back to the atmosphere). For the DTWC, 47 PKSs were derived from the several college textbooks (Aguado & Burt, 2004; Arhens, 2009; Lutgens & Tarbuck, 2010), and from the American Meteorological Society's *The Global Water Cycle* (2001). After the validation of the PKSs by three experts in the fields of meteorology and hydrology, a concept map was constructed to connect all the PKSs and to make sure that all the fundamental water cycle processes were fully integrated into the development of the DTWC. The concept map was used as an organizational tool to represent the relationship between the PKSs. To ensure the content validity of PKSs and the concept map, both the PKSs and the concept map were sent to five specialists in the fields of science education, meteorology, and hydrology.

Stage two of developing the DTWC was about obtaining information on students' scientific misconceptions about the water cycle. Several educational databases (e.g. Academic Search Premier, ERIC, Education Full Text, and Google Scholar) were used to retrieve articles for this literature review, as well as Duit's (2009) literature review on students and teachers' conceptions in science. Several studies related to the water cycle were identified and reviewed thoroughly (e.g. Bar, 1989; Bar & Galili, 1994; Bar & Travis, 1991; Ben-zvi-Assarf & Orion, 2005; Cardak, 2009; Russell, Harlen, & Watt, 1989; Shepardson, Wee, Priddy, Schelleberger, & Harbor, 2009; Taiwo, 1999). Documented misconceptions of the water cycle were noted for use as distracters in the construction of the DTWC pilot test.

During the review of literature for the DTWC, unstructured interviews were conducted with undergraduates who had recently completed instruction concerning their knowledge on the water cycle. Prior to that instruction, the undergraduates were first asked to create their own representation of the water cycle. After examining the students' drawings, the researcher interviewed six undergraduates about their drawings and their understanding of the water cycle. During the interviews, probing questions were asked to encourage the undergraduates to elaborate upon their answers. Field notes of the students' answers were taken and their drawings were collected as artifacts. After instruction, the undergraduates were asked to complete another drawing of the water cycle and to write a reflection on how their drawings changed and why. To finish the second step, a pilot two-tier test was administered to the undergraduates that included 34 multiple choice questions and an open-ended question for the second tier so that the undergraduates could write a reason as to why they selected a particular answer on the content tier. This particular pilot was completed by 51 students. The information provided by this pilot test was essential because it helped establish distracters needed in the development of response selections for both the first and second tiers in the next piloted version of the DTWC. Refinement of the original pilot test continued with a second piloting of the DTWC that included 20 in-service secondary science teachers and community college geology instructors during the summer of 2012. The second piloting of the DTWC had multiple choice selections for both the first and second tiers. The in-service teachers also wrote comments that aided in improving both the content validity and the communication validity of the DTWC. From this piloted test, two questions were eliminated due to ambiguity, reducing the number of questions from 34 to 32. The third and final stage started with designing a specification grid which allowed the DTWC developer to align the PKSs, concept map, and the developed assessment questions in order to make sure that all three essential content portions of the DTWC still had content validity.

Misconceptions

Misconceptions, also referred to as alternative conceptions, are not easily abandoned by individuals (Wandersee, Mintzes, & Novak, 1994). Students' prior experiences influence their understanding of science. Driver and Erickson (1983) concluded that teachers need to acknowledge that their students come into the classroom with

both scientifically correct and incorrect ideas about how the natural world works. Scientific concepts can help identify and relate factors that students use to explain scientific phenomena. This may include many common conceptions as well as misconceptions. It is not that students are not learning or they lack knowledge, but have derived non-scientific interpretations (Munson, 1994). Misconceptions can be defined as inaccurate explanations of science phenomena constructed by students (D'Avanzo, 2003; Stamp & Armstrong, 2005). Misconceptions can be very persistent and may interfere with the learning of new science concepts. However, misconceptions may not be completely exclusive and may even coexist with an accurate understanding of a scientific conception (Smith, di Sessa, & Roschelle, 1993). Sources of misconceptions are varied and complex, and thus can be difficult to resolve (Modell, Michael, & Wenderoth, 2005). Misconceptions can be present both before and after instruction in students across many subject areas and may appear even after the misconceptions are explicitly addressed (Smith et al., 1993). The diagnosis of these misconceptions is essential for the development of more effective teaching strategies and interventions that will help reduce the barriers to reaching scientific literacy. Ausubel (1968) stated that educators should 'find out what the learner already knows and teach him accordingly' (p. 337).

Water and the Water Cycle

Given the importance placed on understanding of water in K-12 schooling, one would expect that undergraduates would possess an adequate understanding of hydrologic systems and processes. However, Phillips (1991) and Henriques (2000) noted that most individuals have insufficient scientific understanding of water and the cycling of water in and out of the atmosphere. Phillips (1991) examined 4–9th graders over 10 years and identified over 50 Earth science misconceptions, 14 of which were related to water. In addition, Phillips' research documented that many of these continued well into adulthood. Henriques (2000) conducted a review focusing on children's misconceptions of weather. She listed misconceptions identified by researchers, and then listed possible reasons why students may have them. Besides the water cycle, her literature review included phase changes of water, the atmosphere and its gases, seasons and heating of the Earth, global warming, and the greenhouse effect. In a similar study, Brody (1993) found that misconceptions about water and water resources were similar across students regardless of their science coursework.

There have been a limited number of research studies focusing upon K-12 students' conceptual understanding of the water cycle (e.g. Bar, 1989; Ben-zvi-Assarf & Orion, 2005; Shepardson et al., 2009; Taiwo, 1999). In addition, Cardak (2009), Morrell and Schepia (2009), and Schaffer (2013) conducted studies on undergraduates' understanding of water cycle. Cardak (2009) examined 156 Turkish science education students' drawings of the water cycle, and then conducted selected interviews. The main misconception found in the students' drawings was that they only included the processes of evaporation and condensation.

Morrell and Schepia (2009) examined 78 elementary pre-service teachers' (PSTs) pre-test and post-test representations/drawings of the water cycle. Following the use of Project WET's (Project WET and Council for Environmental Education, 1995) activity, *The Incredible Journey*, as a classroom intervention, researchers noted limited improvement in the PSTs' water cycle drawings, but still found previously reported scientific misconceptions when examining elementary and middle school students' conceptual understanding of the water cycle. No specific misconceptions were identified, but generalizations were made concerning the PSTs' lack of knowledge of the water cycle.

Schaffer (2013) developed and validated a three-tiered diagnostic test examining elementary education and middle/secondary science education PSTs' scientific understanding and confidence around the water cycle using two-tiered items accompanied by a certainty of response index (CRI) for the third tier (Hasan, Bagayoko, & Kelley, 1999). This 15-item assessment was administered to 77 PSTs. Schaffer found that the PSTs lacked accurate scientific understandings of the water cycle and documented 49 potential alternative conceptions using the '10% and higher' rule (e.g. Chandrasegaran, Treagust, & Mocerino, 2007; Odom & Barrow, 1995; Wang, 2004). Using the '10% and higher' rule, a distracter selected by 10% or more of the students is considered a misconception. However, with the inclusion of the CRI as the third tier, the study reduced that amount and concluded that the PSTs had six genuine scientific misconceptions of the water cycle (Schaffer & Barrow, 2015): (1) water vapor is not a greenhouse gas; (2) carbon dioxide is the principal greenhouse gas instead of water vapor; (3) global sea levels will rise when sea ice melts; (4) the extra water produced due to the melting of sea ice will cause the sea level to rise and flood coastal areas; (5) when a beverage warms and causes water to condense inside the can, the extra water causes too much volume in the can and seeps out; and (6) cooler air provides a lower amount of energy for evaporation (as opposed to raising relative humidity which slows evaporation).

Uncovering Misconceptions through MTDAs

Methodologies for diagnosing misconceptions vary considerably. Open-ended approaches include concept mapping (Novak, 2002, 1995; Ruiz-Primo, 2000) and interviews (Bell, 1995; Carr, 1996; Posner & Gertzog, 1982). While open-ended tasks hold utility in diagnosing misconceptions and conceptual change across many topic areas in science education, these techniques do not lend themselves to large numbers of students. Carefully constructed multiple choice tests are therefore often utilized to diagnose misconceptions in large samples (Treagust, 1995). Many MTDAs and concept inventories (e.g. Chandrasegaran et al., 2007; Odom & Barrow, 1995; Wang, 2004) utilize an incorrect response rate of 10.0% and greater in determining potential student misconceptions. Romine, Barrow, and Folk (2013) describe a statistical method for diagnosing misconceptions on a traditional single-tiered multiple choice test which utilizes a group chi-square test of the null hypothesis that the proportion of students choosing an incorrect response is no greater than chance. While this approach

lends statistical objectivity, the null assumption that all responses are equally likely to be chosen may be overly simplistic. It is further limited in that, while a group inference can be made, the model cannot be used at the individual level to predict whether or not a particular student will hold a particular misconception. These limitations illustrate the advantage of adding additional tiers to multiple choice questions.

Most MTDAs are two-tier, but three-tier MTDAs (e.g. Arslan, Cigdemoglu, & Moseley, 2012; Caleon & Subramaniam, 2010a; Cetin-Dindar & Geban, 2011; Odom & Barrow, 2007; Pesman & Eryilmaz, 2010) also exist as does one four-tier (Caleon & Subramaniam, 2010b) MTDA. Three-tier and four-tier assessments involve the use of a confidence tier, which allows students to rate their strength of confidence in their selected answers to the first- and second-tier levels. While CRIs are not new to the social sciences, their utilization in MTDAs is relatively new in science education, beginning with Odom and Barrow's (2007) revision of the Diffusion and Osmosis MTDA (DODT) (Odom, 1992; Odom & Barrow, 1995).

Caleon and Subramaniam (2010a) developed and validated a three-tiered diagnostic test on the nature and propagation of waves (WADI). The instrument was given to 243 high school students as a pre- and post-test. The pre-test Cronbach's alpha reliabilities were 0.58, 0.64, and 0.88 for the content, reason, and confidence tiers, respectively. The post-test reliabilities were 0.63, 0.69, and 0.93. Cetin-Dindar and Geban (2011) developed and validated a three-tiered assessment of understanding of acids and bases using a sample of 156 high school students. Alpha reliabilities of 0.58, 0.59, and 0.72 were found for the first, second, and third (confidence) tiers, respectively. Arslan et al. (2012) developed and validated a three-tiered MTDA on atmospheric-related environmental problems using 256 PSTs. The reliability of the PSTs' scores was 0.74. Pesman and Eryilmaz (2010) developed and validated a three-tiered MTDA to assess misconceptions of 124 high school students' understanding of simple electrical circuits. The Cronbach's alpha reliability was 0.69.

A four-tiered diagnostic assessment was developed and validated by Caleon and Subramaniam (2010b). The 4WADI is a modification of the WADI, and was used to assess 598 high school students following instruction. Cronbach's alpha was measured at: content tier ($\alpha = 0.40$) with confidence ($\alpha = 0.88$), reason tier ($\alpha = 0.19$) with confidence ($\alpha = 0.91$), and the overall score ($\alpha = 0.50$) and confidence ($\alpha = 0.92$). By using a MTDA, either at the beginning and/or at the completion of study, educators gain a unique perspective on their students' conceptual understandings and misconceptions which they can use to modify instruction.

Methods

Integrating Confidence into Mastery—A Framework

While the idea of MTDAs is not new, relatively few instruments contain a confidence tier with a CRI. Schaffer, Romine, and Barrow (2015) make the case that confidence

needs to be considered as an essential component of mastery of content knowledge. Concept mastery needs to be defined as a product of both correctness and confidence (Schaffer et al., 2015). Adding a confidence tier to the DTWC allows determination of whether a participant's answer is due to correct knowledge, lack of knowledge, or presence of an actual incorrect scientific conception about the water cycle. In the third tier of the DTWC, a four-point Likert scale was used to measure participants' confidence. Participants selected from 'guessing (1),' 'uncertain (2),' 'confident (3),' or 'very confident (4).' A framework for integration of correctness and confidence into mastery as defined in this study is outlined in Table 1.

Utilizing the confidence tier in light of the above construction, we define a 'misconception' as understanding or mastery of a concept that differs from the currently accepted scientific evidence. Students who selected 'certain' or 'very confident' as their responses for the confidence tier of the DTWC were deemed as having high confidence. If this selection occurred with an incorrect response, then this was considered as a 'tenacious misconception' (e.g. Arslan et al., 2012; Caleon & Subramaniam, 2010a; Odom & Barrow, 2007).

Selection of correct responses for both the first and second tiers and a confidence level of 3 or above was considered indicative of proper scientific knowledge of the water cycle concepts. A student's selection of 'guessing' or 'uncertain' on the DTWC was considered low confidence, indicating a lack of knowledge even if the participant had a correct answer.

Validation of a Metric for Understanding of the Water Cycle

Participants. The DTWC was validated using a sample of 130 university students from two colleges. Seventy-seven students were sampled from a large research-extensive university in the midwestern USA comprising over 30,000 students. The remaining 53 students were sampled from a small private liberal arts college from the same region enrolling approximately 1,400 students. Of the 130 participating students, 84 (65%) were female, and 46 (35%) were male. Reported prior experiences with learning the water cycle varied. Fifteen students (12%) reported not seeing material related to the water cycle since elementary school. Forty-five (35%) reported seeing the material in middle school, 39 (30%) reported last learning it in high school, and 29 (22%) reported studying the water cycle in college. About 89%, i.e. 116 of the 130 students, reported that they had never taken an atmospheric science/meteorology class as an undergraduate.

Table 1. CRI matrix for the DTWC

Answer	Low CRI (2.0 or below)	High CRI (above 2.0)
Correct	Lack of knowledge (lucky guess)	Correct conception (mastery of a correct concept)
Wrong	Lack of knowledge	Misconception (mastery of an incorrect concept)

Scoring of the DTWC. A distinct advantage of MTDAs lies in their diversity. Given a teacher's or researcher's interest and practical environmental constraints, a MTDA such as the DTWC can be used independently as a traditional single-tiered, two-tiered, or three-tiered test. We therefore proceed to evaluate the construct validity and reliability of the DTWC with respect to one-, two-, and three-tiered usage styles, and discuss strengths and limitations of respective styles in the context of the Rasch measurement framework. To evaluate the use of the DTWC as a single-tiered test, a correct item response on the first tier was scored a '1' and an incorrect response was scored a '0.' When the DTWC was used as a two-tiered test, a correct response on the first *and* second tiers was required for a '1' score. A student who got the first tier correct but chose an incorrect second-tier response, or chose an incorrect first-tier response with a correct second-tier response, was given a '0' score on the item. When the third [confidence] tier was integrated into the scoring system using the framework presented above (Schaffer et al., 2015), correct first- and second-tier responses and a confidence level above 2 on the four-level CRI was required for a '1' score. Students with correct responses and low confidence (2 or below on the CRI), as well as those with incorrect responses, were given a '0' score.

While nearly all test scores express a student's tendency to express correct knowledge (hence common use of the word, 'ability'), evaluation of misconceptions is an important goal of multi-tiered assessments which can be aided by instead defining the scale in terms of a student's tendency to express incorrect or non-scientific knowledge. In this coding scheme, an incorrect first- or second-tier response with a confidence level above 2 was scored a '1,' meaning that a misconception was expressed. Correct responses or responses with low confidence were scored a '0,' meaning that no misconception was expressed by the student.

The Rasch Framework. We used the Rasch model as a criterion to evaluate the construct validity and reliability of the DTWC scale when it is used as a single-tiered, two-tiered, or three-tiered instrument. The Rasch model is mathematically related to item response theory (IRT), which is itself an extension of CTT. CTT, also called true score theory, models a continuous latent variable like knowledge of the water cycle by treating the item responses as continuous. Given the dichotomous (0 = incorrect; 1 = correct) response set for the DTWC, the assumption that an item response is related linearly to ability is dubious. IRT better addresses the categorical nature of item responses by utilizing the logit (log-odds) transformation such that categorical observations can be rightly treated with a probabilistic framework and used to model the continuous latent variable.

Justification for using the Rasch model. The idea that theories are discovered from data underpins scientific inquiry. Both IRT and CTT emphasize construction of a scale that fits the data (Lord & Novick, 1968). However, this scientific perspective carries the important and inherent assumption that the instrumentation used to collect the data is valid and well calibrated. The goal of this validation study is to

make a case for validity of the DTWC, and to identify areas where validity is lacking or uncertain, through analysis of data. The Rasch measurement framework is well suited to this task by imposing a philosophical criterion for how data produced by a well-constructed, valid test should look. The Rasch model is specified by equating the log-odds of a student answering an item correctly to the difference between the item's difficulty and the student's ability. A high-ability student is likely to get a low-difficulty item correct. On the same token, a low-ability student is likely to miss a high-difficulty item. A natural philosophical extension is that as a student's ability and an item's difficulty get closer together, it becomes more difficult to predict whether or not the student will get the item correct. An argument for construct validity of the DTWC can be made based on satisfactory fit of the data it produced with these criteria which are expressed mathematically through the Rasch model.

An additional advantage of using the Rasch model over the data-dependent traditions of CTT and IRT is that person and item measures can be compared along the same scale. Rasch measures of item difficulty are invariant with respect to student ability, and Rasch measures of student ability are invariant with respect to item difficulty. This resonates with our intuitive sense of quality measurement—the tick marks on a ruler should be invariant regardless of the length or height of the object that the ruler is used to measure (Boone & Scantlebury, 2006).

Item-level construct validity. Using BIGSTEPS software, we fit a dichotomous, or simple logistic, Rasch model (Wright & Stone, 1979), to the data collected through use of the DTWC as a (1) single-tiered test, (2) two-tiered test, and (3) a three-tiered test with a CRI above 2 as a criterion for correctness (Table 1). Mean squares infit and outfit with the Rasch model were used as evidence for or against item-level construct validity. As chi-square statistics, both infit and outfit were calculated as a function of observed and expected responses. However, as opposed to outfit, infit is information-weighted, making it less sensitive to anomalous responses. A fundamental idea behind the Rasch model is that if an item on the DTWC is an effective measure of knowledge of the water cycle, then the probability of a student getting the item correct should increase monotonically with his/her ability. Mean squares fit statistics enabled us to evaluate the conformity of data to this criterion.

Mean squares fit indices have an expected value of 1.0. While values between 0.5 and 1.5 are generally considered productive for measurement (Wright & Linacre, 1994), a stricter criterion of 0.7–1.3 is often used for low-stakes tests (Bond & Fox, 2007). A mean squares fit index above 1.0 indicates greater-than-expected error with respect to the Rasch model. Extreme values (greater than 1.5) are indicative that high-ability students tend to miss the item or that low-ability students are able to guess the correct answer.

A mean squares fit index below 0.5 indicates that data fit the Rasch model suspiciously well. If a student and an item are at the same location along the Rasch scale, then the probability of the student answering that item correctly is 50%. When a student's ability is close to the item's difficulty, the Rasch model is expected to

miscategorize responses with some frequency. A mean squares fit index of 0 would suggest perfect discrimination, or a clean threshold where students below would miss the item and those above would get it correct, or what is called a 'guttman pattern' in the Rasch literature (Wright & Stone, 1979; Linacre, 2002). Closer-than-expected fit with the Rasch model indicates presence of item wording which favors high-ability students or discriminates against low-ability students (Masters, 1988).

Assessment of unidimensionality. Besides item-level validity, we are also interested in the structure and validity of scales constructed from different usages of the DTWC. This includes measurement precision, conformity to the fundamental assumptions of unidimensionality and local independence, and appropriate targeting of items to undergraduate students. We used Rasch person and item reliability (Linacre, 1999) to quantify precision of person and item measures, respectively, along the Rasch continuum. Conformity to the assumption that the DTWC measures a single dimension (knowledge of the water cycle) was tested through principal components analysis (PCA) on residuals with respect to the Rasch model. The idea behind PCA on residuals is that, if all important variances in the data are modeled by a single Rasch model, then residuals should reveal no detectable pattern. Simulation studies by Raiche (2005) and Linacre and Tennant (2009) suggest that a first eigenvalue around 2 items of variance is indicative of randomness in the residuals. However, Galli, Chiesi, and Primi (2008) use a first eigenvalue of 3 as a threshold to indicate that a scale is unidimensional enough to be useful.

Assessment of local independence. Conformity of items to the assumption of local independence was quantified through observation of correlation of item residuals. Local independence implies that there are no commonalities between items that are not accounted for by knowledge of the water cycle (Linacre, 2009). After knowledge of the water cycle is taken into account, item residuals will be non-correlated if they meet the assumption of local independence. While it is seldom the case that these extraneous dependencies between items are completely nonexistent, item residual correlations below 0.7 (indicating less than 50% shared variance) are considered to indicate reasonable conformity with the local independence assumption (Linacre, 2010).

Person–item mapping. In addition to quantifying reliability and conformity to the assumptions of latent variable modeling, we also wished to quantify the extent to which items were well targeted to participants. An item will provide the most information about participants with locations proximal to its location on the Rasch scale. By plotting person and item measures concurrently along the Rasch continuum, we are able to make a visual inference on the extent to which a particular usage for the DTWC makes the test too easy or too difficult for participants, and observe how using the DTWC in one-, two-, and three-tiered fashions changes item and test

functioning. Due to the invariance property of Rasch measures, person–item mapping is also informative in making predictions about understandings and misunderstandings a particular student is likely to have. If a student's measure is located above an item's measure, then that student is likely to get that item correct. On the other hand, students with measures below an item's measure are likely to miss that item. In this way, person–item mapping serves to give students' logit measures qualitative meaning.

Validation of a Metric for Water Cycle Misconceptions. The utility of Rasch as a validation tool is derived from its utility as a predictive model. Using traditional coding (0 = incorrect; 1 = correct), a dichotomous Rasch model enables prediction of whether or not a participant of a given ability will get an item with a defined difficulty correct. Specifically, the Rasch model would predict that a student will get all items with measures at or below his/her ability level correct, but will miss items with measures above his/her ability level. Therefore, the Rasch model can also predict whether or not a student will display a misconception. If a misconception-focused coding scheme is used (1 = misconception; 0 = no misconception), where a misconception is an incorrect answer coupled with a CRI rating above 2 (Table 1), instead of a measure of ability, the Rasch scale provides a measure of a student's tendency to express misconceptions. Under such a transformation, a student higher on the Rasch scale is likely to have a greater number of misconceptions, and items lower on the Rasch scale have a greater tendency to reveal misconceptions. Specifically, if a student's position on the Rasch continuum sits above that of an item, then that student is predicted to express a misconception on that item. While this type of scale does not get at the particular misconception displayed by a student on an item, a simple distracter analysis is sufficient to qualify the incorrect responses that were chosen with the greatest frequency on a particular item.

A key purpose of a diagnostic test is to reveal misconceptions. The Rasch paradigm yields a unique and helpful perspective on the suitability of items for accomplishing this task. Mean squares infit and outfit indices can be used to evaluate the extent to which the distance between an item's and student's locations on the misconception continuum is a predictor of whether or not he/she will express a misconception on that particular item. As with previous coding schemes, PCA on Rasch residuals gives a measure of the extent to which the misconceptions scale is unidimensional. Person and item reliability indices were used to indicate the precision of person and item locations along the misconception scale. These all provide measures of the extent to which the DTWC does what a diagnostic test needs to do, which is to provide a diagnostic for students' misconceptions about the water cycle.

Impact of Prior Instruction. Differential effects of prior instruction on prevalence of misconceptions related to the water cycle were evaluated using a factorial three-way analysis of variance (ANOVA) procedure. Categorical indicators of interest included when students last studied the water cycle, whether or not students had taken an

undergraduate meteorology/atmospheric science class, the college from which students were sampled, and all interactions between these variables. Hence this model contained three main effects, three two-way interactions, and one three-way interaction.

Results

What is the Construct Validity and Reliability of the DTWC across the Three-Tiers?

Item-Level Validity. A majority of items on the DTWC fit well with the Rasch model (Table 2), which is indicative of the utility of items to differentiate students along the scale of understanding of the water cycle. All items under all usages had infit values between 0.7 and 1.3. The lowest infit values were measured at 0.89, 0.81, and 0.88, and the highest values were measured at 1.13, 1.18, and 1.27 for one-, two-, and three-tiered usages, respectively. Outfit values for several items fell outside of this range. Item V28 (density of humid air) displayed outfits of 1.47 and 1.86 for one- and two-tiered usages, respectively. This item fit better when confidence was integrated into mastery; the outfit of this item when used as a three-tier was 0.84. Item V24 (rising air expands and cools), when used as a two-tiered item, had an outfit of 1.62. However, this item fit well with the Rasch model when used as a one- and three-tiered item, with outfit values of 1.06 and 1.00, respectively. Items V3 (energy), V13 (condensation), and V29 (shape of a raindrop) displayed satisfactory outfits when used in a one- and two-tiered fashion, but were measured at 1.49, 1.74, and 1.91, respectively, when three-tiered scoring was used. Item V18 (deposition) displayed an outfit of 1.16 when used as a single-tiered item. However, response patterns associated with two- and three-tiered measurement greatly underfit the Rasch model, with outfit values of 5.66 and 5.39, respectively. Items V19 (greenhouse gas) and V21 (latent heat) overfit the Rasch model when used in the multi-tiered format. While item V19 had an outfit of 0.76 when used as a single-tiered item, use as a two-tiered item caused the outfit to drop to 0.64. When the confidence tier was integrated into scoring, outfit further dropped to 0.55. A similar phenomenon was observed for item V21. When used as a single-tiered item, it displayed an outfit of 0.81. This dropped to 0.76 when used as a two-tiered item and further to 0.49 when the confidence tier was integrated.

Scale Validity. The DTWC scales were found to meet criteria of unidimensionality, local independence, and reliability to an extent that they are useful in certain contexts. However, we find that the utility of the DTWC varies depending upon how it is used. Use of the DTWC as a single-tiered assessment provides a Rasch person measurement reliability of 0.55. When the second tier is integrated, person measurement reliability increases to 0.64. Reliability further increases to 0.73 when the third tier is utilized to integrate confidence into mastery. All usages lead to sufficiently unidimensional scales with locally independent items. The largest item residual correlations were measured at 0.57 (V24 and V25), 0.33 (V15 and V21), and 0.40 (V25 and V26) for one-, two-,

Table 2. Rasch item measures and mean squares fit indices for multi-tiered and misconception scale

	Item difficulty measures (SE)				MNSQ infit measures				MNSQ outfit measures			
Item	1-tier	2-tier	3-tier	Misc.	1-tier	2-tier	3-tier	Misc.	1-tier	2-tier	3-tier	Misc.
V1	0.46(0.18)	0.19(0.20)	0.21(0.27)	−0.45(0.20)	1.01	0.99	0.99	1.02	1.04	1.13	1.13	0.98
V2	−0.87(0.19)	−0.87(0.18)	−1.01(0.21)	−0.17(0.20)	0.91	0.91	0.92	0.90	0.86	0.89	0.94	0.89
V3	−1.3(0.22)	−1.94(0.20)	−1.93(0.22)	1.01(0.27)	1.04	1.01	1.27	0.96	1.24	1.08	1.49	0.91
V4	−1.92(0.26)	0.41(0.21)	0.29(0.28)	−1.11(0.19)	0.97	0.93	1.00	1.02	1.02	0.78	0.81	1.11
V5	−1.79(0.25)	−0.13(0.19)	−0.53(0.22)	−0.78(0.19)	1.01	1.03	1.00	1.05	0.94	1.03	0.94	1.03
V6	−1.92(0.26)	−2.78(0.25)	−2.24(0.19)	1.55(0.33)	0.92	0.90	0.96	0.93	0.85	0.79	0.95	0.73
V7	−1.73(0.24)	−2.72(0.24)	−2.99(0.21)	1.67(0.37)	0.97	1.00	0.90	1.14	0.89	0.93	0.88	1.50
V8	0.13(0.18)	−0.80(0.18)	−0.63(0.22)	0.44(0.24)	0.98	1.01	1.02	1.07	0.95	0.97	0.96	1.02
V9	0.39(0.18)	−0.25(0.19)	−0.37(0.23)	−0.17(0.20)	0.99	0.91	0.93	0.94	0.97	1.04	0.74	0.96
V10	−1.30(0.22)	−1.86(0.20)	−1.85(0.20)	0.55(0.24)	1.05	1.06	1.04	1.03	1.16	1.13	1.09	1.07
V11	−0.29(0.19)	−0.87(0.19)	−0.53(0.22)	1.25(0.29)	1.13	1.08	0.98	1.00	1.14	1.06	0.98	0.97
V12	−2.15(0.28)	−2.91(0.26)	−3.08(0.22)	2.30(0.46)	0.95	0.93	1.03	1.05	0.82	0.99	0.92	2.45
V13	0.46(0.19)	1.62(0.32)	1.32(0.42)	−0.13(0.20)	1.07	0.98	0.95	0.99	1.11	0.87	1.74	0.94
V14	−0.57(0.19)	−1.34(0.18)	−2.01(0.19)	0.14(0.21)	0.98	0.81	0.92	1.01	1.00	0.77	0.85	0.98
V15	0.84(0.20)	2.00(0.38)	1.75(0.50)	1.01(0.28)	1.08	0.99	1.00	1.05	1.05	0.97	0.97	1.14
V16	0.67(0.18)	0.14(0.20)	−0.07(0.25)	0.55(0.24)	1.01	0.97	1.00	1.06	1.16	1.02	1.14	1.00
V17	0.88(0.19)	0.51(0.22)	0.76(0.33)	−0.71(0.19)	0.99	0.95	0.93	0.96	0.99	0.81	0.73	0.94
V18	1.15(0.20)	1.86(0.36)	1.75(0.50)	−0.64(0.19)	0.99	1.02	1.01	0.91	1.16	5.66	5.39	0.84
V19	2.03(0.26)	1.52(0.31)	1.32(0.42)	−0.78(0.19)	0.92	0.93	0.95	1.01	0.76	0.64	0.55	0.99
V20	2.03(0.26)	1.10(0.27)	1.01(0.37)	−2.14(0.21)	1.05	1.07	1.05	1.09	1.00	1.05	0.98	1.21
V21	2.86(0.36)	2.00(0.38)	1.32(0.42)	0.33(0.22)	0.96	0.98	0.92	0.93	0.81	0.77	0.49	0.86
V22	−1.61(0.23)	−0.21(0.19)	−0.07(0.26)	−0.25(0.20)	0.89	1.03	1.06	1.05	0.84	0.97	0.83	1.12
V23	−0.33(0.19)	−1.14(0.19)	−1.01(0.21)	0.79(0.27)	1.07	1.06	1.08	1.12	1.12	1.08	0.99	1.23
V24	0.53(0.18)	0.27(0.22)	0.29(0.28)	−0.33(0.20)	0.97	1.11	0.97	0.97	1.06	1.62	1.00	0.96
V25	0.92(0.19)	1.17(0.27)	1.32(0.42)	−0.33(0.20)	0.92	1.01	0.98	0.97	0.94	1.23	1.02	0.96
V26	2.03(0.26)	2.87(0.57)	3.14(0.99)	−0.89(0.19)	1.02	1.00	1.02	0.95	0.98	0.88	1.06	0.89
V27	0.23(0.18)	1.33(0.29)	1.75(0.50)	−0.89(0.19)	1.01	0.97	1.01	0.97	1.02	0.96	0.84	1.02
V28	2.25(0.28)	2.35(0.46)	3.14(0.99)	−1.70(0.19)	1.02	1.07	1.02	0.99	1.47	1.86	0.84	1.06
V29	0.99(0.20)	0.66(0.23)	1.32(0.43)	−0.71(0.19)	1.03	1.01	1.09	0.89	1.15	1.18	1.91	0.82
V30	−1.20(0.21)	0.02(0.20)	0.65(0.32)	0.49(0.23)	0.95	0.95	0.99	0.93	0.99	1.02	0.94	0.99
V31	−0.64(0.19)	−1.63(0.19)	−2.20(0.19)	0.44(0.24)	0.93	0.92	0.88	1.09	0.91	0.90	0.87	1.18
V32	−1.25(0.21)	-0.60(0.20)	−0.83(0.22)	−0.33(0.20)	0.94	1.18	1.09	0.99	0.89	1.18	0.95	1.01

and three-tiered usages, respectively. A residual correlation of 0.57 implies that about 33% of the variance between items V24 and V25 can be attributed to a factor not accounted for by knowledge of the water cycle when the test is treated as single-tiered. Such low dependency between items does not significantly deteriorate the functioning of the measurement scale (Linacre, 2010). First eigenvalues for PCA on Rasch residuals were measured at 2.61, 2.24, and 2.33 for one-, two-, and three-tiered usages, respectively. These values are slightly above the strict criteria for unidimensionality outlined in simulation studies (Raiche, 2005; Linacre & Tennant, 2009). However, that these are significantly below 3, coupled with low item interdependence, indicates that the DTWC can be treated as a unidimensional measure (Bond & Fox, 2007).

Item measurement reliabilities for one-, two-, and three-tiered usages were measured at 0.97, 0.97, and 0.94, respectively, indicating that a sample size of 130 yields sufficient precision to establish test construct validity. Figures 1 and 2 display comparative person and item measure distributions for one-, two-, and three-tiered usages, and provide an illustration of how item functioning and person distributions change with how the DTWC is used. The horizontal axis of Figure 1 displays locations of person and item Rasch logit measures along the scale of the one-tiered DTWC, where overlap of the person and item measure distributions can be observed. Person measures derived from one-tiered usage are centered about a mean measure of 0.06, with a standard deviation of 0.64. Overlap of person and item measures on the two-tiered DTWC is displayed along the vertical axis of Figure 1 and the horizontal axis of Figure 2. When the DTWC is used as a two-tiered test, the mean of person measures shifts downward to –1.02 and the standard deviation of measures increases to 0.80. When the DTWC is used as a three-tiered test, person measures shift further downward, centering about a mean of –2.19, with a further increased standard deviation of 1.22. With the downward shift of the person measure distribution, person and item distributions (vertical axis of Figure 2) are further offset.

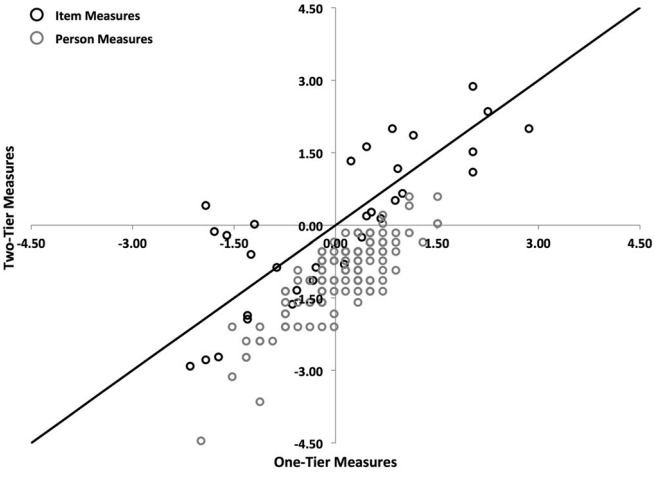

Figure 1. Comparison of one- and two-tiered person and item measures

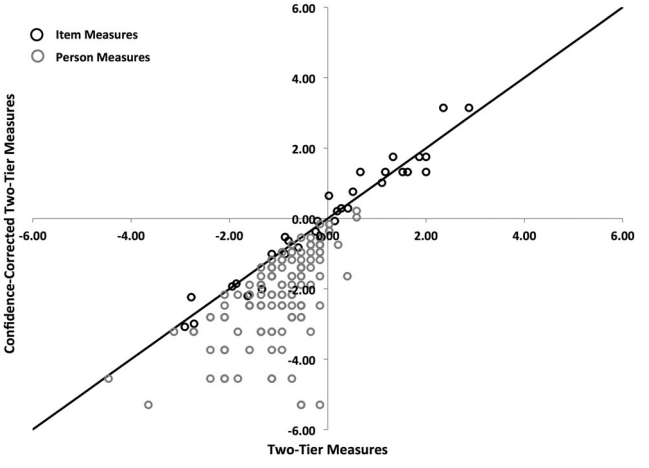

Figure 2. Comparison of two- and three-tiered person and item measures

These collective observations indicate that as additional tiers are added, the DTWC estimates a lower ability level for students despite similar item difficulty distributions. Increase in the standard deviation of the person measure distribution indicates that, as additional tiers are added, the efficacy of the DTWC in differentiating the undergraduates increases. This is also reflected by the observed increase in person measurement reliability as additional tiers are added.

The one-to-one line on Figures 1 and 2 measures the extent to which person and item measure distributions keep their similarity as additional tiers are added. That person measures sit below the line provides an illustration of how the person ability distribution drops as each successive tier is added. While the mean of item measures is centered at 0 for all scales (uniform distribution around the one-to-one line), we observe that changes in item estimates are much more prevalent with the addition of the second tier to the one-tiered test than with the addition of the third tier to the two-tiered test. While both comparisons are positively correlated, two-tiered and one-tiered item measures have a correlation of 0.81, indicating 65% of variance shared between one- and two-tiered item measures. The correlation of 0.97 between two- and three-tiered measures, indicating 95% shared variance, is much higher. This indicates that while adding a second tier changes the nature of the DTWC scale and interpretation of items, adding the confidence tier tends to preserve the functioning of the two-tiered items while enhancing the efficacy of the scale for differentiating students.

How Effective is the DTWC in Revealing Misconceptions about the Water Cycle?

When the Rasch model was fit to data transformed into a coding scheme reflecting the misconception domain (1 = misconception; 0 = no misconception), we found that the items and resultant scale demonstrated adequate construct validity and reliability. In addition, it was an informative predictor of prevalence of misconceptions about the

water cycle in students, and the tendency of individual items to reveal misconceptions. Item infit values ranged between 0.89 and 1.14. With the exception of item V12 (evaporation), which displayed an outfit value of 2.45, and V7 (water reservoirs), which displayed an outfit of 1.50, item outfits fell between 0.73 and 1.23. Collectively, these results indicate excellent item-level construct validity for identifying misconceptions. It is interesting that item V12, when used in a traditional framework, displayed outfits indicative of good fit with the Rasch model. Conversely, item V18, which displayed outfits above 5 when used in the traditional way, displayed an outfit of 0.84 when transformed into the misconception domain.

As with the measurement validity of items, the validity of the ensuing scale also demonstrated potential utility in predicting misconceptions displayed by persons and items. We found that the misconception scale was reasonably unidimensional, indicated by a first eigenvalue of 2.25 derived from PCA on Rasch residuals. We also found that items displayed reasonable local independence, indicated by a maximum inter-item correlation of 0.48 (V24 and V25). We measured reliabilities of 0.75 and 0.94 for person and item locations along this scale, respectively.

Prediction of Misconceptions. With the establishment of validity of items along the misconception scale with respect to the Rasch model, we proceeded to utilize this scale to evaluate the relative efficacy of items in revealing misconceptions, and conversely the tendency of students to express misconceptions. Observing the person-item map (Figure 3), we see that the item distribution ($M = 0.0003$, $SD = 0.98$) sits higher along the scale than the person distribution ($M = -1.25$, $SD = 1.06$). The person-item map demonstrates that items V6 (water reservoirs), V7 (water reservoirs), V11 (relative humidity), and V12 (evaporation) are not likely to reveal a misconception for any of the students. Items V6, V7, and V12 are also among the easier items along the three-tiered ability continuum, with difficulties of –2.24, –2.99, and –3.08, respectively, indicating that students tend to have mastery of concepts measured by these items. On the other hand, item V11 sits in the middle of the three-tiered ability continuum with a difficulty of −0.53 indicating that students choosing the incorrect responses to this item tend to lack confidence while students tend to have greater confidence in their correct responses. Items V6 and V7 address students' understanding of sources of moisture found on Earth. Item V12 measures students' ability to identify an evaporation process, and item V11 asks students to identify the relationship between the temperature of air and its capacity for holding moisture. While these items did not reveal misconceptions, they are informative in their indication that college students tend to have mastery of these topics.

Items V4 (evaporation kinetics), V20 (displacement), and V28 (density of humid air) demonstrated the greatest efficacy in revealing misconceptions. Of the 130 students, the Rasch model predicted that 69 (53%) students would reveal a misconception in item V4, 111 (85%) would reveal a misconception on item V20, and 98 (75%) would reveal a misconception on item V28. While this construction says nothing about particular misconceptions revealed by the instrument, a distracter analysis of these

Figure 3. Person-item map of the Rasch misconceptions scale

items sheds light onto the particular tenacious misconceptions about the water cycle that exist within a significant proportion of college students.

What are the Most Persistent Misconceptions about Hydrologic Processes?

Item V4 probes a student's understanding about the relationship between the rate of evaporation and atmospheric temperature. The correct answer is that *rate decreases as air becomes colder* because *air's capacity to hold water is a function of temperature.* However, 55 (42%) of the students chose the response, *rate decreases as air becomes colder* because *cooler air provides a lower amount of energy for evaporation* with a high level of confidence. Item V28 asks students to compare the density of dry and humid air. The correct answer is that, *compared to dry air, humid air is less dense because water molecules are lighter than the average of other molecules in the atmosphere.* However, 72 (55%) of the students selected the response, *humid air is more dense because hydrogen bonding changes as the air becomes humid* with high confidence. Item V20 asks students whether floating sea ice will cause sea levels to rise, fall, or stay the same. The correct answer is *the melting of floating sea ice will cause no change in sea levels because sea water in both liquid and ice states have the same mass.* However, 84 (65%) of the students selected their response, *the sea level will rise because the extra water produced due to the melting will cause sea level to rise and flood coastal areas.*

What is the Relationship between Students' Prior Experience with the Water Cycle on their Tendency to Express Misconceptions?

We found no significant differences in prevalence of misconceptions in students across the various levels of prior experience. Mean locations on the Rasch misconception scale did not vary significantly based on when students last learned about the water cycle ($F_{3,116} = 0.12$, $p = .95$) or whether or not they had completed an undergraduate meteorology/atmospheric science class ($F_{1,116} = 0.018$, $p = .90$). Further, we found no differential effect of when students last learned about the water cycle between students who had and had not taken an undergraduate meteorology/atmospheric science class ($F_{2,116} = 0.538$, $p = .59$).

We found that the main effect of institution was significant ($F_{1,116} = 5.653$, $p = .019$) at the 95% confidence level. The mean location on the misconceptions scale for students sampled from the large research-extensive university was –1.54 (SE = 0.20) while the mean location for students sampled from the small liberal arts college was –0.90 (SE = 0.23). None of the interactions involving institutional effects were significant, indicating the absence of differential effects of when students last learned about the water cycle, and whether or not students had taken an undergraduate meteorology/atmospheric science class, between institutions. We note that while controlling for the difference between the samples from different institutions was important for ensuring trustworthy statistical tests of the variables of interest, our study design does not permit generalization of institutional effects outside of the specific context of this study.

Discussion

Differential Item- and Scale-Level Validity across Tiers

In this study, we demonstrate the efficacy of a multi-tiered instrument format in the context of a novel assessment of undergraduates' understanding about the water cycle. Proper understanding of the water cycle requires a diverse background cutting across the disciplines of physics, chemistry, biology, and earth/atmospheric science. Consequently, it can be challenging to collect reliable, unidimensional measures. Through the DCWT, we demonstrate that utilization of a multi-tiered measure enables precise, unidimensional measurement of such a diverse construct feasible.

A significant advantage of MTDA formats such as the three-tiered DTWC is their diversity; they allow an instructor or researcher to assess students in a variety of ways according to practical circumstances such as time, classroom setting, the need for descriptive power, or the need for precision. When used as a one-tiered test, we find that the DTWC provides a unidimensional measure constructed of locally independent items which fit well with the Rasch model. High overlap of person and item measure distributions indicates that the one-tiered test is written at a difficulty well targeted to college students. However, the reliability of the one-tier DTWC is only slightly above the minimum value of 0.5 that the What Works Clearinghouse (WWC, 2014) considers adequate for research purposes. We therefore caution against using the DTWC in its one-tiered format due to lack of precision.

When used as a two-tiered test, reliability of the DTWC increases above 0.6, entering a range where person measures can serve the practical end of group comparisons. In addition, we find that unidimensionality of the scale improves. Asking students to justify their first-tier response serves to eliminate outside factors unrelated to knowledge of the water cycle which may influence their measures on the DTWC such as guessing of the correct answer. Zimmerman and Williams (2003) noted that likelihood of guessing can be offset in short tests by including five or more response options. Addition of the second tier to the DTWC increases the number of response options greatly, thereby improving the assessment's reliability and unidimensionality (Zimmerman & Williams, 2003).

However, adding the second tier had some negative effects on test validity from the Rasch perspective. After adding the second tier, items V18, V24, and V28 took extreme outfit values (greater than 1.5), indicating that certain students with extremely low measures got these items correct, and conversely, students with extremely high measures missed these items. While extreme outfit is a much lower threat to test validity than extreme infit, it is nonetheless useful to look at these questions and explore what may be causing the misfit. Item V18 asks, 'On a beautiful morning in late November, you go outside and all the windows on your car are covered with frost. Why did this frost form?' The first-tier choices included; '(1) Condensation (2) Deposition (3) Temperature change or (4) Sublimation.' The correct second-tier response was, 'A direct change from a gas to a solid regardless of temperature.' Analysis of the response pattern for this item indicates a nearly equal proportion of low-ability as

high-ability students getting this item correct, which accounts for its misfit with the Rasch model. Guessing is unlikely the cause for the misfit of V18 given that the introduction of the confidence tier did little to improve model fit. Rather, it appears that students who tended to be familiar with other aspects of the water cycle were not familiar with the process of deposition.

Item V24 asks, 'When air rises, it: (1) expands and cools, (2) expands and warms, (3) compresses and cools, or (4) compresses and warms.' The correct second-tier response is, 'Atmospheric heating causes the air to rise, and then expand.' We see some disconnect between the first and second tiers. The student who selected 'expands and cools' as a first-tier response may have experienced some confusion with the reference to atmospheric heating in the second tier. A closer analysis of the response pattern for item V24 reveals that while several students in the bottom half of the ability distribution chose the correct response set, there were no students from the upper half of the ability distribution who got this item correct. Item V28 asks, 'When compared to dry air, humid air is: (1) less dense, (2) more dense, or (3) of equal density.' The response pattern for V28 indicates a small negative relationship ($\rho = -0.04$) between a student's ability and his/her tendency to get this item correct. Since adding the confidence tier improved the fit of items V24 and V28, we can implicate educated (but unconfident) guessing as a possible reason for misfit of these items in the two-tiered format.

Integration of confidence into mastery does little to change the functioning of the items on the DTWC as observed by the high correlation between item measures when used in two- and three-tiered formats, respectively (Figure 3). However, adding confidence does change how items fit with the Rasch model and the nature of the scale. The addition of a second tier to a single-tiered instrument, introduction of a third tier CRI, and making confidence a necessary condition for knowledge further lowered the distribution of students' ability measures. However, it also increased the variance of the scale, leading to improvement of measurement reliability. Scale reliability above 0.7 is useful for measuring students across a variety of contexts where group comparisons are needed. Despite a higher reliability, it is interesting that integrating confidence into mastery increased the multidimensionality of the assessment slightly. The framework used in this study attempts to combine the dimensions of knowledge and confidence into a single score, and we justify this by making the case that confidence is a necessary component of true knowledge. Change in instrument dimensionality as measured by PCA on Rasch residuals can be used as an objective test as to whether or not this is actually the case. The first eigenvalue from PCA on residuals rose from 2.24 to 2.33; this miniscule change and increase in measurement reliability indicates that while the three-tiered version is not as unidimensional as the two-tiered version, integration of confidence nonetheless increases the utility of the DTWC as a measure for knowledge of the water cycle.

Introduction of confidence into the scale for knowledge, while improving many aspects of scale validity, seemed to damage the validity of certain items with respect to the Rasch framework. Outfit indices for items V13, V18, and V29 indicate high misfit with the Rasch model. That these items fall in the middle of the

misconceptions scale indicates that presence of misconceptions in high-ability students is likely not to blame. Item V13 asks, 'What forms directly due to the process of condensation?' The four first-tier options included, '(1) water vapor, (2) clouds, (3) rain, and (4) snow.' The question proceeded to give second-tier options, '(1) increases in temperature, (2) decreases in temperature, (3) increases in pressure, and (4) decreases in temperature.' In light of the data, it appears that the wording of these options is problematic because they are disconnected from the first-tier options, and tend to leave the student wondering whether this is referring to clouds in general, or conditions for condensation. Since this item in its two-tier usage fits well with the Rasch model, we can conclude that the wording of this item did not directly mislead students, but rather reduced students' confidence in their answers. A distracter analysis indicates that a majority of students choosing the correct response to this item did so with low confidence. The simple revision of adding 'And condensation is caused by … ' to preface the second-tier responses may improve the wording of this item, thereby increasing upper-level students' confidence in their responses. Item V29 reveals a similar phenomenon as item V13 in that it fit well with the Rasch model as a two-tiered item, but misfitted as a three-tiered item. Item V29 asks students to identify the true shape of a raindrop, giving three options, one of which includes the traditional teardrop shape, another of which is a perfect circle, and the correct flattened elliptical shape. The correct reasoning for the flattened elliptical shape was, 'air pressure increases causing the drop to flatten.' We found that only 6 of 23 students choosing the correct first-tier response and second-tier reasoning did so with high confidence. In addition, 58 students selected the teardrop shape with the reasoning that gravity causes the drop to deform as it falls from the sky. However, that 31 of these 58 students chose this response with high confidence is indicative of a tenacious misconception in these students. These factors likely led to misfit of item V29 when confidence was introduced into the scale.

In addition to several items misfitting the Rasch model under two- and three-tiered usages, we also find better-than-expected fit with the Rasch model in two items (V19 and V21). In these items, we observe the general trend that fit with the Rasch model increases as additional tiers are added. Both approached an outfit of 0.5 after confidence was added. Unusually good fit with the Rasch model results when the item measures underlying dimensions positively correlated with knowledge of the water cycle (Masters, 1988). In many testing applications, it can be difficult to specifically pin down the specific additional dimension that is favoring high-ability students, or discriminating against low-ability students. Comparison across tier usages makes it clear that confidence serves as that underlying dimension. Confidence resulting in model overfit implies that the high-ability students tend to be more confident in their responses than low-ability students. Therefore, integration of confidence into mastery serves to bias these items in favor of high-ability students. Whether or not this is a problem depends on one's perspective regarding whether knowledge and confidence should be combined or treated separately.

A Cross-Tier Description of Item Validity

We discussed previously several patterns revealed by the three-tiered testing format that could be considered useful for evaluating the validity and utility of a three-tiered test such as the DTWC. We first observe that adding a second tier to an assessment changes the functioning of items to a significant degree; however, adding a third confidence tier does not tend to shift items along the scale much. We also observe that as additional tiers are added, the student ability distribution drops. Since three different usage styles give three different distributions for student ability, an important question arises: which distribution do we trust? Since MTDAs are built upon constructivist traditions, we can reply to this question with a constructivist answer: it depends on how the researcher defines knowledge. Given a significant increase in reliability and unidimensionality when a second tier is added, we feel a recommendation against using the DTWC as a single-tiered assessment is well justified. Its usage diminishes not only the psychometric integrity of the scale, but also the test's descriptive utility with regard to misconceptions, vitiating a primary purpose of the DTWC. However, given that scales derived from two- and three-tiered usage structures exhibit high similarity, and that the descriptive utility of the second tier is engaged, the decision between two- and three-tiered usages comes down to the question of whether or not confidence should be considered a necessary condition for mastery. From a psychometric perspective, much of this depends upon whether or not knowledge and confidence should be combined, or whether they should be treated as separate dimensions. Rasch analysis of the DTWC sheds some objective insight. We observed that multidimensionality of the test increased slightly when confidence was added. However, reliability of the scale, and its utility for differentiating students, increased. All usages of the DTWC meet the liberal criterion (first eigenvalue below 3 items of variance) for unidimensionality used by Galli, Chiesi, & Primi (2008). We therefore argue that, despite a slight increase in multidimensionality, integrating confidence as an important component of mastery improves the scale by eliminating confounding factors such as guessing that tend to damage instruments' validity and reliability.

Integration of confidence had different effects on different items. Two items (V24 and V28) misfitted with the Rasch model when used as two-tiered, but fitted well when confidence was integrated. This indicates that a number of students were guessing on these items (i.e. choosing the correct answer with low confidence) which significantly damaged their validity, but was corrected when integrating confidence. Conversely, adding confidence appeared to corrupt the validity of three items (V3, V13, and V29) with respect to the Rasch model. Given that good fit with the Rasch model accords with the observation that high-ability students tend to get an item correct, and low-ability students tend to miss the item, this result was indicative of the tendency for high-ability students to choose the correct response with low confidence. This could also result from the presence of a tenacious misconception in high-ability students. There was one instance (V18) where high misfit with the Rasch model resulted from both two- and three-tiered usages. It is probably an

indicative of confusing wording or lack of understanding of deposition that caused high performers to choose the wrong response with high confidence.

Some tentative guidelines for interpreting differential fit with the Rasch model between two- and three-tiered usages are outlined in Table 3. These guidelines are based on the empirical data collected in the context of this study and logic based on how the Rasch model behaves. We recommend that researchers employing three-tiered assessments consider utilizing these guidelines with skepticism, evaluating their efficacy with respect to novel data sets.

Interpretation of DTWC Measures

Scale development and validation is typically undertaken with the positivist notion that there is an underlying truth to be measured. Consequently, the value of an item for measurement is evaluated by the extent to which it gets at the defined underlying truth, such as knowledge of the water cycle. Indeed, the utility of evaluating item validity based on fit with the Rasch model is laid upon the foundation of falsification, which is of fundamental importance to the positivist approach to scientific inquiry (Popper, 1987). The discussion above and the guidelines outlined in Table 3 call for evaluation of item validity based on the more complex constructivist framework, which underlies the necessity for multi-tiered assessment in the first place. For example, if scale reliability and validity were the only thing to consider, then we could implement a positivist perspective and make the case that the item which underfits the Rasch model and reduces scale reliability degrades the measurement scale and

Table 3. Tentative guidelines for interpreting fit with the Rasch model resulting from two- and three-tiered usage styles

Two-tiered fit	Three-tiered fit	Implication
Good	Good	Item is worded non-ambiguously. No imbalance of tenacious misconceptions across student ability. Confidence does not favor high or low-ability students
Good	Underfit	Confidence favors low-ability students. Possible wording issue that makes high-ability students less confident in their response. Possible tenacious misconception in high-ability students
Underfit	Good	Validity of the item damaged by guessing, which is corrected by integrating confidence
Underfit	Underfit	Confusing wording that leads to high-ability students choosing the incorrect answer with high confidence or leading wording that leads low-ability students to the correct response with high confidence. Possible tenacious misconception in high-ability students
Good	Overfit	High-ability students show greater confidence than low-ability students
Overfit	Overfit	Use of complex wording that favors high-ability students. High-ability students show greater confidence than low-ability students

should be excluded from the DTWC. However, we observe in Table 3 that an item which underfits the Rasch model, while degrading the measurement scale, also exhibits interesting properties that make it useful from the perspective of evaluating misconceptions. A prime example is item V29, which misfitted the Rasch model after confidence was integrated into the scale. This does not mean that V29 is a bad item; rather, this item's misfit is indicative of a tenacious misconception in high-ability students. The utility of item V29 in quantifying a student's location along a misconceptions scale is highlighted by its good fit with the Rasch model after the data were transformed into the misconceptions domain.

Out of this analysis arises the practical question: How do we use the DTWC? While the shift to a constructivist model for assessment contraindicates hard and fast rules, the Rasch model nonetheless yields insight into the strengths and weaknesses of particular items for specific uses. For example, if the only goal in using the DTWC in a two-tiered format is to give students a location along the scale of knowledge of the water cycle, then item V29 could be considered productive. If confidence were integrated into the scale, our data show that the utility of V29 diminishes significantly; perhaps V29 could be excluded in this context to the end of improving the measure. On the other hand, if the goal is only to uncover student misconceptions, then V29 can be considered among the most informative items on the DTWC.

If predicting students' prevalence of misconceptions, items V7 (addressing water reservoirs) and V12 (addressing evaporation) which work well in one-, two-, and three-tiered formats actually degrade the Rasch misconceptions scale and could possibly be excluded. In Figure 3, we observe that both V7 and V12 are at the top of the scale well above the person distribution. This indicates a model expectation that none of the students would express misconceptions on these items. However, the data indicate that nine students on V7, and five students on V12, expressed a misconception, and correlations between students' logit measures and responses ($\rho = 0.03$ for V7 and $\rho = 0.06$ for V12) indicate little relationship between a student's general tendency to have misconceptions on the DTWC in general and expressing specific misconceptions on these items. If the goal is to increase precision of students' locations on the misconception scale, then perhaps these items should be removed. Another case for removal of these items could be made based on the fact that a very small proportion of students actually expressed misconceptions on these items. However, these items are informative in the sense that they indicate that college students express fewer misconceptions about water reservoirs and evaporation than other topics of the water cycle. Location of these items near the bottom of the two- and three-tiered scales indicates that non-prevalence of misconceptions on these items is due to students knowing the correct answers to these questions, and not lack of confidence.

Conclusion

We found that our undergraduate students' tendency to express misconceptions about the water cycle did not change significantly with prior academic experience. Our data

validate Brody's (1993) finding, illustrating the ineffectiveness of science coursework in helping students adopt scientific conceptualizations about water. We consider the DTWC a valid and reliable instrument that can be used by science content and method instructors to measure a student's conceptual understanding of the water cycle so that his/her misconceptions can be addressed in an informed and focused way. In addition, the DTWC contributes to what is among the most underrepresented areas in science education assessment—the geosciences (Schaffer, 2013). The DTWC is a powerful tool allowing teachers to account for students' current knowledge of a concept before teaching begins (Driver & Easley, 1978). Proper diagnostics will lead to the development of more pointed and effective teaching strategies and interventions aimed at eliminating students' barriers to mastery of the water cycle.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References

Aguado, E., & Burt, J. E. (2004). *Understanding weather and climate*. Upper Saddle River, NJ: Pearson Education.

American Association for the Advancement of Science (AAAS). (1990). *Science for all Americans*. New York, NY: Oxford University Press.

American Meteorological Society. (2001). *The global water cycle: An introduction*. Boston, MA: Author.

Arhens, C. D. (2009). *Meteorology today: An introduction to weather, climate, and the environment*. Belmont, CA: Brooks/Cole.

Arslan, H. O., Cigdemoglu, C., & Moseley, C. (2012). A three-tier diagnostic test to assess pre-service teachers' misconceptions about global warming, greenhouse effect, ozone layer depletion, and acid rain. *International Journal of Science Education*, *1*, 1–20.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart, and Winston.

Bar, V. (1989). Children's views about the water cycle. *Science Education*, *73*, 481–500.

Bar, V., & Galili, I. (1994). Stages of children's views about evaporation. *International Journal of Science Education*, *16*(2), 157–174.

Bar, V., & Travis, A. (1991). Children's views concerning phase changes. *Journal of Research in Science Teaching*, *28*, 363–382.

Bell, B. (1995). Interviewing: A technique for assessing science knowledge. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 347–364). Mahwah, NJ: Lawrence Erlbaum.

Ben-zvi-Assarf, O., & Orion, N. (2005). A study of junior high students' perceptions of the water cycle. *Journal of Geoscience Education*, *53*, 366–373.

Bond, T. G., & Fox, C. M. (2007). *Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple choice tests. *Science Education*, *90*(2), 253–269.

Brody, M. J. (1993, April). *Student understanding of water and water resources: A review of literature*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Caleon, I. S., & Subramaniam, R. (2010a). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, *32*, 939–961. doi:10.1080/09500690902890130

Caleon, I. S., & Subramaniam, R. (2010b). Do students know what they know and what they don't know? Using a four-tier diagnostic test to access the nature of students' alternative conceptions. *Research in Science Education*, *40*, 313–337. doi:10.1007/s11165-009-9122-4

Cardak, O. (2009). Science students' misconceptions of the water cycle according to their drawings. *Journal of Applied Sciences*, *9*, 865–873.

Carr, M. (1996). Interviews about instances and interviews about events. In D. F. Treagust, R. Duit, & B. J. Fraser (Eds.), *Improving teaching and learning in science and mathematics* (pp. 44–53). New York: Teachers College Press.

Cetin-Dindar, A., & Geban, O. (2011). Development of a three-tier test to assess high school students' understanding of acids and bases. *Procedia—Social and Behavioral Sciences*, *15*, 600–604.

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, *8*, 293–307.

Covitt, B. A., Gunckel, K. L., & Anderson, C. W. (2009). Students' developing understanding of water in environmental systems. *The Journal of Environmental Education*, *40*(3), 37–51.

D'Avanzo, C. (2003). Research on learning: Potential for improving college ecology teaching. *Frontiers in Ecology and the Environment*, *1*(10), 533–540.

Driver, R., & Easley, J. (1978). Pupils and paradigms: A review of literature related to concept development in science. *Studies in Science Education*, *13*, 105–122.

Driver, R., & Erickson, G. (1983). Theories-in-action: Some theoretical and empirical issues in the study of students' conceptual frameworks in science. *Studies in Science Education*, *10*, 37–60.

Duit, R. (2009). *Bibliography—STCSE: Students' and teachers' conceptions and science education*. Retrieved from: ReindersDuit duit@ipn.uni-kiel.de

Galli, S., Chiesi, F., & Primi, C. (2008). The construction of a scale to measure mathematical ability in psychology students: An application of the Rasch Model. *TPM (Testing Psicometria Metodologia)*, *15*(1), 1–16.

Gunckel, K. L., Covitt, B. A., Salinas, I., & Anderson, C. W. (2012). A learning progression for water in socio-ecological systems. *Journal of Research in Science Teaching*, *49*(7), 843–868.

Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education*, *34*(5), 294–299.

Henriques, L. (2000, April). *Children's misconceptions about weather: A review of the literature*. Paper presented at the annual meeting for the National Association for Research in Science Teaching, New Orleans, LA.

Lin, S. W. (2004). Development and application of a two-tier diagnostic test for high school students' understanding of flowering plant growth and development. *International Journal of Science and Mathematics Education*, *2*, 175–199.

Linacre, J. M. (1999). Relating Cronbach and Rasch reliabilities. *Rasch Measurement Transactions*, *13*, 696.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2009). Local independence and residual covariance: A study of Olympic figure skating ratings. *Journal of Applied Measurement*, *11*, 157–169.

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

Linacre, J. M., & Tennant, A. (2009). More about critical eigenvalue sizes (variances) in standardized-residual principal components analysis (PCA). *Rasch Measurement Transactions*, *23*(3), 1228.

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

Lutgens, F. K., & Tarbuck, E. J. (2010). *The atmosphere: An introduction to meteorology*. Upper Saddle River, NJ: Pearson Education.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*(1), 15–29.

Modell, H., Michael, J., & Wenderoth, M. B. (2005). Helping the learner to learn: The role of uncovering misconceptions. *The American Biology Teacher*, *67*(1), 20–26.

Morrell, P. D., & Schepia, A. (2009, January). *Exploring preservice elementary teachers' conceptualization of the water cycle*. Paper presented at the annual conference of the Association for Science Teacher Education, Hartford, Connecticut.

Munson, B. H. (1994). Ecological misconceptions. *Journal of Environmental Education*, *25*(4), 30–35.

NGSS Lead States. (2013). *Next generation science standards: For states, by states. Achieve, Inc. on behalf of the twenty-six states and partners that collaborated on the NGSS*. Washington, DC: National Academy Press.

Novak, J. D. (1995). Concept mapping: A strategy for organizing knowledge. In S. M. Glynn& R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 229–245). Mahwah, NJ: Lawrence Erlbaum.

Novak, J. D. (2002). Concept mapping: A tool for improving science teaching and learning. In D. F. Treagust, R. Duit, & B. J. Fraser (Eds.), *Improving teaching and learning in science and mathematics* (pp. 32–43). New York: Teachers College Press.

Odom, A. L. (1992). *The development and validation of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis* (Unpublished doctoral dissertation). University of Missouri, Columbia, MO.

Odom, A. L., & Barrow, L. H. (1995). Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. *Journal of Research in Science Teaching*, *32*, 45–61. doi: 10.1002/tea.3660320106

Odom, A. L., & Barrow, L. H. (2007). High school biology students' knowledge and certainty about diffusion and osmosis concepts. *School Science and Mathematics*, *107*(3), 94–101.

Pesman, H., & Eryilmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *Journal of Educational Research*, *103*, 208–222.

Phillips, W. C. (1991). Earth science misconceptions. *Science Teacher*, *58*(2), 21–23.

Popper, K. (1987). Science: Conjectures and refutations. In J. A. Kourany (Ed.), *Scientific knowledge: Basic issues in the philosophy of science* (pp. 139–157). Belmont, CA: Wadsworth.

Posner, G. J., & Gertzog, W. A. (1982). The clinical interview and the measurement of conceptual change. *Science Education*, *66*, 195–209. doi: 10.1002/sce.3730660206

Project WET and Council for Environmental Education. (1995). *The incredible journey* (Project WET Curriculum and Activity Guide, pp. 161–165), Bozeman, MT: Water Conversation Council for Environmental Education. Retrieved from www.projectwet.org

Raiche, G. (2005). Critical eigenvalue sizes in standardized residual principle components analysis. *Rasch Measurement Transactions*, *19*(1), 1012.

Romine, W. L., Barrow, L. H., & Folk, W. R. (2013). Exploring secondary students' knowledge and misconceptions about influenza: Development, validation, and implementation of a multiple-choice influenza knowledge scale. *International Journal of Science Education*, *35*(11), 1874–1901.

Ruiz-Primo, M. A. (2000). On the use of concept maps as an assessment tool in science: What we have learned so far. *Revista Electronica de Investigacion Educativa*, *2*(1), 29–53.

Russell, T., Harlen, W., & Watt, D. (1989). Children's ideas about evaporation. *International Journal of Science Education*, *11*, 566–576. doi:10.1080/0950069890110508

Schaffer, D. L. (2013). *The development and validation of a three-tier diagnostic test measuring pre- service elementary education and secondary science teachers' understanding of the water cycle* (Doctoral dissertation), University of Missouri—Columbia. Retrieved from https://mospace.umsystem.edy/xmlui/bitsteam/handle/10355/37809/research.pdf?sequence=2

Schaffer, D. L., & Barrow, L. H. (2015). Alternative conceptions or lack of knowledge: Assessing pre-service teachers' understanding of the water cycle. *Journal of Science Teacher Education*. Manuscript submitted for publication.

Schaffer, D. L., Romine, W. L., & Barrow, L. H. (2015, April). *Misconception or lack of knowledge*: *Using confidence to enhance measurement validity in a multi-tiered assessment*. Paper to be presented at the 2015 annual meeting for the National Association for Research in Science Teaching, Chicago, IL.

Shepardson, D. P., Wee, B., Priddy, M., Schelleberger, L., & Harbor, J. (2009). Water transformation and storage in the mountains and at the coast: Midwest students' disconnected conceptions of the water cycle. *International Journal of Science Education*, *31*, 1447–1471. doi:10.1080/09500690802061709

Smith, III, J. P., di Sessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, *3*(*2*), 115–163.

Stamp, N., & Armstrong, M. (2005). Using "The Power of Story" to overcome ecological misconceptions and build sophisticated understanding. *Bulletin of the Ecological Society of America*, *86* (3), 177–183.

Taiwo, A. A. (1999). Perceptions of the water cycle among primary school children in Botswana. *International Journal of Science Education*, *21*, 413–429. doi:10.1080/095006999290633

Treagust, D. (1986). Evaluating students' misconceptions by means of diagnostic multiple choice items. *Research in Science Education*, *16*, 199–207.

Treagust, D. F. (1988). The development and use of diagnostic instruments to evaluate students' misconceptions in science. *International Journal of Science Education*, *10*, 159–169.

Treagust, D. F. (1995). Diagnostic assessment of students' science concepts. In S. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 327–346). Mahwah, NJ: Lawrence Erlbaum.

Tsai, C. C., & Chou, C. (2002). Diagnosing students' alternative conceptions in science. *Journal of Computer Assisted Learning*, *18*, 157–165. doi:10.1046/j.0266-4909.2002.00223

Wang, J. R. (2004). Development and validation of a two-tier instrument to examine understanding of internal transport in plants and the human circulatory system. *International Journal of Science and Mathematics Education*, *2*, 131–157.

Wandersee, J. H., Mintzes, J. J., & Novak, J. D. (1994). Research on alternative conceptions in science. In D. L. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 177–210). New York: Macmillan.

What Works Clearinghouse. (2014). *WWC procedures and standards handbook* (version 3.0). Princeton, NJ: US Department of Education, Institute of Education Sciences.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

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

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement*, *27*(5), 357–371.