Looking Beyond Observations:

Observed Scores versus Rasch Measures for the Analysis of Efficacy Data

BY

LISA J. LYNN
B.A., University of Illinois at Chicago, Chicago, 2009
M.Ed., University of Illinois at Chicago, Chicago, 2011

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Chicago, 2017

Chicago, Illinois

Defense Committee:

Kimberly Lawless, Ph.D., Chair and Advisor
Scott W. Brown, Ph.D., University of Connecticut
Everett Smith, Ph.D.
Yue Yin, Ph.D.
Mariya A. Yukhymenko, Ph.D., California State University Fresno

ACKNOWLEDGMENTS

"It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment."

- Carl Friedrich Gauss

A dissertation is nothing if not an act of learning, an intellectual adventure, and there are many people to thank for accompanying me on this most fulfilling journey. First is my advisor Kim, who believed in me from the moment I entered the program and guided me every step of the way. My committee members: Yue, Ev, Scott, and Masha. Thank you for giving me the support to thrive and for showing me the respect of insisting upon my best work. I am proud to have worked with each of you.

Besides my committee members, I would like to thank Carol Myford for feedback and encouragement. My appreciation also goes to several anonymous reviewers and an anonymous mentor, for their insights and honesty about the shortcomings of my work and for their advice on focus and persistence.

Thank you to my academic sisters and brothers, especially Kamila, James, and Jeremy. Together we laughed, we cried, we got published, we got rejected, we laughed some more. I have learned so much from you and look forward to a conversation that continues well beyond graduation and for the rest of our careers.

ACKNOWLEDGMENTS (continued)

Thank you to my parents, my brother, my extended family, and my stepdaughters Becca and Emma, for supporting my education and also reminding me that I have a life and a purpose outside the halls of the academy. Thank you to my grandparents for being my biggest fans. I wish you were all here to witness this accomplishment, but I know that I was a success in your eyes long before I had any academic credentials.

Lastly, I must thank my husband Joe, who may not know exactly what a violation of homogeneity of regression slopes is, but knows how to comfort me through every violation, non-significant result, and dead end. You gave me unconditional love and support and, in our most difficult moments, Papa John's pizza. I never would have survived this "act of learning" without you.

LJL

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
APA	American Psychological Association
CTT	Classical test theory
DIF	Differential item functioning
ICC	Intra-cluster correlation
IES	Institute of Education Sciences
PBL	Problem-based learning
WWC	What Works Clearinghouse

SUMMARY

Funders of educational research devote hundreds of millions of dollars each year to support theoretically sound educational interventions and research on these interventions. Unfortunately, relatively few of these interventions have been established as clearly effective; most have been shown to be weak or ineffectual compared with normal educational practice (Coalition for Evidence-based Policy, 2013). The purpose of this study was to explore one possible reason for mixed findings in educational intervention studies: the use of observed scores in parametric statistics, where interval measures should be used. The study compared the statistical conclusions derived from observed scores and Rasch measures of several outcomes, using exemplar data from a large, grant-funded, PBL intervention study.

Consistent with prior research, results showed that for an objective test, attitude measures, and a rater-scored essay, linear relationships between Rasch person measures and observed scores were very strong ($r > .86$). For the seven outcomes tested, Rasch person measures and observed scores consistently agreed on the statistical significance of findings. Based on the dataset and analyses used in this dissertation, there is no evidence to conclude that the use of observed/ordinal scores is a likely culprit for null or mixed findings. This is in contrast with certain earlier findings that showed at least one case wherein Rasch person measures and observed scores provided divergent statistical conclusions. Present findings suggest that, when sample size is adequately large, and when measures are already determined (i.e. Rasch analysis is not being used for scale development) and meet the general standards of reliability and model fit, there is little to be gained by using Rasch modeling to convert ordinal observed scores into interval

Rasch person measures. There is no basis to recommend that researchers use Rasch modeling in this manner to improve the quality of their interpretation of efficacy or conclusion of the impact of an intervention.

Chapter 1: Introduction

A recent report on educational intervention impact evaluations funded by the Institute of Education Sciences (IES) found that only 10-15% were clearly effective; the rest had weak effects or no positive effects compared to what schools were already doing (Coalition for Evidence-Based Policy, 2013). This is somewhat surprising, given that IES is one of the largest and most prestigious funding agencies for educational research. IES, the research arm of the U.S. Department of Education, has a budget of over $200 million and a mission to "provide rigorous and relevant evidence on which to ground education practice and policy and share this information broadly" (IES, n.d.a). IES is a major force in funding educational research, and thereby in determining the direction that educational research follows on a national scale. IES also administers the What Works Clearinghouse (WWC), a project that seeks to collect and evaluate evidence on educational interventions for researchers, educators, and policymakers (IES, n.d.b). IES and its subsidiary WWC are key organizations that influence what interventions and research are executed (by way of IES funding) and whether interventions are believed to "work" (by way of WWC evaluations). WWC includes in its mission the goal of providing "scientific evidence" for educational interventions (IES, n.d.b) and has set out detailed standards for what constitutes scientific evidence with emphasis on randomized controlled studies and quantitative outcomes. Ideally, IES and WWC together would fund effective educational intervention studies and promote widespread adoption of the best of them. However, with so few interventions found to be effective, educational practitioners, administrators, and policymakers are left with a paucity of proven effective interventions from which to choose. (There are numerous other funding agencies, including the National Science

Foundation and National Institutes of Health, which contribute substantially to education research, as well as innumerable educational research journals that disseminate the findings of educational intervention impact studies. The findings of this dissertation are intended to apply broadly to educational intervention impact studies, not to be limited to studies funded by IES or evaluated by WWC.)

Researchers have weighed in with various possible reasons for the apparent ineffectiveness of these interventions that were thought to be promising. Dosage is potentially an issue; if students are not actually getting much exposure to an intervention due to excessive absences or demands on classroom time that leave little time for a new program, they are likely not to benefit from it (Chaney, 2015; Viadero, 2009). Varying school and community contexts may play a role in findings of effectiveness, as well. A school or district that lacks funds to begin with will have limited funds to support an intervention, for example (Chaney, 2015; Viadero, 2009). Dosage and context both play a role generally in fidelity of implementation, or the similarity between the intervention as it was designed and the intervention as it is carried out in schools and classrooms (Song & Herman, 2010). In addition to the factors of how interventions play out in schools and communities, research design may influence whether findings are trustworthy and accurate. Poorly designed randomization or vague control procedures are detrimental to findings; it is crucial in randomized controlled designs that treatment and control groups are clearly defined, that all students in the intervention group and no students in the control group receive the intervention (Song & Herman, 2010; Viadero, 2009).

Several have pointed to measurement and analysis as potential reasons for lackluster findings. Where intervention and control groups are analyzed only in

aggregate, results may mask important findings for subgroups, such as certain populations or students with higher or lower self-efficacy; an intervention that appears ineffective "overall" might be extremely effective for certain groups of students (e.g. Chaney, 2015). Granger (2015) suggests that more innovative survey research, such as small, frequent measures done via cell phone, and sophisticated mediation analysis could improve the accuracy of findings and shed more light on the causal logic of interventions.

Two researchers involved in setting the WWC standards and evaluating studies addressed measurement issues and acknowledged the WWC initially set a "low bar" for measures, requiring face validity and minimal reliability, even when researchers are using their own measures created specifically to measure the impact of their intervention (Song & Herman, 2010, p. 361). Consequently, study results may suffer from a mismatch between desired outcomes and measures, most commonly by using measures that are much narrower than the constructs they intend to represent (Song & Herman, 2010). Low statistical power is also a possible detriment to intervention findings (Song & Herman, 2010). Statistical power refers to the likelihood of finding a correct significant result for a given sample size, effect size, and alpha level (Song & Herman, 2010). A study that has low statistical power runs the risk of incorrectly concluding that an intervention has no effect. When alpha is held at .05, a larger sample size or a larger effect size makes it more likely a study will find a correct significant result. Sample sizes in educational interventions can range from dozens to thousands; however, cluster randomization can reduce the power of even a very large study. In cluster randomization, classrooms or schools, rather than individual students, are assigned to treatment or control conditions. Because the pre-existing groups are more homogenous than a random sample of the

population would be, the use of clusters reduces the effective sample size, thereby reducing the statistical power (Hedges & Hedberg, 2007). Several years prior, Andrich (2002) contributed to the evaluation literature with an explanation of using desired educational outcomes to carefully define constructs and design precise measures aligned to outcomes in a Rasch framework. However, it is clear from reviews of IES grant proposals and WWC intervention reports that much of the most prominent research in education is using the simplest and most straightforward method of measurement: observed achievement scores, often gleaned from researcher-created measures. Among intervention reports for science programs on the WWC website, all five qualified studies had potential measurement issues: researcher-created assessments, marginal inter-rater agreement (alpha and agreement near 75%), and use of observed scores for parametric statistics (WWC, 2015).

As this dissertation will show, observed scores are not true measures of an academic or affective construct such as science knowledge, science interest, or writing quality (e.g. Andrich, 2002; Wright, 1999; Wright & Linacre, 1989; Wright & Stone, 1979), and there is a possibility that the use of observed scores, which are ordinal data, could mislead researchers by producing spurious significant results or masking significant effects. The conceptual and mathematical tools for creating true measures suited to use in parametric statistics have existed for decades (Rasch, 1960/1980; Wright & Stone, 1979). Dr. Benjamin Wright argued in 1999, “Today there is no methodological reason why social science cannot become as stable, as reproducible and as useful as physics.” However, these tools have not yet been widely adopted in education research.

Problem Statement

The federal government and private foundations spend billions of dollars funding research on educational interventions. Many of these interventions are promising; they are based on a solid theoretical framework and are well designed and executed. However, the educational research corpus is filled with interventions that yield null and mixed results.

In trying to understand how theoretically strong interventions fail to yield strong results, program evaluation researchers have proposed numerous potentially problematic factors, including individual and contextual differences among recipients, variations in the implementation of interventions, and several study design and analysis issues.

One factor that has not been discussed in the evaluation literature is the common mismatch between ordinal measures and parametric statistics, which require interval measures. This dissertation proposes Rasch modeling as a method for rescaling ordinal measures to interval data and seeks to demonstrate, using real data from a large, federally-funded educational intervention study, that true interval measures may reveal substantively different results compared to raw observed data.

Chapter 2: Review of Literature

The federal government and private funding agencies spend hundreds of millions of dollars annually funding educational interventions and research on educational interventions. Some of these interventions have clearly proven to be effective; however, studies of many interventions have failed to establish their efficacy. For example, a recent report on educational intervention impact evaluations funded by one agency, the Institute of Education Sciences (IES) found that only 10-15% were clearly effective; the rest had weak effects or no positive effects, compared to what schools were already doing (Coalition for Evidence-Based Policy, 2013).

The results of educational intervention efficacy studies are typically made public in one of dozens of peer-reviewed journals. The sheer number of published studies makes it a daunting task to monitor the findings of efficacy studies. In order to consolidate research across numerous journals, IES subsidiary What Works Clearinghouse (WWC) seeks to collect and evaluate published evidence on educational interventions for researchers, educators, and policymakers (IES, n.d.). WWC includes in its mission the goal of providing "scientific evidence" for educational interventions (IES, n.d., para. 1) and has set out detailed standards for what constitutes scientific evidence with emphasis on randomized controlled studies and quantitative outcomes. Ideally, WWC would promote widespread adoption of the most effective educational interventions. However, with a majority of interventions found to yield null or mixed results and so few interventions found to be clearly effective, educational practitioners, administrators, and policymakers are left with a paucity of proven interventions from which to choose.[1]

[1] There are numerous other funding agencies (e.g., Bill & Melinda Gates Foundation, MacArthur Foundation, National Science Foundation, and National Institutes of Health), which contribute substantially

The purpose of this review of literature is to examine the issue of null and mixed results in educational interventions. This review is meant to reflect issues of measurement in educational intervention studies generally. However, a review of the entire corpus of educational intervention efficacy research would be impractically broad; therefore, this review will focus on a single pedagogical approach, problem-based learning, which has shown promising results in certain contexts but remains controversial theoretically and empirically. This review will first define problem-based learning, and then will review prior research related to null and mixed findings in problem-based learning. Lastly, this review will explore a potential competing hypothesis for an explanation of null and mixed results: that the use of observed scores, which are ordinal data, may be degrading to the conclusion validity of educational intervention studies.

Problem-based Learning

Problem-based learning (PBL) is a constructivist approach to learning based on the philosophy that solving problems, especially problems encountered in real life or professional practice, provides a path to learning (Barrows & Tamblyn, 1980; Boud & Feletti, 1998; Hmelo-Silver, 2004; Hung, Jonassen & Liu, 2008; Savery & Duffy, 1995). In a PBL activity, students face an authentic, realistic, ill-structured problem, such as making a medical diagnosis given a difficult set of medical symptoms and data. Students work in small groups to define the problem and develop plans for potential solutions; students then work individually to gather information needed to solve the problem. Individuals share information with their small groups, and groups integrate the individual

---

to education research, as well as innumerable educational research journals that disseminate the findings of educational intervention impact studies. The findings of this literature review are intended to apply broadly to educational intervention impact studies, not to be limited to studies funded by IES or those evaluated by WWC.

contributions into potential solutions. During both group and individual work, students set goals, develop strategies, and assess their own progress at the group and individual levels. Teachers do not disseminate information, but rather support students in their self-directed learning. At the end of the unit, students present their proposed solutions and reflect on their process and learning (Hung et al., 2008).

Mixed Results in Problem-based Learning

Since it was pioneered as a medical school curriculum beginning in the 1970s (Barrows & Tamblyn, 1980; Neufeld & Barrows, 1974), PBL has found support both within and beyond the medical education community. Its success, compared to traditional lecture-based medical education (Albanese & Mitchell, 1993; Hmelo, 1998; Norman & Schmidt, 1992; Strobel & van Barneveld, 2009; Vernon & Blake, 1993), led to its adoption in other educational contexts, including being adapted for K-12 schools beginning in the 1990s (Dods, Segers, Van den Bossche, & Gijbels, 1997; Gallagher & Stepien, 1996; Gallagher, Stepien, & Rosenthal, 1992; Hmelo, Holton, & Kolodner, 2000). However, support for PBL has not been unequivocal. Opponents suggest that well-structured learning with strong guidance is better suited to human cognitive architecture, while less structured learning environments may lead to incomplete, disorganized, or partially incorrect knowledge (Kirschner, Sweller, & Clark, 2006). There is some empirical support for this argument: a meta-analysis of 43 PBL studies in higher education showed mixed results, with positive effects for acquisition of skills but potentially negative effects for the acquisition of knowledge, including many studies that yielded null results (Dochy, Segers, Van den Bossche, & Gijbels, 2003).

Researchers have pointed to numerous possible factors that account for the lack of consistency in PBL studies. In a direct response to Kirschner et al. (2006), Hmelo-Silver, Duncan, and Chinn (2007) challenged the assumption that PBL is not well supported by instructional guides and demonstrated that PBL employs extensive scaffolding to support complex learning. Related to scaffolding, Leary, Walker, Shelton, and Fitt (2013) studied the issue of support in PBL interventions by examining the differential effects of tutor experience and found that tutor training was associated with positive PBL effects (no tutor training $g = -.01$; tutor training $g = .29$), but tutor experience was associated with smaller PBL effects (inexperienced tutor $g = .31$; experienced tutor $g = .19$).

Other studies have looked beyond scaffolding to a variety of other factors that may impact the reported efficacy of PBL interventions. Walker and Leary (2009) conducted an expansive review of 82 studies in K-12 and higher education and found differential effects based on several factors: effects varied by discipline or subject matter, by the type of problem used, and by the type of assessment used. PBL was especially successful in teacher education, medical and health education, and social sciences, but not in engineering or science. Design-type problems and diagnosis-type problems showed positive effects, but dilemma-type problems showed negative effects. Lastly, assessments of principles and applications showed positive effects for PBL, but assessments of concepts (declarative knowledge) did not. The lack of impact on declarative knowledge partially agrees with the earlier finding of negative impact on knowledge acquisition (Dochy et al., 2003). With a focus specifically on assessment, Gijbels, Dochy, Van den Bossche and Segers (2005) examined 40 PBL studies in higher education for differential effects by assessment levels. They categorized assessments as one of three types: 1)

concepts, defined as declarative knowledge of facts; 2) principles, which are the relationships among concepts; or 3) applications, meaning the use of concepts and principles, including in new situations. Their meta-analysis found positive effects of PBL on principle knowledge, but not for concept knowledge or applications, and in fact a tendency toward negative effects of PBL on concept knowledge. To the extent that concept knowledge represents declarative knowledge, these results corroborate other research that has also found null or negative effects of PBL on declarative knowledge (Dochy et al., 2003; Walker & Leary, 2009).

Despite their varied findings, the meta-analyses reviewed above are consistent in one respect: all of them found heterogeneity of effects within their study samples, based on statistically significant Cochran's Q statistics (Dochy et al., 2003; Gibels et al., 2005; Leary et al., 2013; Walker & Leary, 2009). This is clear evidence that the body of extant PBL studies shows mixed results based on such diverse factors as discipline of study, level and type of assessment, problem type, study design, tutor experience and training, and likely numerous other factors that have thus far gone unexamined.

Assessment in Problem-based Learning

In any evaluation of educational interventions, assessment is a crucial consideration. Evaluation, defined as a judgment of the merit of an intervention, is critically concerned with gathering information in order to make the judgment (Fitzpatrick, Sanders, & Worthen, 2004). The gathering of information about students' knowledge, behaviors, and attitudes takes the form of assessment (Airasian & Russell, 2007). Assessment may take the form of qualitative information such as observations and interviews, or it may take the form of quantitative measurement, defined as the numerical

representation of assessment information (Airasian & Russell, 2007). Because this review of literature is primarily concerned with forms of measurement, the discussion of assessments is predominantly limited to assessments that can be quantified and used as measurements; this is not to discount the role of qualitative assessments.

Forms of assessment play an integral role in PBL. One of the tenants of PBL is that students engage in ongoing self-assessment, peer assessment, and group assessment (Hung et al., 2008). Unlike a traditional classroom format, where teachers assess individual students, often using artificial forms such as tests or questioning that are meant to elicit knowledge at specific time points (Airasian & Russell, 2007), PBL is designed to foster students' continuous, interactive, and personalized assessment of their own learning and their progress toward a goal at the individual and group levels (Hung et al., 2008). However, when the time comes to evaluate PBL effects on student outcomes, researchers must use individual assessments that can be summarized, compared, and communicated to a broader audience. Although many published studies include rich descriptions of students' individual and group work products, interviews with students, and classroom observations (e.g., Drake & Long, 2009; Hmelo et al., 2000; Tarhan, Ayar-Kayali, Urek, & Acar, 2008), this qualitative data is not easily transformed to a conclusion about the efficacy of PBL. There may be a fundamental mismatch between the goals and integral assessment forms in PBL and the conventions of educational efficacy research (e.g., WWC, 2014).

Hmelo-Silver (2004) addressed this apparent mismatch in describing a study that found positive knowledge outcomes for PBL students, "provided that the assessments measure knowledge in problem-solving contexts rather than in the context of multiple-

choice examinations." (p. 250). The concern about assessment is well founded, but problem-solving and multiple-choice contexts are not necessarily a dichotomy. Belland, French, and Ertmer (2009) implicitly challenged the conflation of certain types of knowledge and skills with certain test forms by categorizing assessments by the knowledge or skills they intended to assess, not the form (multiple-choice, written essay, etc.) they may have taken; for example, Chang (2001) used multiple-choice items to assess application by requiring students to respond to a novel scenario. With this in mind, the following review of assessments used in PBL studies is organized by assessment form because of the measurement characteristics of each form (the objective assessments reviewed provide dichotomous data; the rater-scored assessments reviewed provide polytomous data2), not because the forms necessarily assess particular types of knowledge or skills. In the case of attitudes, the form of assessment (often Likert-type rating scales) and what is being assessed (affective outcomes) are closely linked; Likert-type rating scales are used to assess self-described affective outcomes or personal characteristics exclusively (Likert, 1974).3 It is established that different forms of assessments play a role in mixed PBL effects (Dochy et al., 2003; Gijbels et al., 2005; Walker & Leary, 2009). However, there is no consensus or established practice as to what type(s) of assessment PBL researchers should use. PBL studies use a wide variety of assessments and measures depending on the goals of their research.

Objective assessments. Objective assessments refer to assessments with correct and incorrect answers agreed upon by subject matter experts, very commonly selection-

[2] This alignment between assessment types and data characteristics is not necessarily always true, but it is true for the assessments described in this literature review.

[3] The converse is not true: affective outcomes may be assessed qualitatively or with methods other than Likert-type scales, such as the "Draw a Scientist" task used to elicit attitudes about science (Drake & Long, 2009).

type items such as multiple-choice and true/false (Airasian & Russell, 2007). Objective achievement tests are a common type of measurement for effects of an intervention, as they are easily aligned to curricular goals or state standards, are easily scored, and their scores are readily used in parametric significance tests. In many cases, items already exist in published tests or textbooks. Several PBL studies have used objective assessments closely aligned to the curricular goals of the intervention. An example is Akınoğlu and Tandoğan (2007), in which 50 middle-school students in Turkey were randomly assigned to either PBL or normal lecture-based curriculum for a ten-week unit about force and energy. The primary pre and post assessments were made up of 25 closed-ended items from Turkey's national curriculum, to which the PBL intervention and the control curriculum were aligned. The PBL group performed higher than the control group in the post-assessment ($p < .05$), providing a counter-example to the overall findings of Walker and Leary (2009), which suggested negative effects of PBL on knowledge acquisition. The close alignment between the national curriculum, the PBL intervention, the control condition, and the objective test make the positive results from this study seem very promising in terms of supporting the efficacy of PBL.

Another PBL intervention with an objective assessment that was closely aligned to curricular objectives was carried out by Drake and Long (2009) in a fourth-grade science classroom. Two classes were cluster randomized, one to receive PBL instruction and one to receive traditional instruction in a unit about electricity, both administered by a university professor. The curricula in both the PBL and control classrooms were closely aligned to state science objectives. The content assessment administered before and after the intervention consisted of 16 objective items (multiple-choice, true/false, matching,

and labeling) also aligned to the state science objectives and reflecting declarative knowledge. The PBL group had a significantly higher gain score than the control group from pre to post ($p < .05$); however, in a delayed post-test four months later, the groups were not significantly different. This lack of retention stands in contrast to the finding in Dochy, et al. (2003) that PBL students gained less knowledge but retained more of it.

A pair of studies that may shed some light on the way objective assessments influence the reported outcomes of PBL interventions are Mergendoller, Maxwell, and Bellisimo (2000) and (2006). Both studies were approximately two-week PBL units for twelfth-grade economics classes, which were cluster randomized to participate in PBL or traditional economics curriculum. In the first study (Mergendoller et al., 2000), three objective assessments were used: tests of specific content knowledge for each of two PBL units, one on fiscal policy and one on supply and demand, as well as a test of general economics knowledge. Items on all three tests were gleaned from a nationally normed economics test and a high school economics textbook. The PBL and control groups did not have significantly different post-test scores for either of the two specific content tests; the control group students outperformed the PBL students on the general economics knowledge test. The second study (Mergendoller et al., 2006) used a different and larger sample of twelfth-graders and one of the two PBL units used in Mergendoller et al. (2000). Only one objective test was administered: a test of specific content knowledge for the PBL unit, which was similar but not identical to the one used in Mergendoller et al. (2000). Mergendoller et al. (2006) showed that PBL students had greater unit-specific knowledge gains than students in the control group. Based on the published studies, it is difficult to know all the ways in which these two iterations of the

same PBL intervention may have differed. For instance, teacher effects may also play a role in the efficacy of PBL: Mergendoller et al. (2006) describes differential results by teacher, in which PBL works better for some teachers, while traditional instruction works better for others. Although it is impossible to rule out other explanations such as teacher effects, it is interesting that the revised unit content knowledge test seems to have detected a positive effect of PBL that was not detected in an earlier study. It is worth noting that the first iteration of the fiscal policy content assessment had middling reliability (alpha = .71) (Mergendoller et al., 2000), and reliability was not reported for the revised instrument (Mergendoller et al., 2006). It may be that a revised instrument with higher reliability was more sensitive to the change in PBL students relative to the control group.

Reliability is a measure of an instrument's consistency, which can be conceptualized over time (test-retest reliability) or, more relevent to the current discussion, internally (Cronbach's alpha, KR-20, KR-21) (DeVellis, 2003). Generally, reliability levels between .65 and .70 are considered "minimally acceptable" and levels from .70 to .80 "respectable"; these instruments are reasonably useful for measuring variables for research purposes and group-level analyses (DeVellis, 2003). Reliability between .80 and .90 is considered "very good" in the context of research purposes and group-level analyses (DeVellis, 2003). Even if an instrument has effective items, some outside factors may cause reliability to be lower than desired, such as group heterogeneity, the situation in which the instrument is administered (e.g. a room that is hot, cold, or noisy may be distracting), and characteristics of the individuals responding, including how well they understand the items based on their age, reading level, and

facility with English (Cohen & Spenciner, 2007). Therefore, a study that finds "low" reliability may not be an indictment of an instrument, but could be due to factors outside the knowledge or control of researchers.

Furthermore, although reliability may have played a role in the different effects in the two Mergendoller et al. studies, a low level of reliability does not necessarily mask intervention effects in all situations. At least two studies have found significant intervention effects using instruments with alphas of .70 and below. A PBL intervention dealing with human biology used a curriculum-aligned multiple-choice test of 25 items developed by researchers, with an alpha of .70, and found that PBL students scored higher than control students in the post assessment (Sunger, Tekkaya, & Geban, 2006). Another example used a research design with three time points (pre, post, and four weeks later) and a knowledge test with alphas of .63 (pre) to .66 (post), which found that the PBL group scored higher than the control group at post and also had retained more knowledge than the control group after four weeks (Son & VanSickle, 1993). The results of Sunger, Tekkaya, and Geban (2006) and Son and VanSickle (1993) demonstrate that low reliability instruments can and do find significant intervention effects, and conversely, high reliability instruments do not guarantee a flawless assessment or a finding of significant effects. Chang (2001) used assessments of knowledge, comprehension, and application with high reliability (KR21 = .77 to .81) and found favorable effects of PBL for knowledge and comprehension, but not for application. This is unexpected, given other findings that PBL students often do well with applications (Walker & Leary, 2009). The author speculates that the application items may have been too few (seven in total) and too difficult (Chang, 2001). Indeed, items that are

inappropriately easy or difficult could mask significant effects, even when reliability is high. Taken together, Sunger, Tekkaya, and Geban (2006), Son and VanSickle (1993), and Chang (2001) show that there is no simple relationship between assessment reliability and efficacy findings for PBL interventions.

Subjective assessments. Subjective assessments are those that require some amount of human judgment (subjectivity) in their scoring (Airasian & Russell, 2007). The most common types allow students to supply their own answers to open-ended questions or prompts; examples range from closely restricted short-answer items to longer essays and projects that allow students creative freedom. Given the emphasis of PBL on process (e.g., Hung et al., 2008), some researchers have turned to open-ended, subjective assessments as a supplement to objective assessments.

Hmelo et al. (2000) carried out a PBL design intervention in which sixth graders learned about the human respiratory system by designing an artificial lung; a control group learned by traditional lecture methods. Besides an objective true/false quiz on respiration, students were assessed using their drawings of the respiratory system on the outline of a body. Drawings were assessed with a holistic rubric of five levels of advancing model sophistication, where the lowest level was simply a hollow structure in the chest and increasingly complex drawings showed the connections between the lungs, blood vessels, and brain. Both the PBL and control students showed increases in the sophistication of their models from pre to post, with PBL students more often scoring at the highest levels of sophistication. The shifts were interpreted as a significant improvement for PBL students, more so than for control students (Hmelo et al., 2000). In accordance with the ordinal nature of the five-level holistic scoring rubric, non-

parametric tests were used to test significance. The results appear promising, but they are a bit more difficult to interpret than numeric gain scores (evaluated using parametric tests) would be. For instance, the authors report that the control group scored most frequently at levels 2 and 3 and only one student scored a 5; for the PBL group, levels 2 and 3 were also the most frequent scores, but four PBL students scored a 5 (Hmelo et al., 2000). Relative to the sample size ($N = 42$) this is a substantial improvement; however, it is not as readily communicated as a mean score for each group would be.

In order to measure long-term comprehension and application following PBL units in social studies, Wirkala and Kuhn (2011) employed a posttest-only design carried out nine weeks after the units, administered to both PBL and control groups. The assessments were purely open-ended and scored by raters. In the comprehension assessment, students were asked to define seven concepts from the units; definitions were scored dichotomously for correctness and then on a five-point scale for the depth of the definition. In the application assessments, students read a scenario that required a new application of the concepts they learned in the PBL unit. Their written responses were coded for the use of seven concepts; concepts were scored dichotomously for presence in their written application response and then on a five-point scale for the depth of explanation. Thus the two assessments had parallel scoring methods but elicited PBL concepts in two different contexts, one being cued recall and one being a novel application. For the ordinal levels of explanation for concepts, a chi-square test was used, which is appropriate for the ordinal data, but requires a bit of work in interpretation. For instance, the control and PBL groups had approximately equal numbers of students who only reached level 2 of explanation, but the control group was overrepresented in the

bottom levels, and the PBL group was overrepresented in the top levels (Wirkala & Kuhn, 2011). Additionally, researchers collapsed the original five levels of scoring into four levels to reach required cell sizes for the chi-square analysis, which may have threatened the precision of the scoring method.

Hmelo et al. (2000) and Wirkala and Kuhn (2011) both used rubrics oriented to detect increasing levels of sophistication. In a somewhat opposite method, Chang and Barufaldi (1999) used a rubric to detect increasing levels of misconceptions about mountain formation. In a pre-post design, researchers asked students to explain how mountains form and then scored their responses on a five-level rubric, ordered from correct factual understanding to having pervasive misconceptions. For each time point (pre and post), the numbers of students at each level of understanding were compared across the two groups (PBL and control) using a chi-square test. The two groups were not significantly different at the pre time point and were significantly different at the post time point. However, it requires close examination of the distribution of scores to determine that the PBL group had fewer misconceptions after the intervention. For instance, in the post-test, the PBL and control groups had similar numbers of students who scored at the lowest level (*n*'s = 24 and 27, respectively), but the PBL group had more students score at the highest level (*n* = 18), compared to the control group (*n* = 3) (Chang & Barufaldi, 1999). Although the difference is apparent and the significance is established by the chi-square test, the reader must look to the distribution across cells to identify the pattern of change.

Hmelo et al. (2000), Wirkala and Kuhn (2011) and Chang and Barufaldi (1999) all used appropriate statistics for their ordinal scoring systems, and all presented results

that are convincingly in favor of PBL; however, because their subjective, open-ended assessments produced no single scores to compare directly across groups, they can be difficult to interpret. Further, readers know very little about the consistency of the rubrics used for scoring and the raters who performed the scoring. All studies published descriptions of the levels used for scoring. However, Hmelo et al. (2000) did not publish any reliability data. Wirkala and Kuhn (2011) published percent agreement (86% to 89%) and Cohen's kappa (.70 to .87) between two raters. Chang and Barufaldi (1999) published the correlation (r = .93) among raters. Although these are generally considered acceptable levels of reliability, the levels of correspondence and agreement do not provide much information about potential systematic differences between raters; for instance, whether one rater was consistently more severe than the other or if a rater tended more toward middle-level ratings (Myford & Wolfe, 2004).

As was the case with objective measures, subjective measures have found mixed results for PBL. Following a civil rights PBL unit, Saye and Brush (1999) had students write a structured persuasive essay. The essay instructions guided students to use particular forms of knowledge in each paragraph (i.e. summarize in paragraph 1, argue each side in paragraphs 2 and 3, and support one side in paragraph 4). Each paragraph was scored separately based on the number of factual statements (paragraph 1), persuasive arguments (paragraphs 2 and 3), and for dialectical reasoning (in paragraph 4). Results showed that the PBL group produced more factual statements and had better dialectical reasoning, but there was no difference between groups in their general reasoning. The mix of significant and non-significant results within the components of a single essay reveal the importance of careful alignment between the skills being assessed

and the task students are asked to do. Other studies found inconsistencies within their own findings by assessment type: Akınoğlu and Tandoğan (2007) found significant differences between PBL and control groups in an objective achievement test, but no significant differences by group in open-ended questions.

Attitude measures. Besides the academic outcomes of knowledge and skills, many researchers are interested in how PBL affects students' attitudes. Attitudes such as interest and self-efficacy are components of intrinsic motivation (Deci, Vallerand, Pelletier, & Ryan, 1991), which is one of several PBL learning goals outlined by Hmelo-Silver (2004). Attitudes were considered important in the earliest meta-analyses of PBL in medical school programs, which found that PBL students enjoyed their programs more than students in traditional programs (Alabanese & Mitchell, 1993; Vernon & Blake, 1993), and attitudes continue to be measured and reported in many PBL studies. The inconsistency found in academic outcomes, using both objective and subjective assessments, is also present in measures of attitudes. For the purposes of making the measurement implications clear, only those attitude measures that used Likert-type rating scales are reviewed.

Hernandez-Ramos and De La Paz (2009) used several attitude measures before and after a middle-school United States history unit and found that both the control and PBL groups had increases in perceived knowledge and testing self-efficacy, with no significant differences between the groups in their growth patterns. In attitude toward social learning, the control group declined slightly and the PBL group increased; the same pattern was evident in attitude toward social studies. Repeated-measures ANOVAs show that the time-by-treatment effects are significant (social learning $p = .039$; social

studies $p = .030$) such that the two groups have different growth patterns. This pattern can be seen in other findings about subject-area interest as well: Akınoğlu and Tandoğan (2007) reported declining levels of interest in science for both PBL and control groups, but the PBL group's interest declined less. These results are contrary to what is generally seen in academic outcomes, where both PBL and control groups have learning gains, but the gains for one group may be larger. In the case of attitudes, it seems that the most common trend may be for interest to decline or remain flat over time, and an effective intervention reduces the decline or causes a slight increase in interest.

In another example of mixed results within a study, Sunger and Tekkaya (2006) reported several subscales from the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich, Smith, Garcia, & McKeachie, 1991). PBL had favorable effects, compared to a control group, on the subscales for intrinsic goal orientation and task value subscales, but there were no group differences on subscales for extrinsic goal orientation, control of learning beliefs, self-efficacy, or test anxiety (Sunger & Tekkaya, 2006). Although the effects on intrinsic goal orientation and task value are promising, it is still the case in this study that null effects were prevalent among the motivation variables.

One study that hints at a shortcoming of even widely accepted attitude scales is Liu, Hsieh, Cho, and Schallert (2006). Researchers conducted a PBL study with a pre-post design and no control group. Because of the lack of control group, their findings say little about the relative efficacy of PBL, compared to other instructional techniques, but their results for attitude toward science and science self-efficacy suggest one potential reason for null effects in attitude measures. Science self-efficacy was measured using eight items adapted from the MSLQ, and attitude toward science was measured using 14

items from the Attitude Toward Science in School Assessment (ATSSA; Germann, 1988). These instruments both require students to respond using rating scales of 1 to 5, where 5 indicates more self-efficacy or a more positive attitude. Both of these published scales are widely used and boast high reliability (alphas > .90). In this study, students were found to have higher self-efficacy following the PBL unit ($p < .001$); however, there was no change in attitude toward science (Liu et al., 2006). The authors note that "attitude scores were all above the mid-point of the scale" (p. 234) even at the pre time point, and they suggest that a ceiling effect for attitude scores may have played a role in the lack of significant findings (Liu et al., 2006).

The issue of a potential ceiling effect in attitude scores (e.g., Liu et al., 2006) is somewhat similar to the issue of overly difficult items as discussed in Chang (2001). The ceiling effect suggests that the items used to assess attitude toward science were too *easy* for the students in the study; in this case, it was too easy for students to agree or agree strongly with many of the statements about science. It is possible that students did like science even more following the PBL unit; however, if many of them already strongly agreed with statements such as "science is fun," there is no way for them to agree even more strongly after the PBL unit. Increasing the number of scale points (e.g., from 1-5 to 1-7 or 1-10) does not necessarily address the problem, as students might simply choose 7's or 10's where they had chosen 5's. Writing more difficult items such as "I like to read about science in my free time" or "a science experiment is one of my favorite things to do" would be a potential solution, as would considering that items on the scale have a range of difficulty and giving more weight to agreement with more difficult items.

Summary of Assessment Issues in Problem-based Learning Interventions

The foregoing review has shown that studies of PBL interventions in K-12 contexts provide mixed results, as has been found by numerous meta-analyses of PBL programs in K-12 and higher education. The mixed results are present in academic outcomes measured by objective assessments, academic outcomes measured by subjective, rater-scored assessments, and attitude outcomes measured by response scales.

At least one research synthesis has pointed out assessment issues prevalent in PBL research. Belland et al. (2009) reviewed 33 PBL studies for the type of outcome each study measured (deep content learning, problem solving, or self-directed learning) and the type of assessment used (multiple-choice, essay, clinical judgment, and many others). They concluded that "few studies included 1) theoretical frameworks for the assessed variables and constructs, 2) rationales for how chosen assessments matched the constructs measured, or 3) other information required for readers to assess the validity of authors' interpretations" (Belland et al., 2009, p. 59). Importantly, the authors also point out that many of the PBL studies they reviewed would fall short of WWC "meets criteria" standards based on the measures reported (Belland et al., 2009). Thus, this body of PBL studies could suggest mixed or null effects possibly based on flawed assessments; additionally, a large portion of PBL research would not qualify for review by WWC at all, and therefore would not reach the wider audience of practitioners and policy-makers that WWC strives to inform.

One possible reason for mixed results in PBL interventions that has not been discussed in the PBL literature is the widespread use of observed scores, which are ordinal data, in parametric statistics that require interval data. The following section

explains the distinction between ordinal and interval data, the controversy around the importance of this distinction, and how the use of ordinal data could pose a threat to statistical conclusion validity of single studies and the larger body of educational research.

Ordinal Data in Educational Research

The assumptions of parametric statistics. Statistics textbooks caution readers that there are several assumptions embedded in parametric statistics (e.g., Field, 2005; King & Minium, 2008). Parametric statistics include some of the statistics most commonly used in social science research to compare groups, treatments, and time points, including *t*-tests and analysis of variance (ANOVA). Parametric statistics are so named because they are based on the assumption that the data under study are drawn from a population with a normal distribution, from which inferences can be made about the parameters of the population (King & Minium, 2008). It is rare for data sets to follow a strict normal distribution (Micceri, 1989); however, non-normality of a data set may not indicate non-normality of the population, and sample non-normality is generally not considered a problem for samples larger than $n = 30$ (King & Minium, 2008).

Parametric statistics also assume that populations being compared have similar amounts of variance within each population; when group sample sizes are approximately equal, common parametric statistics tend to be robust to violations of this assumption (e.g., Glass, Peckham, & Sanders, 1972; Markowski & Markowski, 1990). A third assumption of parametric statistics is independence of observations, meaning that groups that are being compared are sufficiently separate such that neither group is able to influence or contaminate the other and especially that no individual is a member of both

groups (Field, 2005). This third assumption is primarily a research design consideration rather than a statistical or measurement issue.

The assumption of interval data. The final assumption and the one germane to the present study is the assumption that data are on an interval or ratio scale of measurement (Stevens, 1946). Stevens (1946) outlined four forms of measurement and their statistical properties. Nominal data are classifications with no numerical correlates and should only be analyzed using their counts; gender and race are common examples. Ordinal data are classifications with numerical order but not precise numerical value and therefore can be divided into ranked groups (highest, lowest, median) but should not be used in parametric statistics. A classic example of ordinal data is rankings in a race. A first-place runner might be only a fraction of a second ahead of a second-place runner, but the third-place runner trails by several seconds; thus, although first, second, and third places have a true order, there are not equal intervals between first and second and between second and third. Interval data, however, have numerical values that represent true intervals, where the numbers assigned represent an amount of something that can be compared to other amounts of the same thing. Temperature in degrees Fahrenheit is an example: the difference between 60 and 70 degrees is the same as the difference between 80 and 90 degrees. Lastly, ratio data are data that, besides having the quality of interval data, also have a meaningful zero point that represents a complete lack of the thing being measured (Stevens, 1946). Mass in grams is an example of ratio data; the difference between 2 and 4 grams is the same as the difference between 7 and 9 grams, and 0 grams means a complete lack of mass. Ratio data is most commonly found in physical weights and measures and is not a requirement for most parametric statistics.

Stevens (1946) made an important distinction between the statistics that are permissible to use with ordinal data and interval data: mean and standard deviation (which are the basis for most well-known parametric statistics) should only be applied to interval data, never to ordinal data. The reason for this is that ordinal data, which has unequal intervals, is subject to distortion when it is subjected to calculations like the mean. For instance, the example of race finishers, it would be wrong to assume that the second place runner finished exactly equidistant between the first and third place runners, which is exactly what the arithmetic mean would suggest. Despite the logical and mathematical soundness of this prohibition, Stevens (1946) provided a "pragmatic sanction" for using ordinal data in statistical calculations when it provides "fruitful results" (p. 679). If ordinal measures and ordinal data are the best that are available to social scientists, Stevens (1946) allows them to proceed in calculating statistics that are inappropriate to the scale of measurement, with the warning that unequal intervals between ordinal data points will introduce error into statistical conclusions.

Since 1946, researchers have continued to debate the appropriateness of the use of ordinal data in parametric statistics. The requirement of interval data has been repeated in statistics textbooks (e.g., Field, 2005; King & Minium, 2008), and there is no shortage of scholarly journal articles that reiterate it and point out the common misuse of ordinal data. The problem with ordinal data can be understood conceptually, as in the examples from rehabilitative medicine provided by Merbitz, Morris, and Grip (1989). In one of their examples, mobility is rated on a scale of 0 to 4, with 0 being bedridden and 4 walking unaided. Moving from using a wheelchair (level 2) to walking with a cane (level 3) is certainly progress, and walking unaided (level 4) represents even more progress.

The ordinal nature of the scale is supported by clinical experience and common sense. However, it would be very controversial to say that advancing from wheelchair use to walking with a cane is *exactly the same amount* of progress as advancing from use of a cane to walking unaided (Merbitz, et al., 1989). The implication that the amounts of progress are equal is the fallacy that occurs when numerical levels of progress are considered to be interval data. The problems are compounded when several ordinal items, each with idiosyncratic properties, are considered together as a scale and/or used to make statistical inferences (Merbitz, et al., 1989).

There is another shortcoming of observed scores, related to their nature as ordinal data. In reply to Merbitz, et al. (1989), Wright and Linacre (1989) made the distinction between observations (which are necessarily ordinal, as discussed above) and measures. A true measure (as opposed to an observation) must measure the same way, regardless of whom it is measuring or which particular measure is being used. Take the standard ruler as an example: a ruler can be used to measure the length of any thing (a pen, a notebook, a smartphone), and a thing can be measured with any ruler. This is the principle of objectivity. Test-free person measurement states that there is a universe of items that define a variable, and within that universe of items, any sample of items should work equally well for measurement. Sample-free item calibration states that the items that comprise a scale must function the same way for any group within the universe of people for whom it is intended (not necessarily to function the same for all humans, but for all American eighth-graders, for example). Using observed scores from tests or rating scales, it is impossible to make sound inferences about the items divorced from the particular

sample of individuals responded to them, and impossible to make sound inferences about the persons divorced from the particular items to which they responded.

Many commentators have raised the alarm regarding the mismatch of data and analysis technique as a threat to the validity of conclusions in education, especially concerning Likert-type rating scales (e.g., Jamieson, 2004). Kuzon, Urbanchek, and McCabe (1996) named the use of ordinal data in parametric statistics "Sin 1" in a list of seven deadly sins of statistical analysis. It is a crucial problem because, if there is a theoretical mismatch between observed scores and parametric statistics, this may constitute a threat to conclusion validity, an instrument's ability to lead to the correct statistical conclusion, i.e. whether an effect is statistically significant or non-significant (Shadish, Cook, & Campbell, 2002).

However, the use of ordinal data is persistent in educational research and has been defended vehemently by those who say parametric statistics are robust to imperfect data (e.g., Norman, 2010). In response to Jamieson (2004), Pell (2005) replied that the original scale of measurement is unimportant as long as the distribution of data meets the assumptions of normality and equality of variance. Carifio and Perla (2008) agreed with Pell's (2005) "intervalist" view of Likert-type data and suggested that even if a single Likert-type item produces ordinal data, a sum score of many of them will produce interval data because the entire scale has "emergent properties" that are different from the properties of the original items.

Notwithstanding the psychometric merits or shortcomings of this debate, the purpose of recounting it is to illuminate that even within the last decade, there are researchers who use ordinal data in parametric statistics and furthermore defend their use

of ordinal data because, put very simply, it appears to work. The many intervention studies that have correctly found statistically significant effects using observed scores seem to provide evidence for the robustness of parametric statistics, and the finding of "fruitful results" (Stevens, 1946) largely serves as its own pass for this statistical transgression. (However, as Merbitz, et al. (1989) point out, "simply because statistical calculations are performed on a set of numbers does not mean that the results of the calculation can be meaningfully interpreted.") It is not surprising that observed scores appear to "work" in parametric statistics: the relationship between observed scores and a corresponding interval measure is monotonic and in the shape of an ogive, and near the center of the ogive, the relationship is nearly linear (Wright & Linacre, 1989). When there are more extreme high or low scores, the observed scores will be more distorted compared to interval measures; when scores are more central, observed scores will be more likely to yield accurate conclusions.

The crucial question for the debaters of observed scores and interval measures is, does it matter? Do ordinal and interval measures of the same construct lead to different conclusions? Wright and Linacre's (1989) explanation suggests that they can and will lead to different conclusions in certain situations. If the answer is yes, even if not in all circumstances, then the use of observed scores poses a threat to statistical conclusion validity (a threat that may be more dire in some situations than in others), and the prevalent use of observed scores may be a source of mixed results in PBL intervention research and in educational intervention research generally.

Creating Interval Measures from Observed Scores

There are numerous ways to handle the issue of inadequate data (such as the use of observed scores) in educational interventions. Researchers could favor qualitative methods or nonparametric statistics; however, WWC standards favor quantitative measures, as do other overarching evaluation methods such as meta-analyses. As was demonstrated in the previous discussion of holistic rubric measures, studies that report changes in categorical scores can demonstrate significance using the chi-square test, but the exact nature of the changes requires some work in interpretation (e.g., Chang & Barufaldi, 1999, Hmelo et al., 2000, Wirkala & Kuhn, 2011). Researchers also could favor alternative measurement techniques such as Bayesian methods, item-response theory (IRT), latent trait theory, etc. There are several techniques for rescaling ordinal data to interval data, including logarithmic linear transformation (Field, 2005), IRT (e.g., Harwell & Gatti, 2001), multidimensional scaling (Kruskal, 1964), and Markov chain Monte Carlo modeling (Granberg-Rademacker, 2009).

Rather than attempt to address all possible research methods and measurement models, this literature review is grounded in quantitative measurement and parametric statistics in particular, as are commonly used to evaluate educational interventions. The following section addresses Classical Test Theory, as the most common measurement model for educational research, and one alternative, Rasch modeling, as a solution to the problem of using observed scores in educational intervention efficacy studies. Rasch modeling (Rasch, 1960/1980; Wright & Stone, 1979) is a method that produces interval measures of persons that are separable from the particular items to which they responded. Integrally, Rasch modeling also produces interval measures of items separable from the

persons who responded to them. Therefore, the Rasch model produces interval data that is ideally suited for use in parametric statistics.

Classical Test Theory

When several items on a response scale or achievement test are summed for a total score, as is typically done in classroom assessment (Airasian & Russell, 2007) and as prescribed by Carifio and Perla (2008), the total scores are observed scores, also known as raw scores or simply "scores." The measurement model underlying observed scores is Classical Test Theory (CTT). In CTT, numeric scores are assumed to have a meaningful numeric value, and points are summed to create a total observed score that includes a true score plus measurement error (Crocker & Algina, 1986).

$$X = T + E \tag{1}$$

where observed scores (X) theoretically include a true score (T) plus some amount of measurement error (E).

This additive model has implications for data from objective tests, affective scales, and rater-scored assessments. In the case of objective tests where items are scored dichotomously, it implies that all items have the same difficulty and answering any one item correctly is the same as answering any other item correctly. For example, the difference between a score of 11 and 12 should be the same as the difference between 15 and 16. It is easy to see how this may not be a safe assumption: if the sixteenth item is much more difficult than the twelfth item in the example above, then moving from 15 to 16 is more difficult than moving from 11 to 12, and the intervals between the numeric scores are *not* equal. Most objective tests have items within a range of difficulty. For

example, multiple-choice item stems such as “what is freshwater?” and “what is desalinization?” might both be appropriate for a water unit test, but it is very likely that the second one is more difficult to answer correctly. In the case of affective scales, where responses often take the form of a number on a scale of 1 to 5 or similar, the CTT model implies 1) that all items are of equal difficulty and 2) that the intervals between responses are equal, e.g., that the difference between “1 = strongly disagree” and “2 = disagree” is the same as the difference between “4 = agree” and “5 = strongly agree.” Lastly, in the case of rater-scored assessments, the CTT model implies 1) that all rubric criteria are of equal difficulty, 2) that intervals between rating scale levels are equal and 3) that all raters are applying the rubric in the same way.

The Rasch Model

The dichotomous Rasch model. Rasch models are able to produce estimates of person ability and item difficulty on an interval scale when data fit model requirements (Wright & Stone, 1979). The simplest form, shown in Equation 2, is for dichotomous data. The Rasch model states that the probability of a person responding correctly to an item is a logarithmic function of the person ability and the item difficulty (Rasch 1960/1980). Whenever the estimated person ability is greater than the estimated item difficulty, it is likely that the person will respond correctly, and the likelihood of a correct response increases as the person ability increases relative to the item difficulty. Mathematically, the Rasch model as shown in Equation 2 is identical to the one-parameter IRT model (Crocker & Algina, 1986), and many researchers refer to Rasch and one-parameter IRT as one and the same (e.g., Embretson, 1996; Kohli, Koran, & Henn,

2014). However, many adherents of Rasch consider it to be philosophically different from IRT despite the superficial resemblance (Andrich, 2004).

$$\Pr\{X_{ni} = x\} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)} \quad (2)$$

where the probability (Pr) of person *n* answering item *i* correctly (i.e. score = 1) is a function of the person ability (β_n) and the item difficulty (δ_i).

The person ability and item difficulty may be compared directly because they are represented on a common logit scale. Logits are a representation of probability where values above zero indicate probability more than 50%. Higher logit values for person ability represent more ability (or more of a construct such as interest in a subject), meaning a greater likelihood of responding correctly to more difficult items. Higher logit values for item difficulty represent more difficult items.

The polytomous Rasch models. The model shown in Equation 2 applies when each item can have one of two values, 0 for incorrect or 1 for correct. When items have a range of possible values, as in the case of a rating scale (in which responses can range, for example, from 1 to 5), an additional parameter is added to the model that represents the threshold between two responses. Like person ability and item difficulty, threshold values are represented in logits, where a higher logit value means that it is more difficult to move from one response to the next (e.g., from 4, agree to 5, strongly agree).

$$\Pr\{X_{ni} = x\} = \frac{\exp \sum_{k=0}^{x} (\beta_n - \delta_i - \tau_k)}{\sum_{h=0}^{K} \exp \sum_{k=0}^{h} (\beta_n - \delta_i - \tau_k)} \quad (3)$$

where the probability of person *n* getting a score of *x* on item *i* is a function of the person ability, the item difficulty and the threshold (τ_k) between adjacent scores.

The model shown in Equation 3 is known as the rating scale model (Andrich, 1978). Although the thresholds may vary between levels on the rating scale (e.g., levels 2 and 3 could be closer together than levels 4 and 5), the same threshold structure is modeled for each item. The partial credit model shown in Equation 4 allows for each item to have a different threshold structure (Masters, 1982). The partial credit model is useful with polytomous items that do not necessarily share a rating scale structure. Examples are multiple choice items where there is one definitely correct response and other responses that represent varying amounts of partial knowledge, as well as rater-scored items with levels of partial credit that may vary across items.

$$\Pr\{X_{ni} = x\} = \frac{\exp \sum_{k=0}^{x} (\beta_n - \tau_{ki})}{\sum_{h=0}^{K_i} \exp \sum_{k=0}^{h} (\beta_n - \tau_{ki})} \quad (4)$$

where the probability of person *n* getting a score of *x* on item *i* is a function of the person ability and the threshold (τ_k) between adjacent scores on the item. Because the threshold is calculated for each category and each item, the threshold and difficulty are one and the same in the partial credit model.

Many-facet Rasch measurement. In the example above of the partial credit model used with rater-scored items, it is assumed that rating scales across items might

vary; for instance, if scoring levels are “no credit,” “partial credit,” and “full credit,” partial credit might be closer to no credit for some items and closer to full credit for other items. This depends upon how the rubric is written for each item and also upon how the raters interpret the rubric and items. However, the partial credit model does not take into account the way different raters might behave. In many-facet Rasch measurement (MFRM), the severity of each individual rater is considered along with person ability, trait difficulty (analogous to item difficulty), scale thresholds, and rater severity (Linacre, 1989). As shown in Equation 5, rater severity is represented in logits, where a higher value indicates a rater more likely to give lower scores (a more “severe” rater). This model can be expanded to additionally examine interactions among raters and persons, traits, and/or categories; for instance, to allow category structures to vary by trait as does the partial credit model.

$$\Pr\{X_{nikj} = x\} = \frac{\exp \sum_{k=0}^{x} (\beta_n - \delta_i - \tau_k - \alpha_j)}{\sum_{h=0}^{K_i} \exp \sum_{k=0}^{h} (\beta_n - \delta_i - \tau_k - \alpha_j)} \tag{5}$$

where the probability of person *n* getting a rating of *x* on trait *i* from rater *j* is a function of the person ability, the trait difficulty, the threshold between adjacent ratings, and the severity of the rater (α_j).

For each of the above models, software such as Winsteps (Linacre, 2015b) or Facets (Linacre, 2015a) works through iterations to find the best estimates of ability for each person, difficulty for each item, threshold for each category, and severity for each rater in the data set to fit the pre-determined model; all estimates are expressed in logits. The estimates and corresponding fit statistics for items, thresholds, and raters provide rich

information about a test or scale. Estimates of ability for each person are on a true interval scale where the difference between scores of -2.0 and -1.0 is the same as the difference between scores of 0 and 1.0 if the data fit model requirements. These true interval measures are ideally suited for use in parametric statistics; the measures can be transferred to a statistics package such as SPSS (IBM, 2015) and used for determining pre-post effects, between-groups differences, correlations, etc.

Empirical Comparisons between CTT and Rasch Models

The theoretical advantages of Rasch models; namely, the creation of true interval scales, are well documented in Rasch textbooks and handbooks (e.g., Bond & Fox, 2007; Smith & Smith, 2004; Wright & Stone, 1979). The present literature review focuses on studies that have attempted to demonstrate a practical advantage (or disadvantage) to using Rasch models. These fall into three large categories: 1) studies that describe the practical advantages of using the Rasch model for diagnostic purposes in instrument development; 2) studies that execute Rasch and CTT methodologies on the same instruments with the goal of establishing one or the other method empirically superior for instrument development; 3) studies that assess the conclusion validity of Rasch and CTT methods by comparing the use of Rasch and observed scores in parametric statistics.

Advantages of the Rasch model in scale development. A large body of literature demonstrates the practical advantages of using the Rasch model for diagnostic purposes in scale development in a variety of settings including medical/clinical and educational. In the case of dichotomous scales, Rasch modeling aids scale development by providing difficulty information for each item, and because items and persons are measured on the same scale (logits), they can be compared directly to determine how

well suited items and persons are. That is, are the items difficult enough, but not too difficult, to accurately measure the population in question? Boone and Scantlebury (2006) describe the benefits of using Rasch analysis for a dichotomous (multiple-choice) state science assessment. Besides determining whether items and persons are a good match, the Rasch model provides model fit statistics for items and information about differential item functioning that may reveal a race, gender, or other bias in a test.

The benefits provided by the dichotomous Rasch model hold for the more complex models, which also provide additional information about the features of an instrument. The rating scale model provides information on how rating scales are being used. Smith, Wakely, de Kruif, and Swartz (2003) demonstrated this with a writing self-efficacy scale taken by fourth- and fifth-grade students. The scale originally had 10 points along a 100-point scale labeled, e.g., 10 (not sure) to 100 (really sure). Their rating scale analysis found that although all the categories generally reflected a trend where higher numbers reflected more self-efficacy, there were several categories that were never the most likely response for a student of any ability level. Thus, the categories were collapsed into a 4-point rating scale that better reflected the capability of fourth- and fifth-graders to delineate their own confidence in carrying out writing tasks.

The partial credit model provides information about the use of rating scales that is flexible for each item. This allows for examination of the rating structure for each item individually; particularly with items scored by raters, rating scales might appear homogenous but function differently for different items. Eggert and Bögeholz (2010) demonstrated this in constructing and evaluating a socio-scientific decision-making instrument. Some items that had the same *a priori* rating structure (e.g., possible scores of

0, 1, 2) did not all support the same rating structure; that is, some supported a 3-point scale while some only supported a 2-point scale. In the latter cases, threshold information revealed that there was no meaningful difference between two categories for a certain item.

The many-facet Rasch model provides additional information about how raters interact with traits (analogous to items), rating categories, and persons. An evaluation of a state writing assessment found that, despite extensive rater training, raters showed significant differences in their severity, as well as overly predictable patterns for some raters and overly erratic rating patterns for others (Englehard, 1992). The Rasch model can also be used to identify misfitting persons, as Englehard (1992) demonstrates; researchers can then inspect misfitting persons' responses for unusual elements such as a pattern of responses in the case of a test or rating scale, or illegible handwriting or some other distracting factor in the case of rater-scored responses.

All of the above examples are to demonstrate that Rasch methodology provides rich information about item difficulty relative to person ability, rating scale structure, thresholds between scored levels, and rater behavior. Rasch includes several statistics that have correlates in CTT, such as scale reliability and item difficulty; however, Rasch also provides several statistics for which there is no CTT equivalent, such as item and person fit statistics and rating scale thresholds. Taken together, studies such as those cited above have led many researchers to conclude that Rasch analysis is more informative than CTT for instrument development and refinement.

Empirical comparisons between Rasch and CTT methods. A related line of research has set out to establish an empirical advantage for one methodology over the

other by comparing the two in some way, but no clear conclusion has been achieved. When used for scale reduction, Rasch and observed scores often produce visibly different scales, but if both scales are closely related to an external criterion (such as the original scale), there is no way to determine using real data that one method produced a better reduced scale than the other. Prieto, Alonso, and Lamarca (2003) used Rasch and CTT in parallel to reduce the length of a dichotomous, health-related quality of life scale. CTT using exploratory factor analysis identified four factors with a total of 22 items; Rasch identified two factors with a total of 20 items; 12 items were the same across the two scales. Although the two scales looked fairly different, both reduced scales correlated highly with the original 38-item scale (r = .97 for both) and with each other (r = .95). A similar study used CTT and Rasch in parallel to reduce the length of a health-related rating scale instrument (Nijsten, Unaezeà, & Stern, 2006). CTT using exploratory factor analysis identified three factors with a total of 10 items; Rasch identified a single factor with 11 items; six items were the same across the two scales. The two scales both correlated highly with the original scale (r = .94 for CTT; r = .96 for Rasch) and with each other (r = .87).

A comparison of methods in a children's healthcare scale compared CTT analysis (maximizing Cronbach alpha) with the Rasch Partial Credit Model (using item fit statistics) to determine which scale items to retain (Erhart et al., 2010). The two analyses yielded somewhat different item sets: 13 items in the CTT analysis and 11 items in the Rasch analysis, with nine items in common between the two scales and highly correlated scores (r = .93). When correlations are used to compare item and person statistics for identical Rasch and CTT instruments, the similarities seem even greater. Fan (1998)

found highly correlated item difficulty values (r = .95 to .99 for various sub-groups of persons) and also highly correlated person measures (r = .98 to 1.00 for various sub-groups). Similar results were replicated in a simulated data study (item and person rs > .98) (MacDonald & Paunonen, 2002).

Although some researchers express a preference for Rasch scales (e.g., Nijsten, Unaezeà & Stern, 2006), none of these studies were able to establish a clear empirical benefit, in the sense that both Rasch and CTT scales showed good reliability and validity characteristics in their own paradigms (Rasch scales met Rasch criteria; CTT scales met CTT criteria), and Rasch and CTT scales were both closely related to the original scale from which they were derived.

There are many studies that directly compare CTT and Rasch for analysis of dichotomous and rating scale items, including some that use rater scoring of individual items (e.g., Lynn & Lawless, 2015); these tend to report reliability statistics for raters and then do not take rater effects into further account (e.g., Eggert & Bögeholz, 2010). In the case of performance assessments (such as a written essay) that are scored by raters, many studies use observed scores without any correction for rater effects (e.g., Brown & Lawless, 2014). In a review of writing assessments, Huot (1990) describes various scoring methods such as holistic scoring, analytic scoring, and weighted category scores, all of which are various types of additive observed scores.

Studies using MFRM, however, are definitely and often singularly concerned with rater effects. As Rasch dichotomous, rating scale, and partial credit models are often used primarily for the development of instruments, MFRM is often used primarily for assessing rater effects in order to modify rubrics, rater training, or scoring procedures

(e.g., Myford & Wolfe, 2004). Therefore, few researchers have undertaken direct comparisons between CTT, which does not provide information on rater effects, and MFRM.

There have been several studies comparing MFRM and generalizability theory (GT). GT is an extension of CTT in which the variance in scores is divided into proportions, such as a proportion of variance due to person, proportion due to different tasks, and proportion due to rater effect, as well as interactions among these factors (Cronbach, Nageswari, & Gleser, 1963). Thus, like MFRM, GT seeks to identify sources of variance and control for them in providing a true score for each individual. However, like CTT observed scores, GT scores are still ordinal data, whereas MFRM provide corrected scores that are interval data when the data fit model requirements (Sudweeks, Reeve, & Bradshaw, 2004). One study that directly compared CTT and MFRM found, not surprisingly, that MFRM provides much more information about sources of variance, making it superior for assessments of rater effects (Haiyang, 2010). A CTT analysis of reliability (Cronbach's alpha) suggested poor reliability, but only with more detailed MFRM analysis could the specific sources of error be located.

Sudweeks, Reeve, and Bradshaw (2004) compared GT and MFRM for detecting effects of raters and different rating occasions in a large university writing assessment. The two analyses agreed on the relative contributions of the sources of variance; neither appeared to show a clear advantage, and MFRM logit measures and G scores had a strong linear relationship (r = .99) (Sudweeks, Reeve, & Bradshaw, 2004). A study of a smaller group of fourth-grade students also found GT and MFRM analyses to agree on the relative contributions of sources of variance (Smith & Kulikowich, 2004). MacMillan

(2000) also found that GT and MFRM agreed on the amount of variance attributable to raters (although MFRM detected more discrepant raters). MFRM scores and linear-scale corrected scores were highly correlated ($r = 1.00$).

The forgoing studies have all compared Rasch and CTT (or GT) with the goal of establishing that one or the other has as clear advantage. If neither has an advantage, then choosing a measurement model may be an issue of appropriateness for a particular purpose (e.g., Smith & Kulikowich, 2004), using both models in a complementary fashion (e.g., Sudweeks, Reeve, & Bradshaw, 2004), or choosing the simpler model (e.g., Fan, 1998). Several studies have found Rasch models to provide richer data for evaluating and modifying instruments, especially when compared to very simple CTT methods (e.g., Haiyang, 2010).

However, to this point, none of the data discussed has been able to support that using Rasch produces better instruments or more accurate or useful scores. Many researchers conclude that because Rasch and observed score measures are highly correlated, they are essentially the same. In the case of instrument reduction, it is not surprising that two scales reduced from the same larger set of items are both similar to the original; these are similar measures of one construct, with data drawn from the same population. In the case of observed scores and the resultant Rasch measures, it is not surprising, either conceptually or statistically, that the correlations are high between observed scores and Rasch measures. Conceptually, the observed score and Rasch measure each represent a measure of the same construct, operationally defined using the same items and measured in the same population. Statistically, it is a property of the Rasch model that observed scores and Rasch person measures advance monotonically;

each observed score corresponds to exactly one person measure, and any higher observed score will have a higher Rasch person measure than a lower observed score (Linacre, 1992; Wright, 1999; Wright & Linacre, 1989). Therefore, correlation studies tell us little about substantive and potentially consequential differences between observed scores and Rasch person measures.

Comparisons of Rasch and CTT for conclusion validity. Few studies have gone beyond the correlational research described above to examine conclusion validity (Shadish, Cook, & Campbell, 2002). Conclusion validity is an instrument's ability to lead to the correct statistical conclusion, i.e. whether an effect is statistically significant or non-significant, and represents a profound impact of different measurement models such as Rasch and CTT.

Cheema (2013) compared Rasch (partial credit model) and observed scores for mathematics self-efficacy data and, replicating many prior results, found the correlation between the two sets of scores was high (r = .96). The scores were then entered into linear regression models to assess the impact of mathematics self-efficacy on mathematics achievement scores. This simulates how self-efficacy scores might be used in a study and how different scores might lead to different conclusions. The two linear regression models were almost identical (Rasch person measure $\beta = 50.49$; observed score $\beta = 50.59$) and explained very similar amounts of variance (Rasch person measures $R^2 = 30.2\%$; observed scores $R^2 = 30.0\%$). Based on these results, there was no basis for recommending Rasch over CTT scoring methods.

A demonstration of Rasch rating scale methodology on teacher science self-efficacy data included a test of conclusion validity: using Rasch and observed scores in a

paired-samples t-test with pre and post data (Boone, Townsend, & Staver, 2011). Results of the t-test using observed scores indicated a statistically significant change ($p < .05$), but the t-test using Rasch person measures indicated no significant effect ($p = .08$). Thus, the authors establish the importance of using Rasch analysis not only for scale development but also for creating interval scores that perform well in parametric statistics. Using ordinal scores when interval scores are assumed could lead to an incorrect conclusion of statistical significance.

Stewart (2012) performed a comparison of Rasch and observed scores using two different tests, one that is known to have high reliability (.91) and one with lower reliability (.75). There were high correlations between Rasch and observed scores for both of the tests (r's = 1.00). The data were then entered into an analysis of variance (ANOVA) in groups by classroom teacher. In the case of the high-reliability test, F values were high for both measures, and p values were very low (<.0001). The observed score model explained slightly more variance than the Rasch person measure model (R^2 = 24% and 23%, respectively). In the case of the low-reliability test, the observed score and Rasch person measure models both explained only small amounts of variance (R^2 = 6.6% and 6.6%, respectively), and both had low F values and non-significant p values ($p = .072$ and .069, respectively). However, a third model, a two-parameter IRT model explained a bit more variance (R^2 = 8.0%) and had a significant p value ($p = .02$). The results establish that Rasch and observed score models will not always produce different conclusions, but different measurement models used with the same data may produce results different enough to support divergent conclusions. It seems that divergent

conclusions may be more likely when an instrument has low reliability and/or when the parametric test has a *p* value near the conventional cutoff point of .05.

Another study of a low-reliability instrument corroborates the findings of Stewart (2012). Lynn and Lawless (2015) examined a science assessment with a high level of rater agreement (ICC > .90) but a low Cronbach alpha (.76). When used in an analysis of covariance (ANCOVA), observed scores suggested no statistical significance ($p = .132$) and Rasch person measures found a significant result ($p = .032$) (Lynn & Lawless, 2015). Lynn, Yukhymenko and Lawless (in preparation) used similar methodology with a set of five interest and self-efficacy subscales, and despite all of them having acceptable reliability (.79 to .94), Rasch person measures revealed significant effects (*p's* = .029 and .035) in two subscales where observed scores had non-significant *p* values (.059 and .055). Another approach to assessing conclusion validity is via the impact on individuals of different measurement models. In a comparison of MFRM and GT methods for rater-scored essays, MacMillan (2000) considered how Rasch scoring would impact the grades of individuals. Compared to their original uncorrected scores, 5.5% of students would see their score reduced enough to get a lower letter grade; 3.7% would receive a higher letter grade.

In all of these studies of conclusion validity, it is difficult to know empirically whether the observed score or the Rasch person measure is the correct one. In the case of MacMillan (2000), students whose grades went up would be in favor of Rasch; students whose grades went down would prefer to keep their original observed scores. For researchers, it might be tempting to try statistical tests using both Rasch and observed

scores and choose the more favorable result. The only way to judge which set of scores is more accurate is by carefully considering how they are different.

Boone and Scantlebury (2006) provided an illustrative example of how Rasch person measures and observed scores differ. In their dataset based on a 28-item test (plotted in Figure 1), a student who scored 26 differs from a student who scored 27 by 1 point in observed score and also by 1 logit. However, a student who scored 15 differs from a student who scored 16 by 1 point in observed score, but by only 0.10 logits (Boone & Scantlebury, 2006). This highlights the issue of interval data: the true intervals between observed scores can vary by a factor of 10 in this example. To extend the example, if post-intervention scores on this same test showed that students had scored, on average, one point higher than on a pre-test, it would matter very much whether those one-point gains had occurred at the high or low ends of the scale or near the center.

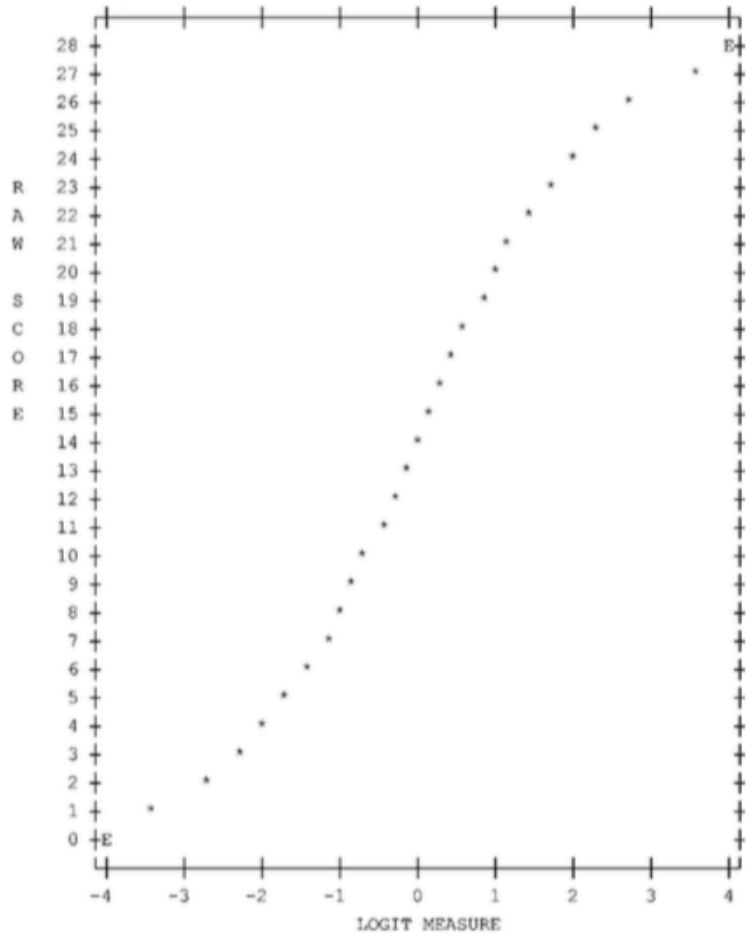

Figure 1. Observed scores (y-axis) and equivalent Rasch logit measures (x-axis) for a science test. The plot illustrates the non-equal intervals between observed scores. The data on this plot is from a particular dataset, but the curve shown is characteristic of any set of observed scores and Rasch person measures. Reprinted from Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2), p. 257. Copyright 2005 by Wiley Periodicals, Inc.

Figure 1 also illuminates the inadequacy of using correlation data to compare observed scores and Rasch person measures. The observed scores and Rasch person

measures are rank-ordered the same; higher observed scores always correspond to higher logit scores. Thus, despite the relationship being visibly curvilinear rather than linear, the two sets of scores will always have a strong positive correlation.

The ideal method for empirically comparing Rasch measures and observed scores is by using simulated data. Simulated data studies use data that fits a certain set of conditions to make broadly generalizable statements about statistical applications; for instance, Davison and Sharma (1990) established that in many situations, ordinal scores can be used in place of an interval measure of a latent variable in t-tests, correlation, and multiple regression, but not in factorial ANOVA. Extensions of this work show that when observed scores are used in factorial ANOVA, spurious interactions may appear to be present in certain situations, and interaction effect sizes may be underestimated in certain situations, related to the appropriateness of the test difficulty for each group within this research design (Embretson, 1996; Romanoski & Douglas, 2002). These studies, taken together, suggest that in many situations, observed scores are inadequate for use in parametric statistics.

Summary of Rasch and CTT Comparisons

A question was posed previously about the use of observed scores in parametric statistics: does it matter? The review of literature suggests that in many situations it does matter. Studies using simulated observed score data with a variety of group characteristics and test characteristics have found that observed score data will, under many circumstances, produce spurious main effects and interaction effects or mask true effects (Davison & Sharma, 1990; Embretson, 1996, 2006; Romanoski & Douglas, 2002). However, these findings and their important implications for educational research

have not been widely disseminated outside the measurement literature, as evidenced by the preponderance of educational researchers using observed scores to evaluate interventions. There are many possible reasons for this: it may be that theoretical arguments, even when supplemented with simulated data, are not very accessible to researchers with limited mathematics and statistics training; it may be difficult for a principal investigator planning an intervention with thousands of students and millions of dollars in funding to give adequate attention to the practical implications of the scale of measurement they are using. It is possible that a demonstration of the specific and concrete effect of using actual observed scores versus actual interval measures in an educational intervention efficacy study will help bridge the gap between the theoretical and simulated data research of the measurement literature and the broader community of educational intervention researchers. The use of real (not simulated) data in this demonstration is purposive: it brings the complexity of real data (as pointed out by Macmillian, 2000), including uneven distributions, possible outliers, and missing data. Additionally, situating a demonstration of a theoretical statistical principle within a specific context of an intervention, participants, measures, and outcomes may have educative value to researchers who would struggle to fully understand the applicability of a simulated study.

Gaps in the Literature

The foregoing review of literature has identified significant gaps in the research that warrant further exploration. Educational intervention studies indicate mixed effects, as was demonstrated using PBL as an exemplar pedagogical approach to educational interventions. After two decades of implementations of PBL in K-12 settings, it is still

not clear exactly whether PBL offers advantages as an instructional format over and beyond the learning students achieve in traditional classroom activities. There are many potential reasons for the mixed findings, including the outcomes that are assessed and the forms of assessment used; however, no study has yet explored the implications of using observed scores, which are ordinal data, in typical parametric statistics. There are theoretical reasons and some empirical justification for the concern about the use of ordinal data in parametric statistics. The controversy about whether it is acceptable to use ordinal data in parametric statistics has tended to be one of theory versus pragmatism: theoretically, ordinal data is not appropriate for parametric statistics (Jamieson, 2004; Stevens, 1946); however, if ordinal data "works" in parametric statistics, it is difficult to justify the extra effort involved in creating interval scales (Norman, 2010; Stevens, 1946).

Rasch measurement provides a rescaling method to create interval data from ordinal observed scores and offers the additional benefits of providing rich information about item functioning in objective and attitude assessment and, using MFRM, examining rater effects in subjective assessments. Although many studies have compared Rasch and CTT methods for scale development and compared the resultant scores by correlation, only a few studies have examined the differential conclusions produced by Rasch measures and observed scores. Even fewer studies have used data from actual interventions to determine whether the interval data produced by Rasch measurement "works" better than observed scores. To date, no studies have compared observed scores and MFRM person measures to determine how interval scale data corrected for rater effects performs differently from observed scores in a test of intervention efficacy.

The purpose of the present study is to add to the line of conclusion validity literature that uses real data by comparing observed score and Rasch measures in the context of a large, grant-funded, PBL intervention study, which serves as an example of a well-developed, theoretically sound educational intervention. The exemplar intervention study is a randomized controlled efficacy trial with adequate statistical power and demonstrated to meet WWC evidence standards for randomization, attrition, implementation, and measurement (Lawless & Brown, 2012; WWC, 2014). These study characteristics rule out poor design and low power as competing hypotheses and focus on the role of the measurement scale in analysis. Factors that have been shown to impact the effectiveness of PBL interventions such as domain or subject matter (Walker & Leary, 2009), problem type (Walker & Leary, 2009), grade level (Hmelo-Silver, 2004), and group format (Hmelo-Silver, 2004) can also be ruled out because they are static across the study.

The results are not limited in their applicability to one particular PBL intervention or to PBL as a pedagogical approach, but are meant to make a methodological contribution that is broadly applicable to educational interventions with similar measurement issues. The present study will examine correlations between Rasch person measures and observed scores to replicate prior research and to establish that in this dataset, as in any dataset, there is a very high correlation between the corresponding data. The present study will further examine the conclusions that each set of scores/measures provides using dichotomous data, rating scale data, and rater-scored performance assessment data in order to highlight the crucial impact of measurement scale: that of influencing whether or not an intervention is found to be effective.

The research questions for the study are as follows:

1. How closely correlated are corresponding sets of observed scores and Rasch person measures for each of the following outcome assessments used in a PBL intervention: (a) objective test, (b) attitude measure, and (c) rater-scored essay?

2. Do observed scores and Rasch person measures lead to different conclusions about the impact of a PBL intervention on each of the following student outcomes: (a) science knowledge, (b) interest and self-efficacy, (c) writing quality?

Chapter 3: Method

Participants

The total sample of participants is 1,979 seventh- and eighth-grade students who agreed to participate in a large, grant-funded, PBL intervention study known as GlobalEd 2 (Brown, Lawless, & Boyer, 2013; Lawless & Brown, 2015) in fall 2014. Of those, 1,066 participants were in urban schools and 913 in suburban schools, located in or near a large Midwestern city and in or near a large Northeastern city in the United States. Participant classrooms were randomly assigned to GlobalEd 2 (n = 998) or a control condition (n = 980). The total sample was approximately evenly divided across gender (male n = 923; female n = 1,005) and grade (seventh grade n = 1,018; eighth grade n = 957). Participants' median age was 13 years at the beginning of the school year. Forty-two percent of the participants identified as White, 13% Black, 29% Hispanic or Latino/a, 6% Asian or Pacific Islander, 9% other races/ethnicities or multiple races/ethnicities.

Intervention

GlobalEd 2 is an online, problem-based learning simulation in which students research and discuss a real-world socio-scientific issue such as global freshwater shortage (Brown, Lawless, & Boyer, 2013; Lawless & Brown, 2015). Each classroom represents a country other than the United States, and within each classroom/country, students form issue groups to focus on health, environment, human rights, and economics. The goal is for each classroom/country to form an international water treaty with at least one other country. The simulation begins with a six-week research phase. During the research phase, students research the geography, politics, and culture of their country, as well as

the greater global implications of freshwater shortage. Each issue area group works in depth to understand the specific concerns of their country relative to freshwater shortage and their issue area topic. Students have access to curated online resources and may also use their own resources (textbooks, library resources, websites, etc.) to perform research. The culmination of the research phase is the opening statement, in which each group writes and posts a formal position statement for other countries to read, outlining their specific goals for a treaty.

The opening statements and subsequent interactions are purposefully text-based to promote writing skills within the context of a socio-scientific issue area. In particular, students are directed and supported to use a CER form of argumentation in their interactions (McNeill & Krajcik, 2006; 2008; Toulmin, 1958). Interactions among classrooms/countries take place within an ICONS simulation, which is a closed online environment where students are known by anonymized handles that indicate only their country and issue area. Interactions are supervised by a trained adult facilitator known as Simcon, who ensures that interactions are professional and on-topic and also provides continuous feedback to students on their argumentation and the content of their postings.

After opening statements are posted, students begin to interact across classrooms using asynchronous messaging (similar to email) and several live, real-time web conferences. The goal of the interactive phase is for countries to form alliances and to make progress in addressing some of the problems of freshwater shortage they have discovered. This phase is student-driven in the sense that students may choose to focus on conservation, technology, economic incentives, or any other appropriate pathway to addressing global freshwater shortage.

After six weeks of interaction, each group posts a closing statement summarizing the progress they have made with negotiations. Finally, a two-week debriefing phase supports students in reflecting on their experience and thinking about its applications in other contexts.

Besides acting as facilitators for their students' PBL scenario, teachers receive intensive, ongoing professional development in GlobalEd 2. Professional development begins in the summer before the intervention and continues throughout the simulation. The first phase, which is for new GlobalEd 2 teachers, is a three-week, online, self-paced course of video modules and activities that cover the principles of PBL, the social studies and science content germane to the topic of global freshwater shortage, argumentative writing, and assessment in PBL. Both new and veteran GlobalEd 2 teachers participate in a one-day, live workshop that includes discussion of past challenges, ideas for the upcoming year, and concludes with a "mini" simulation that mimics the longer simulation in which students will participate.

The intervention study is funded as an IES Goal 3 grant for efficacy and replication, which specifies that the intervention is carried out under ideal conditions, including substantial support for teachers (Lawless & Brown, 2012). Throughout the student simulation, teachers are supported with a variety of resources and ongoing professional development. The teacher resources website is regularly updated with comments from the curriculum support team, worksheets, handouts, lesson plans, podcasts, and other resources. Communications among teachers and the curriculum support team are maintained throughout the simulation with a simulation newsletter that addresses the concerns of the upcoming week. Teachers complete a weekly activity log

where they are encouraged to reflect on their practice and list any resources they need or questions they have. Teachers also have direct access to an assigned GlobalEd 2 teacher liaison who can provide individual support.

Instruments

Instruments were developed to evaluate the impact of the intervention as part of a multi-year efficacy trial funded by an IES grant. The format, content, and rating procedures for each instrument are reported below.

Science knowledge. Factual knowledge of science was assessed using a multiple-choice quiz of 18 items, with a focus on recall of water cycle vocabulary and essential facts, such as, "What word means the change of state from liquid to a gas?" (evaporation) and, "How much of Earth's water is fresh water?" (3%). The science knowledge quiz is reproduced in appendix A.

Affective outcomes. Interest and self-efficacy scales were adapted from scales that have been used previously in GlobalEd 2 and have been found to have high reliability (Cronbach's alpha = .79 to .94) (Lynn, Yukhymenko, & Lawless, 2015). The interest scale consisted of three subscales: science interest (5 items, e.g., "Learning about science topics interests me"), science career interest (5 items, e.g., "I am interested in pursuing a science career in the future"), and social studies interest (4 items, e.g., "I am interested in other countries/cultures"). Each item used a five-point Likert-type response scale of (1) Strongly Disagree, (2) Disagree, (3) Undecided, (4) Agree, (5) Strongly Agree. The interest scales are reproduced in appendix B.

The self-efficacy scale consisted of two subscales: writing self-efficacy (5 items, e.g., "How confident are you that you can write a well organized essay on a given

topic?") and socio-scientific self-efficacy (4 items, e.g., "How confident are you that you can learn how science and social studies are related?"). Students responded using a five-point Likert-type scale of (1) Not confident to (5) Extremely confident. The self-efficacy scales are reproduced in appendix C.

Argumentative writing. To assess argumentative writing, students were given 30 minutes to respond to the following prompt: "The world is in danger of running out of freshwater. Do you think this is true? Do you agree or disagree with this statement? Why?" Students were further instructed to choose a position, use evidence and reasoning to support their position, organize their writing, and proofread for spelling, punctuation, and grammar. Essays were scored using a rubric based on the claim/evidence/reasoning (CER) chain of argumentation (McNeill & Krajcik, 2006; 2008; Toulmin, 1958). The rubric also included elements of organization, science content, and social studies content. The argumentative writing prompt is reproduced in appendix D.

Raters were 13 educational psychology graduate students across the two study sites who were trained in a group setting on use of the rubric (reproduced in appendix E). Essays from each site were randomly assigned to raters at that site. Raters were blinded to the identity of the student, the condition (treatment or control), and whether the essay was a pre or post. In order to ensure consistent ratings, each essay was assigned to two raters; each rater was blinded to the identity of the other rater and the scores the other rater awarded to an essay. In cases where the two raters disagreed on a score for any category, discrepant scores were resolved by the following rules: If the two raters assigned scores above zero that were within one point, the mean of the two scores was retained (e.g. rater A assigned 1 for science content and rater B assigned 2, the final score is 1.5). If one rater

assigned a zero and the other assigned a non-zero or if the two scores were more than one point apart (e.g. rater A assigned 1 for science content and rater B assigned 3), a third rater was dispatched to score the essay. The mostly closely agreeing scores among the three raters were retained (e.g. rater A assigned 0 for science content, rater B assigned 1, rater C assigned 1, the final score is 1; rater A assigned 0 for science content, rater B assigned 1, rater C assigned 2, the final score is 1.5). Nominal scores as to the student's overall essay position (agrees with the prompt, disagrees with the prompt, takes both positions, or takes no position) required agreement by at least two raters.

Inter-rater agreement of at least 80% across all raters was monitored and maintained throughout the scoring period. Every other week, all raters across both sites scored 10 essays in common. For each category in each essay, rater responses were compared to the mode response. The percent agreement for each category was the number of rater agreements with the mode divided by the total number of raters (for the purposes of inter-rater agreement, non-zero scores within one point were considered an agreement). The mean percent agreement across categories represented the agreement for a given essay, and the mean percent agreement across the sample of 10 essays was the overall inter-rater agreement. Mode responses were reviewed for adherence to the rubric. Additionally, agreement was calculated for individual raters, and raters who dropped below 80% agreement received notes on their disagreements and areas for improvement. Exact inter-rater agreement was primarily used for ongoing training purposes. For the instrument overall (total numeric score) intra-class correlation coefficients (one-way random model) ranged from .649 (single measures) to .969 (average measures).

Procedure

Teachers ($N = 53$) were recruited in the spring and summer to have two of their seventh- and/or eighth-grade social studies classrooms participate in GlobalEd 2 the following fall. Recruitment focused on teachers who had participated in prior instantiations of GlobalEd 2, word-of-mouth, email, and social media. Teachers provided informed consent, and principals at each school provided written approval to have classrooms participate. Teachers received a stipend for their participation in an online professional development seminar and for their participation; schools also received a stipend for participating.

After permission was obtained from parents and assent was obtained from students, pre-assessments were administered by teachers over the course of three school days. Teachers were instructed to administer the essay assessment first, followed but the remaining assessments in any order and spread over the three days in whatever manner would suit their class schedule.

After pre assessments were administered, each teacher was randomly assigned one classroom to participate in the intervention and one to serve as a control. The purpose of this design is to increase the power of the study by including two classrooms per teacher (Lawless & Brown, 2012). This design means that control and treatment groups are not fully independent (as they share a set of teachers); there is a possibility of some contamination across the groups. However, contamination is believed to be minimal, because teacher professional development is situated in the context of the online GE2 environment, and students in control classrooms do not have access to this environment. Additionally, the study includes an assessment of fidelity of implementation to determine

whether classroom practices were qualitatively different in the control and treatment classrooms; this will aid in illuminating any contamination effects. Further discussion and study of contamination effects are outside the scope of this dissertation.

At the end of the 14-week intervention, teachers administered post assessments that were identical to the pre assessments and using the same procedures that were used for the pre assessments.

After the simulation was completed and post assessments were collected and analyzed, one teacher's classroom data was dropped from the dataset due to extensive plagiarism detected in a majority of post assessments from the teacher's treatment and control classrooms. The final number of teachers included in the dataset was $n = 52$.

Analysis

The purpose of the analysis is to compare the observed scores on each measure to the equivalent Rasch measures and to compare the statistical conclusions that observed scores and Rasch measures produce. In order to demonstrate a common practice, analyses were first conducted at the student level and a post-hoc correction applied to account for randomization at the classroom level; classroom level aggregated analyses will also be performed (Song & Herman, 2010; WWC, 2014). The following section details how total observed scores and Rasch person measures were calculated for each assessment. Scale and item characteristics are reported as detailed below; however, as all items and scales have been used previously in published studies and item/scale analysis is not a main focus of this study, items and scales were not revised and analysis proceeded even if some items or scales displayed less than optimal characteristics.

Observed Scores

The science knowledge scale produced dichotomous data. The only possible scores for each item were 0 or 1. Items skipped were scored incorrect. Total scores were calculated as the sum of all correct items out of 18 possible. Students who did not turn both a pre and post science knowledge assessment were excluded listwise. Item difficulty (p) was reported for each item and Cronbach's alpha was reported for the scale.

The interest and self-efficacy scales produced polytomous data. Each item had a range of 5 points. Students who did not respond to all of the items on a scale at both pre and post time points were excluded listwise for that scale. Item difficulty (mean scores) were reported for each item and Cronbach's alpha was reported for the scale.

The writing rubric consisted of seven traits with varying category structures. The essay's claim/evidence/reasoning structure was scored in three separate traits as claim (0-1-2), evidence (0-1-2-3), and reasoning (0-1-2). The essay was also scored for organization (0-1-2), addressing opposition (0-1-2), science content (0-1-2-3), and social studies content (0-1-2-3). Each essay was scored by two raters who were blind to the student's identity, the student's placement in treatment or control condition, whether the essay was pre or post, and the other rater's identity. Where the two raters agreed on a score, that score was retained. Where the two raters were within one point and neither scored the trait 0, the mean of the two scores was retained (e.g. scores of 1 and 2 become a final score of 1.5). Where the two raters differed by more than one point or one scored a 0 and the other scored above 0, a third rater, blind to the first two raters' identities and scores, rescored the essay as a tie-breaker. The score given by two raters was the final score (e.g. scores of 0, 1, and 0 became a final score of 0). The total essay quality score

will be calculated as the sum of scores on all traits. Students who did not turn in both a pre and post essay were excluded listwise. Difficulty (mean scores) were reported for each category.

Rasch Person Measures

All Rasch person measures were estimated using FACETS software (Linacre, 2015). Dichotomous data from the science knowledge assessment will be analyzed using the Rasch dichotomous model (Rasch 1960/80). Rating scale data from the interest and self-efficacy scales were analyzed using the partial credit model (Masters, 1982). As detailed in Chapter 2, the partial credit model allows for items to have more than two possible outcome values and also allows for each item to have a unique rating scale structure. Data from the essays was analyzed using the Many-Faceted Rasch Model, which allows multiple outcome values, flexible rating scale structure, and also accounts for the characteristics of raters (Linacre, 1989). The models are specified below.

Dichotomous model (science knowledge assessment). The science knowledge data were analyzed using the following model, which produces a measure of ability for each participant, as well as a difficulty measure for each item:

$$\Pr\{X_{ni} = x\} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)} \quad (6)$$

where the probability (Pr) of person *n* answering item *i* correctly (i.e. score = 1) is a function of the person ability (β_n) and the item difficulty (δ_i).

Partial credit model (interest and self-efficacy assessments). The interest and self-efficacy data were analyzed using the following model, which produces a measure of

ability for each participant, a difficulty measure for each item, and threshold measures for the rating scale categories for each item:

$$\Pr\{X_{ni} = x\} = \frac{\exp \sum_{k=0}^{x} (\beta_n - \tau_{ki})}{\sum_{h=0}^{K} \exp \sum_{k=0}^{h} (\beta_n - \tau_{ki})} \quad (7)$$

where the probability of person *n* getting a score of *x* on item *i* is a function of the person ability and the thresholds (τ_{ki}) between adjacent scores on each item.

Many-facet Rasch measurement model (MFRM) (writing assessments). The writing assessment data were analyzed using Myford and Wolfe's Hybrid Model 3 (2004), which produces a measure of ability for each participant, a difficulty measure for each item, threshold measures for each rubric category as applied by each rater, and a measure of severity for each rater:

$$\Pr\{X_{nikj} = x\} = \frac{\exp \sum_{k=0}^{x} (\beta_n - \delta_i - \tau_{kij} - \alpha_j)}{\sum_{h=0}^{K_i} \exp \sum_{k=0}^{h} (\beta_n - \delta_i - \tau_{kij} - \alpha_j)} \quad (8)$$

where the probability of person *n* getting a rating of *x* on trait *i* from rater *j* is a function of the person ability, the trait difficulty, the threshold between adjacent ratings as applied by each rater, and the severity of the rater (α_j).

In the case of writing assessment data, ratings from only the first two independent raters were used.

FACETS produces Rasch person measures (in logits), which are estimated by fitting items and persons to the selected Rasch model. Before estimating person measures, several diagnostic statistics were examined to ensure the collected data fit the Rasch model. As a general measure of internal consistency for each scale, Rasch person reliability was reported and can be interpreted the same way Cronbach's alpha or KR20

would be interpreted. Values above .80 indicate the scale can discriminate at least three separate levels of performance and are ideal; values of .70 to .80 discriminate two separate levels of performance and are not ideal but still useful (DeVellis, 2003; Smith, E.V., 2004). For each item, infit and outfit mean-squares were reported; values above 1.5 are flagged as suggesting poor item fit (Linacre, 2003; Smith, R.M., 1996b; Wright & Linacre, 1994). Each item was verified to have a positive point-measure correlation; a negative correlation implies a miscoded or severely off dimension item (Smith, E.V., 2002; Wright, 1992). For polytomous items (interest and self-efficacy rating scale items and writing categories), the above referenced indicators of model fit were reported along with the following: number of observations in each category (values below 10 are flagged), average measures (disordered average measures are flagged), thresholds (disordered thresholds are flagged), and category outfit mean-squares (values above 2.0 will be flagged), following the recommendations of Smith, Wakely, De Kruif, and Swartz (2003) and Linacre (2004). Each scale is theoretically unidimensional; this is both the intent of each scale and a requirement of the Rasch model. Any combination of negative point-measure correlation, poor item fit (misfit or overfit), and/or high eigenvalue (generally greater than 2.0 [Linacre, 2015]) for the first component of a PCA of standardized residuals may suggest multidimensionality or otherwise poor fit between the data and the Rasch model (Linacre, 1992). Therefore, eigenvalues greater than 2.0 are reported as suggesting poor model fit and possibly multidimensionality.

Revision or removal of items to improve scale functioning is generally outside the scope of this study; therefore, analysis proceeded even in the presence of items or scales that do not display optimal characteristics.

When using Rasch measures for pre-post analysis, it is important to verify that the pre time-point instrument and the post time-point instrument have invariant items and thresholds; otherwise, it is unclear whether students have actually gained knowledge or whether their interpretation of the instrument has changed somehow (Wright, 1996). If the instrument functions differently at the post time-point than it did at the pre time-point (for instance, if some items are substantially more difficult or less difficult, or if students interpret the rating scales differently), the person measure estimates from the two time points are not directly comparable.

Therefore, before proceeding with pre-post analysis, pre and post item difficulties and thresholds were verified to be invariant. For dichotomous data (science knowledge assessment), item difficulty levels in the pre and post instruments were examined. Item difficulties were estimated independently for each pre and post instrument. The item difficulty measure and standard error were used to calculate a 95% confidence interval for the difficulty of each item. If the 95% confidence intervals of an item on the pre instrument and the corresponding item on the post instrument overlapped, they were not considered significantly different (Wright & Stone, 1979). For rating scale data (interest and self-efficacy assessments) the same overlapping confidence interval procedure was followed for item difficulty and also applied to category thresholds. Finally, for MFRM data (writing assessments), the procedure was expanded and applied to trait difficulty, individual category thresholds for each trait, and the severity of each rater.

In cases where functioning differs from pre to post, corrective procedures were used to create instruments that are identical at each time point. Procedures are adapted from Mallinson (2001), Wright (1996, 2003), and Wolfe and Chiu (1999) to be consistent

for dichotomous, rating scale, and MFRM data. For each assessment where functioning differs from pre to post, a common item structure was estimated by creating a dataset with pre data from a randomly selected half of participants and post data from the other half; this avoids a violation of local independence that would be created by putting each participant in the dataset twice (Mallinson, 2001). This stacked dataset was used to estimate item difficulties and rating scale threshold structure (in the case of rating scale data) that were common across the two time points. Then, these item difficulties and category thresholds were used as anchor values (specified *a priori*) in both the pre and post datasets; pre and post person measures were estimated separately and are directly comparable because they share a common item and rating scale structure.

Where rater severity differed among pre and post essays, the same procedure used to determine common item difficulties and rating scale thresholds was applied to rater severity and the common rater severity applied as an anchor value for pre and post data. In the case of the rater-scored essays, all participants were entered into the stacked dataset, as all participants are already included at least twice in the original pre and post datasets.

Finally, the person measures were exported to an SPSS-compatible file (IBM, 2015) for use in parametric statistics.

Data Normality and Missing Data

For all observed scores and Rasch person measures, data normality and missing data are reported. Normality is reported in terms of skewness, kurtosis, and overall normality evaluated by the Kolmogorov-Smirnov test (Field, 2005). Missing data is

reported as a percentage of all data and evaluated using Little's test for data missing completely at random (MCAR) (Tabachnick & Fidell, 2001).

Research Questions: Comparison of Measurement Models

RQ1: Correlation of scores. The first research question addresses the similarity of observed scores and Rasch person measures on each of the instruments examined as part of this study (i.e., science, interest, self-efficacy, and writing). The procedure for answering this question in numerous prior studies has been to conduct Pearson product-moment correlations between two corresponding sets of scores (e.g., Fan, 1998; MacDonald & Paunonen, 2002). Therefore, I replicated prior work by examining Pearson correlations between scores. Where correlations between scores are high, positive, and significant, the observed scores and Rasch person measures are very similar or identical in their rank ordering of students.

RQ2: Conclusion validity. The second research question asks whether observed scores and Rasch person measures lead to different conclusions in terms of statistical significance. To answer this question, analyses of covariance (ANCOVAs) were conducted using observed scores and Rasch person measures. For each ANCOVA, the post score was the dependent variable, the pre score was the covariate, and treatment status (intervention or control classroom) was the fixed independent variable. Levene's test was used to ensure that variances were equal between treatment and control groups for any one assessment (Field, 2005). Additionally, homogeneity of regression slopes was ensured by examining the interaction between the independent variable and the covariate (Field, 2005).

The value of alpha is .05 (*p* values less than .05 are considered statistically significant) for each outcome. When performing multiple hypothesis tests, WWC recommends using the Benjamini-Hochberg correction to guard against inflated Type I error (WWC, 2014). The correction is applied when multiple hypothesis tests are used for multiple measures of the same domain (WWC, 2014). In the present study, each outcome represents a different domain; in an evaluation of the intervention, each outcome measure would answer a different research question. Therefore, the uncorrected alpha value of .05 is retained for each ANCOVA and ANOVA.

Correction for cluster randomization. When random assignment is carried out at the cluster level (e.g. classrooms are assigned randomly to treatment or control) and analyses are carried out at the individual (student) level, there is a mismatch between research design and analysis. Although multilevel analysis is an ideal solution, it is still common for researchers to use individual-level analyses to determine the impacts of cluster-randomized studies (Song & Herman, 2010). In fact, this practice is so common that WWC has developed a corrective procedure that takes into account the intra-cluster correlation (ICC) in order to correct for similarities within the cluster groups (WWC, 2014). ICCs were calculated for each measure as the proportion of within-cluster variance that makes up the total variance (Killip, Mahfoud, & Pearce, 2004). WWC uses the ICC to compute clustering-corrected *p* values (WWC, 2014). This dissertation follows WWC protocol in reporting clustering-corrected *p* values as shown in Figure 1. First, the effect size (Hedges' *g*) was calculated as the standardized mean difference

between group post scores, adjusted for the covariate.[4] Next the *t*-statistic for the effect size, ignoring clustering, was calculated. Then the corrected *t*-statistic and associated degrees of freedom (*df*) was calculated using the number of groups (*M*) and the intra-class correlation coefficient (ICC) (ρ). The corrected *t*-statistic and corrected degrees of freedom were used to obtain the corrected *p*-value.

$$\rho = \frac{s_b^2}{(s_b^2 + s_w^2)} \tag{9}$$

Compute ICC (ρ) as a function of variance between clusters (s_b) and variance within clusters (s_w).

$$g = \frac{y'_i - y'_c}{\sqrt{\frac{(n_i-1)s_i^2+(n_c-1)s_c^2}{n_i+n_c-2}}} \tag{10}$$

Compute effect size *g* as a function of covariate-adjusted group mean scores (y_i' and y_c'), and pooled standard deviation.

$$t = g\sqrt{\frac{n_i n_c}{n_i+n_c}} \tag{11}$$

Compute *t*-statistic for effect size, ignoring clustering, as a function of the effect size (*g*) and the size of the intervention group (n_i) and the control group (n_c)

[4] WWC protocol multiplies effect size *g* by a factor of ω to correct for small sample sizes (Hedges, 1981; WWC, 2014). In this dataset, the multiplier ω is extremely close to 1 (.9996). Therefore, ω is disregarded in this study.

$$t_a = t\sqrt{\frac{(N-2)-2\left(\frac{N}{M}-1\right)\rho}{(N-2)\left[1+\left(\frac{N}{M}-1\right)\rho\right]}} \quad (12)$$

Compute corrected *t*-statistic, taking into account the number of groups (*M*) and ICC (ρ).

$$df = \frac{\left[(N-2)-2\left(\frac{N}{M}-1\right)\rho\right]^2}{(N-2)(1-\rho)^2+\frac{N}{M}\left(N-2\frac{N}{M}\right)\rho^2+2\left(N-2\frac{N}{M}\right)\rho(1-\rho)} \quad (13)$$

Compute corrected degrees of freedom (*df*).

Figure 2. WWC protocol for calculating corrected *p*-values when individual effect sizes are reported in a cluster-randomized study (WWC, 2014).

Additionally, to correct for possible contamination across control and treatment groups, which share a set of teachers, a classroom-level, two-way, repeated-measures ANOVA was conducted. Classroom-level variables were created for two different time points (pre and post) and two different treatment conditions (treatment and control) by taking the average of all student scores within a classroom and time point. A significant time x treatment effect indicates that GE2 had an impact on change in scores over time. As with the student-level ANCOVA, this analysis was repeated for observed scores and Rasch measures.

For the student-level analysis, where observed score ANCOVA *p*-values and Rasch person measure ANCOVA *p*-values are on opposite sides of the .05 threshold, after *p*-values are corrected for clustering, they are considered to provide different conclusions. For the classroom-level analysis, where observed score ANOVA *p*-values and Rasch person measure ANOVA *p*-values are on opposite sides of the .05 threshold,

they are considered to provide different conclusions; no correction for clustering is necessary.

Chapter 4: Results

Science Knowledge

Observed score analysis. The science knowledge pre scale had a range from 0 to 18 points, with an overall mean score of 9.87 (*SD* = 3.17). The data had a significant negative skew (skewness = ‑.24, *SE* = .06) and were significantly platykurtic (kurtosis = ‑.37, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1925) = .08$, $p = .000$. Cronbach's alpha for the 18-item scale was .66, and all items had a positive corrected item-total correlation. Difficulty ranged from .83 (least difficult) to .17 (most difficult). Item difficulties and discrimination are listed in Table 1. Data from 54 students (approximately 3%) was missing. Missing pre score data was associated with younger students, schools with higher percentages of free and reduced lunch, and a lower score on the post scale. Missing data were more common among urban and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally common among treatment and control groups, $\chi^2(1) = .04$, $p = .95$. Patterns of missing data were comparable for treatment and control groups.

The science knowledge post scale had a range from 0 to 18 points, with an overall mean score of 10.61 (*SD* = 3.35). The data had a significant negative skew (skewness = ‑.40, *SE* = .06) and were significantly platykurtic (kurtosis = ‑.33, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1881) = .09$, $p = .000$. Data from 98 students (approximately 5%) was missing. Cronbach's alpha for the 18-item scale was .71, and all items had a positive corrected

item-total correlation. Difficulty ranged from .83 (least difficult) to .19 (most difficult). Item difficulties and discrimination are listed in Table 1. Missing post score data was associated with a higher free-and-reduced-lunch population and a lower score on the pre scale and was more likely to be missing from urban students, non-White students, male students, and eighth graders. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally common among treatment and control groups, χ^2 (1) = .90, p = .34.

Table 1

Difficulty and Discrimination of Science Knowledge Scale Items

	Pre scale		Post scale	
Item	Difficulty	Discrimination	Difficulty	Discrimination
1	.32	.175	.33	.162
2	.43	.077	.58	.195
3	.77	.263	.80	.338
4	.47	.298	.59	.224
5	.17	.152	.19	.152
6	.78	.338	.83	.398
7	.60	.317	.66	.371
8	.67	.217	.68	.262
9	.49	.250	.56	.321
10	.67	.330	.74	.422
11	.76	.238	.70	.274
12	.70	.418	.73	.445
13	.72	.180	.67	.246
14	.33	.171	.37	.191
15	.19	.135	.23	.187
16	.83	.350	.85	.409
17	.53	.350	.59	.394
18	.46	.325	.52	.272

Rasch analysis. The science knowledge pre scale had person reliability of .68, which is below the desired reliability of .80. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The pre scale first-contrast eigenvalue was 1.4, which met the requirement of being below 2.0. Item fit statistics are reported in Table 2.

The science knowledge post scale had person reliability of .70, which is below the desired reliability of .80. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The pre scale first-contrast eigenvalue was 1.4, which met the requirement of being below 2.0. Item fit statistics are reported in Table 2.

Table 2

Science Knowledge Item Fit Statistics

	Pre scale			Post scale		
Item	Infit mean square	Outfit mean square	Point-measure correlation	Infit mean square	Outfit mean square	Point-measure correlation
1	1.07	1.12	.17	1.11	1.28	.16
2	1.17	1.29	.08	1.11	1.13	.20
3	0.99	0.95	.26	0.94	0.89	.34
4	0.97	0.97	.30	1.08	1.17	.22
5	1.02	1.25	.15	1.05	1.39	.15
6	0.93	0.81	.34	0.88	0.72	.40
7	0.96	0.92	.32	0.94	0.88	.37
8	1.04	1.08	.22	1.04	1.08	.26
9	1.02	1.03	.25	0.98	0.98	.32
10	0.94	0.90	.33	0.87	0.83	.42
11	1.01	1.04	.24	1.02	1.00	.27
12	0.86	0.79	.42	0.86	0.74	.44
13	1.07	1.16	.18	1.06	1.05	.25
14	1.08	1.10	.17	1.09	1.21	.19
15	1.05	1.20	.14	1.05	1.20	.19
16	0.90	0.75	.35	0.86	0.65	.41
17	0.92	0.91	.35	0.91	0.86	.39
18	0.95	0.93	.32	1.03	1.02	.27

Note. All items met specified fit criteria.

Where the 95% confidence intervals around the pre and post difficulty measures of an item overlapped, the item was considered to have stable difficulty over time. Five items had significantly different difficulty measures in the pre and post scales, as reported in Table 3. Therefore, combined difficulty measures were created based on a stacked file (half pre data and half post data, to avoid local dependence [Mallinson, 2001]), and the combined difficulty measures were used as anchor values for the pre and post datasets (Wright 1996, 2003). The use of anchor values did not substantially alter person reliability or item fit statistics.

Table 3

Science Knowledge Item Difficulties

	Pre scale difficulty		Post scale difficulty		Anchored
Item	Measure (SE)	95% CI	Measure (SE)	95% CI	measure
1*	1.14 (.05)	1.04, 1.24	1.38 (.05)	1.28, 1.48	1.24
2*	0.61 (.05)	0.51, 0.71	0.10 (.05)	0.00, 0.20	0.40
3	-1.19 (.06)	-1.31, -1.07	-1.24 (.06)	-1.36, -1.12	-1.26
4*	0.42 (.05)	0.32, 0.52	0.06 (.05)	-0.04, 0.16	0.29
5	2.15 (.07)	2.01, 2.29	2.28 (.06)	2.16, 2.40	2.22
6	-1.23 (.06)	-1.35, -1.11	-1.42 (.07)	-1.56, -1.28	-1.26
7	-0.22 (.05)	-0.32, -0.12	-0.32 (.05)	-0.42, -0.22	-0.29
8	-0.61 (.05)	-0.71, -0.51	-0.46 (.05)	-0.56, -0.36	-0.56
9	0.30 (.05)	0.20, 0.40	0.17 (.05)	0.07, 0.27	0.22
10	-0.59 (.05)	-0.69, -0.49	-0.78 (.06)	-0.90, -0.66	-0.62
11*	-1.14 (.06)	-1.26, -1.02	-0.58 (.06)	-0.7, -0.46	-0.94
12	-0.74 (.05)	-0.84, -0.64	-0.73 (.06)	-0.85, -0.61	-0.82
13*	-0.87 (.06)	-0.99, -0.75	-0.37 (.05)	-0.47, -0.27	-0.61
14	1.08 (.05)	0.98, 1.18	1.13 (.05)	1.03, 1.23	1.14
15	1.94 (.06)	1.82, 2.06	1.93 (.06)	1.81, 2.05	1.96
16	-1.62 (.07)	-1.76, -1.48	-1.58 (.07)	-1.72, -1.44	-1.62
17	0.12 (.05)	0.02, 0.22	0.03 (.05)	-0.07, 0.13	0.09
18	0.45 (.05)	0.35, 0.55	0.39 (.05)	0.29, 0.49	0.41

*Difficulty changed significantly pre to post.

Rasch person estimates for the science knowledge pre scale had a range from -4.52 logits to 4.66 logits, with an overall mean score of .24 (*SD* = 1.02). The data had a significant negative skew (skewness = ‑.21, *SE* = .06) and were significantly leptokurtic (kurtosis = 1.14, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1925) = .06$, $p = .000$. Rasch person estimates for the science knowledge post scale had a range from -4.52 logits to 4.67 logits, with an overall mean score of .49 (*SD* = 1.11). The data had a significant negative skew (skewness = ‑.13, *SE* = .06) and were significantly leptokurtic (kurtosis = .84, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1880) = .07$, $p = .000$.

Science Interest

Observed score analysis. The science interest pre scale had a range from 5 to 25 points, with an overall mean score of 17.06 (*SD* = 4.39). The data had a significant negative skew (skewness = ‑.30, *SE* = .06) and were significantly platykurtic (kurtosis = ‑.32, *SE* = .12). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1794) = .06$, $p = .000$. Cronbach's alpha for the 5-item scale was .87, and all items had a positive corrected item-total correlation. Mean scores for each item ranged from 3.82 (least difficult to endorse) to 2.81 (most difficult to endorse). Item mean scores are listed in Table 4. Data from 185 students (approximately 9%) was missing. Missing pre score data was associated with students in schools with higher percentages of free and reduced lunch. Missing data was more common among younger students, urban students, and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables,

missing data were equally common among treatment and control groups, $\chi^2(1) = 1.64$, $p = .20$. Patterns of missing data were comparable for treatment and control groups.

The science interest post scale had a range from 5 to 25 points, with an overall mean score of 16.51 ($SD = 4.83$). The data had a significant negative skew (skewness = ‑.27, $SE = .06$) and were significantly platykurtic (kurtosis = ‑.54, $SE = .12$). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1818) = .06$, $p = .000$. Cronbach's alpha for the 5-item scale was .89, and all items had a positive corrected item-total correlation. Mean scores for each item ranged from 3.70 (least difficult to endorse) to 2.77 (most difficult to endorse). Item mean scores are listed in Table 4. Data from 161 students (approximately 8%) were missing. Missing post score data was associated with students in schools with higher percentages of free and reduced lunch. Missing data was more common among urban students, eighth graders, male students, and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally common among treatment and control groups, $\chi^2(1) = 0.09$, $p = .77$.

Table 4

Mean Scores of Science Interest Scale Items

Item	Difficulty in pre scale	Difficulty in post scale
C1	3.82	3.70
C2	2.81	2.77
C3	3.55	3.42
C4	3.21	3.07
C5	3.66	3.56

Pre and post score data were MCAR with respect to each other (Little's MCAR test $\chi^2(1) = 2.70$, $p = .26$). Additionally, individual item data was MCAR (Little's MCAR test $\chi^2(182) = 118.77$, $p = 1.00$).

Rasch analysis. The science interest pre scale had person reliability of .87. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The pre scale first-contrast eigenvalue was 1.7, which met the requirement of being below 2.0. Item fit statistics are reported in Table 5. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. All categories had outfit mean-squares below 2.0 except for one: category 1 of item 1, "I enjoy going to science class," which had an outfit mean-square of 2.2. In general, this was a very easy item for students to endorse. A brief examination of unexpected responses to this item suggests that there are a certain number of students who like science but do not like their science class and used category 1 to indicate their dislike. Category statistics are reported in Table 6.

Table 5

Science Interest Item Fit Statistics

	Pre scale			Post scale		
Item	Infit mean square	Outfit mean square	Point-measure correlation	Infit mean square	Outfit mean square	Point-measure correlation
C1	1.10	1.11	.65	1.23	1.24	.66
C2	0.96	0.96	.70	0.96	0.94	.74
C3	0.78	0.77	.76	0.71	0.69	.81
C4	0.86	0.86	.74	0.81	0.81	.79
C5	1.31	1.33	.60	1.30	1.33	.66

Table 6

Science Interest Item Category Statistics

	Pre scale				Post scale			
Item	Observations	Average measures	Outfit mean square	Threshold (SE)	Observations	Average measures	Outfit mean square	Threshold (SE)
Item C1								
1	38	-1.38	2.2[a]		73	-1.69	1.9	
2	101	-0.97	1.0	-3.08 (.25)	145	-0.94	1.4	-2.97 (.19)
3	521	0.72	1.1	-1.69 (.12)	525	0.46	1.1	-1.56 (.10)
4	705	2.32	1.0	1.18 (.07)	636	2.19	1.0	1.09 (.07)
5	506	3.90	1.2	3.60 (.07)	474	3.65	1.4	3.43 (.08)
Item C2								
1	235	-3.07	1.2		312	-3.28	1.1	
2	539	-1.72	0.8	-3.23 (.09)	485	-1.77	0.7	-2.99 (.09)
3	617	-0.07	0.8	-0.99 (.07)	564	-0.15	0.8	-1.09 (.07)
4	311	1.38	0.9	1.26 (.08)	313	1.39	0.9	1.12 (.08)
5	171	2.35	1.1	2.96 (.12)	179	2.26	1.3	2.96 (.12)
Item C3								
1	82	-2.64	1.0		125	-2.99	0.7	
2	216	-1.17	0.7	-2.91 (.15)	274	-1.28	0.6	-2.97 (.13)
3	540	0.25	0.7	-1.26 (.09)	523	0.17	0.6	-1.11 (.08)
4	637	1.84	0.8	0.87 (.07)	556	1.80	0.6	0.90 (.07)
5	382	3.57	0.8	3.27 (.08)	368	3.55	0.9	3.18 (.08)
Item C4								
1	134	-3.08	0.9		199	-3.20	1.0	
2	293	-1.42	0.8	-2.99 (.12)	359	-1.66	0.7	-3.05 (.11)
3	665	-0.01	0.8	-1.51 (.08)	623	-0.11	0.7	-1.36 (.07)
4	516	1.68	0.8	1.03 (.07)	443	1.70	0.8	1.05 (.07)
5	228	3.02	1.1	3.47 (.10)	216	2.91	1.0	3.37 (.11)
Item C5								
1	87	-2.11	1.3		133	-2.14	1.5	
2	193	-0.62	1.3	-2.54 (.15)	231	-0.99	1.1	-2.50 (.13)
3	480	0.45	1.3	-1.15 (.09)	451	0.41	1.5	-1.04 (.09)
4	580	1.76	1.4	0.86 (.07)	545	1.65	1.2	0.76 (.07)
5	501	3.16	1.3	2.83 (.07)	482	3.12	1.4	2.77 (.08)

Note. No category difficulties changed significantly pre to post.

[a] Response category does not meet specified fit criteria.

The science interest post scale had person reliability of .88. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The post scale first-contrast eigenvalue was 1.4, which met the requirement of being below 2.0. Item fit statistics are reported in Table 4. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. All categories had outfit mean-squares below 2.0, including category 1 of item 1, which showed some misfit in the pre data. Category statistics are reported in Table 6.

Where the 95% confidence intervals around the pre and post difficulty and/or threshold measures of an item overlapped, the item was considered to be stable over time. All items had stable difficulty measures and threshold structures from pre to post; therefore, no stacking or anchoring was performed. Item difficulty measures are reported in Table 7.

Table 7

Science Interest Item Difficulty Statistics

	Pre scale		Post scale	
Item	Difficulty measure (SE)	95% CI	Difficulty measure (SE)	95% CI
C1	-1.16 (.04)	-1.24, -1.08	-1.07 (04)	-1.15, -0.99
C2	1.39 (.04)	1.31, 1.47	1.26 (.04)	1.18, 1.34
C3	-0.28 (.04)	-0.36, -0.20	-0.28 (.04)	-0.36, -0.20
C4	0.53 (.04)	0.45, 0.61	0.61 (.04)	0.53, 0.69
C5	-0.48 (.04)	-0.56, -0.40	-0.52 (.04)	-0.60, -0.44

Note. No item difficulties changed significantly pre to post.

Rasch person estimates for the science interest pre scale had a range from -6.04 logits to 6.3 logits, with an overall mean score of .98 (*SD* = 2.23). The data had no significant skew (skewness = .01, *SE* = .06) and were significantly leptokurtic (kurtosis = .43, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1872) = .06$, $p = .000$. Rasch person estimates for the science interest post scale had a range from -5.94 logits to 6.21 logits, with an overall mean score of .69 (*SD* = 2.42). The data had no significant skew (skewness = ‑.03, *SE* = .06) and were significantly leptokurtic (kurtosis = .24, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1853) = .05$, $p = .000$.

Science Career Interest

Observed score analysis. The science career interest pre scale had a range from 5 to 25 points, with an overall mean score of 11.02 (*SD* = 5.47). The data had a significant positive skew (skewness = .78, *SE* = .06) and were significantly platykurtic (kurtosis = ‑.24, *SE* = .12). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1818) = .14$, $p = .000$. Cronbach's alpha for the 5-item scale was .93, and all items had a positive corrected item-total correlation. Mean scores for each item ranged from 2.36 (least difficult to endorse) to 1.90 (most difficult to endorse). Item mean scores are listed in Table 7. Data from 161 students (approximately 8%) was missing. Missing pre score data was associated with students in schools with higher percentages of free and reduced lunch. Missing data was more common among younger students, urban students, and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables,

missing data were equally common among treatment and control groups, $\chi^2(1) = 1.40$, $p = .24$. Patterns of missing data were comparable for treatment and control groups.

The science career interest post scale had a range from 5 to 25 points, with an overall mean score of 11.28 ($SD = 5.65$). The data had a significant positive skew (skewness = .74, $SE = .06$) and were significantly platykurtic (kurtosis = -.37, $SE = .12$). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1823) = .13$, $p = .000$. Cronbach's alpha for the 5-item scale was .94, and all items had a positive corrected item-total correlation. Mean scores for each item ranged from 2.42 (least difficult to endorse) to 1.93 (most difficult to endorse). Item mean scores are listed in Table 8. Data from 156 students (approximately 8%) was missing. Missing post score data was associated with students in schools with higher percentages of free and reduced lunch. Missing data was more common among urban students, eighth graders, male students, and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally common among treatment and control groups, $\chi^2(1) = 0.01$, $p = .91$.

Table 8

Mean Scores of Science Career Interest Scale Items

Item	Difficulty in pre scale	Difficulty in post scale
C6	1.90	1.93
C7	2.31	2.36
C8	2.31	2.36
C9	2.13	2.21
C10	2.36	2.42

Pre and post score data were MCAR with respect to each other (Little's MCAR test $\chi^2(2) = 5.44$, $p = .07$). Additionally, individual item data was MCAR (Little's MCAR test $\chi^2(168) = 160.82$, $p = .64$).

Rasch analysis. The science career interest pre scale had person reliability of .85. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The pre scale first-contrast eigenvalue was 1.4, which met the requirement of being below 2.0. Item fit statistics are reported in Table 9. All other items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. All categories had outfit mean-squares below 2.0 except for two: category 5 of item 7, which had an outfit mean-square of 2.2 and category 5 of item 10, "I am interested in pursuing a college degree in science." A brief examination of unexpected responses suggests a possible reason for these category misfits: both of these items were presented

such that the science content of the item was on a second line; students may have read and agreed with, “When I graduate, I would like to work with people…” and “I am interested in pursuing a college degree…” rather than the complete science career items. Category statistics are reported in Table 10.

The science career interest post scale had person reliability of .87. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The post scale first-contrast eigenvalue was 1.4, which met the requirement of being below 2.0. Item fit statistics are reported in Table 8. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. All categories had outfit mean-squares below 2.0 except for two: category 5 of item 7 and category 5 of item 10, as reported for the pre data. Category statistics are reported in Table 10.

Table 9

Science Career Interest Item Fit Statistics

	Pre scale			Post scale		
Item	Infit mean square	Outfit mean square	Point-measure correlation	Infit mean square	Outfit mean square	Point-measure correlation
C6	0.97	0.94	.81	0.91	0.87	.84
C7	1.40	1.41	.75	0.71	0.70	.89
C8	0.77	0.79	.86	1.34	1.34	.79
C9	0.70	0.67	.87	0.80	0.85	.87
C10	1.14	1.15	.80	1.23	1.23	.80

Table 10

Science Career Interest Item Category Statistics

	Pre scale				Post scale			
Item	Observations	Average measures	Outfit mean square	Threshold (SE)	Observations	Average measures	Outfit mean square	Threshold (SE)
Item C6								
1	945	-3.83	1.3		915	-4.30	1.2	
2	423	-2.21	0.8	-2.67 (.08)	411	-2.46	0.7	-3.06 (.08)
3*	30	-0.45	0.7	-0.90 (.08)	304	-0.61	0.8	-1.20 (.09)
4	101	1.02	1.0	1.26 (.12)	123	1.17	0.8	1.09 (.12)
5*	88	2.05	1.1	2.31 (.19)	87	2.73	1.0	3.17 (.22)
Item C7								
1	605	-2.94	1.6		586	-3.36	1.4	
2	522	-2.01	1.1	-3.57 (.09)	497	-2.11	1.2	-3.75 (.09)
3	421	-0.14	1.2	-0.78 (.08)	409	-0.29	1.2	-1.00 (.08)
4	193	1.40	1.3	1.40 (.09)	224	1.63	1.3	1.24 (.10)
5	126	2.04	2.2 [a]	2.95 (.15)	132	2.43	2.3 [a]	3.50 (.15)
Item C8								
1	669	-3.37	0.9		636	-3.45	0.9	
2	474	-1.76	0.7	-3.01 (.09)	463	-1.95	0.8	-3.22 (.09)
3	360	-0.07	0.7	-0.62 (.08)	376	-0.10	0.8	-0.80 (.08)
4	205	1.37	0.8	1.12 (.09)	198	1.53	0.7	1.30 (.10)
5	159	2.71	1.0	2.51 (.13)	179	3.27	1.0	2.72 (.13)
Item C9								
1	771	-3.80	0.7		729	-3.85	0.8	
2	460	-1.82	0.5	-2.88 (.08)	451	-2.02	0.5	-3.07 (.08)
3	351	-0.12	0.5	-0.78 (.08)	359	-0.09	0.6	-0.89 (.08)
4	156	1.23	0.8	1.22 (.10)	161	1.52	0.7	1.33 (.10)
5	122	2.40	1.2	2.45 (.15)	150	2.87	1.4	2.62 (.15)
Item C10								
1	633	-3.22	1.0		611	-3.28	1.2	
2	437	-1.71	0.9	-3.02 (.09)	416	-1.91	0.9	-3.20 (.09)
3	419	-0.02	0.9	-0.85 (.08)	412	-0.11	1.1	-1.07 (.08)
4	202	1.40	1.2	1.34 (.09)	244	1.47	1.3	1.16 (.09)
5*	164	2.23	2.2 [a]	2.54 (.13)	169	2.77	2.2 [a]	3.11 (.13)

*Difficulty changed significantly pre to post.

[a] Response category does not meet specified fit criteria.

Where the 95% confidence intervals around the pre and post difficulty and/or threshold measures of an item overlapped, the item was considered to be stable over time. Item 6 had a significantly different difficulty in the post data than in the pre data, as reported in Table 11. Therefore, a stacking and anchoring procedure was performed similar to the one recommended by Wolfe and Chiu (1999, using half pre data and half post data to avoid local dependence (Mallinson, 2001). First, a stacked dataset was used to create stable category structure measures. These measures were used as an anchor for the pre data, from which pre person measures and item difficulties were estimated. The item difficulties were used to anchor the invariant items in the post data, from which post person measures were estimated.

Table 11

Science Career Interest Item Difficulty Statistics

	Pre scale		Post scale		Anchored
Item	Difficulty measure (SE)	95% CI	Difficulty measure (SE)	95% CI	difficulty measure
C6*	0.92 (.04)	0.84, 1.00	1.13 (.05)	1.03, 1.23	1.01
C7	-0.28 (.04)	-0.36, -0.20	-0.26 (.04)	-0.34, -0.18	-0.29
C8	-0.34 (.04)	-0.42, -0.26	-0.39 (.04)	-0.47, -0.31	-0.37
C9	0.18 (.04)	0.10, 0.26	0.06 (.04)	-0.02, 0.14	0.16
C10	-0.48 (.04)	-0.56, -0.40	-0.53 (.04)	-0.61, -0.45	-0.51

*Difficulty changed significantly pre to post.

Rasch person estimates for the science career interest pre scale had a range from -6.28 logits to 5.67 logits, with an overall mean score of -2.12 (*SD* = 3.08). The data had a significant positive skew (skewness = .32, *SE* = .06) and were significantly platykurtic (kurtosis = ‑.55, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, *D*(1872) = .11, *p* = .000. Rasch person estimates for the science career interest post scale had a range from -6.21 logits to 6.07 logits, with an overall mean score of -1.96 (*SD* = 3.18). The data had a significant positive skew (skewness = .41, *SE* = .06) and were significantly platykurtic (kurtosis = ‑.37, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, *D*(1854) = .11, *p* = .000.

Social Studies Interest

Observed score analysis. The social studies interest pre scale had a range from 4 to 20 points, with an overall mean score of 15.07 (*SD* = 3.33). The data had a significant negative skew (skewness = ‑.53, *SE* = .06) and no significant kurtosis (kurtosis = ‑.19, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, *D*(1842) = .10, *p* = .000. Cronbach's alpha for the 4-item scale was .77, and all items had a positive corrected item-total correlation. Mean scores for each item ranged from 4.05 (least difficult to endorse) to 3.14 (most difficult to endorse). Item mean scores are listed in Table 11. Data from 137 students (approximately 7%) was missing. Missing pre score data was associated with students in schools with higher percentages of free and reduced lunch. Missing data was more common among younger students, urban students, seventh graders, and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were

equally common among treatment and control groups, $\chi^2(1) = 0.04$, $p = .85$. Patterns of missing data were comparable for treatment and control groups.

The social studies interest post scale had a range from 4 to 20 points, with an overall mean score of 14.64 ($SD = 3.64$). The data had a significant negative skew (skewness = ‑.46, $SE = .06$) and were significantly platykurtic (kurtosis = ‑.27, $SE = .11$). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1829) = .09$, $p = .000$. Cronbach's alpha for the 4-item scale was .81, and all items had a positive corrected item-total correlation. Mean scores for each item ranged from 3.90 (least difficult to endorse) to 3.07 (most difficult to endorse). Item mean scores are listed in Table 12. Data from 150 students (approximately 8%) was missing. Missing post score data was associated with students in schools with higher percentages of free and reduced lunch. Missing data was more common among urban students, eighth graders, male students, and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally common among treatment and control groups, $\chi^2(1) = 0.16$, $p = .69$.

Table 12

Mean Scores of Social Studies Interest Scale Items

Item	Difficulty in pre scale	Difficulty in post scale
C11	4.05	3.90
C12	4.00	3.88
C13	3.88	3.79
C14	3.14	3.07

Pre and post score data were MCAR with respect to each other (Little's MCAR test $\chi^2(2) = 2.05$, $p = .359$). Additionally, individual item data was MCAR (Little's MCAR test $\chi^2(105) = 87.39$, $p = .89$).

Rasch analysis. The social studies interest pre scale had person reliability of .75. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The pre scale first-contrast eigenvalue was 1.4, which met the requirement of being below 2.0. Item fit statistics are reported in Table 13. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations.

Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. All categories had outfit mean-squares below 2.0 except for one: category 1 of item 11, "I like my social studies class," which had an outfit mean-square of 2.2. Category statistics are reported in Table 14.

The social studies interest post scale had person reliability of .78. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The post scale first-contrast eigenvalue was 1.4, which met the requirement of being below 2.0. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds, and outfit mean-squares below 2.0. Category statistics are reported in Table 14.

Table 13

Social Studies Interest Item Fit Statistics

	Pre scale			Post scale		
Item	Infit mean square	Outfit mean square	Point-measure correlation	Infit mean square	Outfit mean square	Point-measure correlation
C11	1.25	1.30	.46	1.13	1.17	.57
C12	0.88	0.86	.62	0.89	0.88	.66
C13	0.92	0.90	.61	1.01	0.98	.62
C14	0.95	0.94	.59	0.96	0.96	.64

Note. All items met specified fit criteria.

Table 14

Social Studies Interest Item Category Statistics

	Pre scale				Post scale			
Item	Observations	Average measures	Outfit mean square	Threshold (SE)	Observations	Average measures	Outfit mean square	Threshold (SE)
Item C11								
1	32	-0.23	2.5 [a]		62	-0.84	1.6	
2	103	-0.15	1.3	-1.98 (.21)	126	-0.56	1.0	-1.82 (.17)
3	331	0.68	1.2	-0.97 (.11)	403	0.50	1.1	-1.13 (.10)
4	684	1.77	1.2	0.45 (.07)	601	1.61	1.3	0.58 (.07)
5	718	2.72	1.2	2.50 (.06)	657	2.60	1.2	2.37 (.07)
Item C12								
1	40	-1.03	1.3		64	-1.54	0.9	
2	124	-0.52	0.7	-1.95 (.19)	136	-0.51	0.9	-1.88 (.17)
3	342	0.51	0.7	-0.86 (.10)	390	0.40	0.7	-1.05 (.10)
4	641	1.66	0.9	0.47 (.07)	618	1.53	0.8	0.50 (.07)
5	712	2.81	1.0	2.34 (.06)	632	2.78	1.0	2.43 (.07)
Item C13								
1	61	-1.34	1.0		86	-1.41	1.1	
2	149	-0.50	0.9	-1.82 (.16)	159	-0.61	0.9	-1.76 (.14)
3	401	0.39	0.7	-0.96 (.09)	408	0.38	0.9	-1.04 (.09)
4	608	1.50	0.9	0.53 (.07)	600	1.36	0.9	0.45 (.07)
5	640	2.59	1.0	2.25 (.07)	596	2.58	1.1	2.35 (.07)
Item C14								
1	244	-1.98	1.1		285	-2.12	1.0	
2	330	-1.14	0.7	-1.82 (.09)	323	-1.21	0.9	-1.80 (.08)
3	540	-0.20	0.7	-1.10 (.07)	545	-0.25	0.8	-1.23 (.07)
4	440	0.97	0.8	0.52 (.07)	377	0.87	1.0	0.63 (.07)
5	314	1.59	1.4	2.41 (.10)	321	1.77	1.1	2.40 (.11)

Note. No category thresholds changed significantly pre to post.

[a] Response category does not meet specified fit criteria.

Where the 95% confidence intervals around the pre and post difficulty and/or threshold measures of an item overlapped, the item was considered to be stable over time. All items had stable difficulty measures and threshold structures from pre to post; therefore, no stacking or anchoring was performed. Item difficulty measures are reported in Table 15.

Rasch person estimates for the social studies interest pre scale had a range from -7.00 logits to 5.21 logits, with an overall mean score of 1.45 (*SD* = 1.85). The data had a significant positive skew (skewness = .28, *SE* = .06) and no significant kurtosis (kurtosis = .01, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1871) = .11$, $p = .000$. Rasch person estimates for the social studies interest post scale had a range from -4.59 logits to 5.19 logits, with an overall mean score of 1.27 (*SD* = 2.01). The data had a significant positive skew (skewness = .28, *SE* = .06) and no significant kurtosis (kurtosis = ‑.08, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1853) = .11$, $p = .000$.

Table 15

Social Studies Interest Item Difficulty Statistics

Item	Pre difficulty measure (SE)	95% CI	Post difficulty measure (SE)	95% CI
C11	-0.62 (.04)	-0.70, -0.54	-0.51 (.04)	-0.59, -0.43
C12	-0.48 (.04)	-0.56, -0.40	-0.46 (.04)	-0.54, -0.38
C13	-0.17 (.03)	-0.23, -0.11	-0.23 (.03)	-0.29, -0.17
C14	1.27 (.03)	1.21, 1.33	1.20 (.03)	1.14, 1.26

Note. No item difficulties changed significantly pre to post.

Writing Self-efficacy

Observed score analysis. The writing self-efficacy pre scale had a range from 5 to 25 points, with an overall mean score of 17.59 (*SD* = 4.09). The data had a significant negative skew (skewness = ‑.50, *SE* = .06) and no significant kurtosis (kurtosis = .16, *SE* = .12). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1809) = .08$, $p = .000$. Cronbach's alpha for the 5-item scale was .84, and all items had a positive corrected item-total correlation. Mean scores for each item ranged from 3.76 (least difficult to endorse) to 3.35 (most difficult to endorse). Item mean scores are listed in Table 16. Data from 170 students (approximately 9%) was missing. Missing pre score data was associated with students in schools with higher percentages of free and reduced lunch. Missing data was more common among younger students, urban students, and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally common among

treatment and control groups, $\chi^2(1) = 3.54$, $p = .06$. Patterns of missing data were comparable for treatment and control groups.

The writing self-efficacy post scale had a range from 5 to 25 points, with an overall mean score of 17.68 ($SD = 4.46$). The data had a significant negative skew (skewness = -.51, $SE = .06$) and no significant kurtosis (kurtosis = .02, $SE = .12$). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1797) = .08$, $p = .000$. Cronbach's alpha for the 5-item scale was .87, and all items had a positive corrected item-total correlation. Mean scores for each item ranged from 3.77 (least difficult to endorse) to 3.34 (most difficult to endorse). Item mean scores are listed in Table 16. Data from 182 students (approximately 9%) was missing. Missing post score data was associated with students in schools with higher percentages of free and reduced lunch. Missing data was more common among younger students, urban students, male students, and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally common among treatment and control groups, $\chi^2(1) = 0.001$, $p = .97$.

Table 16

Mean Scores of Writing Self-efficacy Scale Items

Item	Difficulty in pre scale	Difficulty in post scale
D1	3.47	3.46
D2	3.46	3.51
D3	3.55	3.60
D4	3.35	3.34
D5	3.76	3.77

Pre and post score data were MCAR with respect to each other (Little's MCAR test $\chi^2(2) = 1.73$, $p = .42$). Additionally, individual item data was MCAR (Little's MCAR test $\chi^2(229) = 259.42$, $p = .08$).

Rasch analysis. The writing self-efficacy pre scale had person reliability of .84. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The pre scale first-contrast eigenvalue was 1.6, which met the requirement of being below 2.0. Item fit statistics are reported in Table 17. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. All categories had outfit mean-squares below 2.0. Category statistics are reported in Table 18.

The writing self-efficacy post scale had person reliability of .85. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The post scale first-contrast eigenvalue was 1.7, which met the requirement of being below 2.0. Item fit statistics are reported in Table 17. Category

structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. All categories had outfit mean-squares below 2.0. Category statistics are reported in Table 18.

Table 17

Writing Self-efficacy Item Fit Statistics

	Pre scale			Post scale		
Item	Infit mean square	Outfit mean square	Point-measure correlation	Infit mean square	Outfit mean square	Point-measure correlation
D1	0.89	0.89	.68	0.83	0.83	.75
D2	0.81	0.81	.72	0.77	0.77	.77
D3	0.88	0.88	.68	0.94	0.93	.71
D4	1.31	1.31	.54	1.36	1.38	.59
D5	1.11	1.11	.60	1.09	1.05	.66

Note. All items met specified fit criteria.

Table 18

Writing Self-efficacy Item Category Statistics

	Pre scale				Post scale			
Item	Observations	Average measures	Outfit mean square	Threshold (SE)	Observations	Average measures	Outfit mean square	Threshold (SE)
Item D1								
1	86	-2.10	1.3		106	-2.28	1.0	
2	201	-1.18	0.8	-2.60 (.15)	190	-1.23	0.8	-2.52 (.14)
3	605	0.15	0.8	-1.47 (.08)	601	.02	0.7	-1.63 (.09)
4	707	1.58	0.8	0.67 (.06)	624	1.71	0.7	0.79 (.06)
5	271	2.89	1.0	3.40 (.09)	315	3.09	1.0	3.36 (.09)
Item D2								
1	75	-2.32	1.1		93	-2.25	0.9	
2	222	-1.13	0.7	-2.87 (.16)	203	-1.22	0.8	-2.70 (.15)
3	611	0.20	0.7	-1.33 (.08)	543	.09	0.6	-1.40 (.09)
4	674	1.65	0.8	0.79 (.06)	664	1.71	0.7	0.67 (.06)
5	270	3.02	0.9	3.41 (.09)	328	3.28	0.9	3.42 (.09)
Item D3								
1	70	-2.13	1.0		92	-1.95	1.3	
2	196	-0.93	0.8	-2.69 (.16)	180	-1.04	0.8	-2.47 (.15)
3	578	0.35	0.8	-1.28 (.09)	524	.26	0.8	-1.37 (.09)
4	662	1.64	0.9	0.84 (.06)	622	1.69	0.9	0.77 (.07)
5	345	3.07	1.0	3.13 (.08)	409	3.14	1.0	3.07 (.08)
Item D4								
1	125	-1.81	1.5		147	-1.94	1.6	
2	296	-0.82	1.3	-2.48 (.12)	268	-.89	1.2	-2.43 (.12)
3	558	0.23	1.2	-1.00 (.07)	572	.08	1.4	-1.25 (.08)
4	567	1.25	1.2	0.70 (.06)	498	1.47	1.3	0.88 (.07)
5	309	2.35	1.4	2.78 (.08)	339	2.35	1.5	2.81 (.09)
Item D5								
1	59	-1.59	1.3		76	-1.62	1.4	
2	168	-0.60	1.0	-2.52 (.18)	151	-.83	0.9	-2.38 (.17)
3	448	0.55	1.0	-0.99 (.09)	431	.45	1.0	-1.19 (.10)
4	689	1.78	1.0	0.68 (.06)	652	1.83	0.9	0.64 (.07)
5	499	2.95	1.2	2.82 (.07)	519	3.14	1.3	2.93 (.07)

Note. No category difficulties changed significantly pre to post. All categories met specified fit criteria.

Where the 95% confidence intervals around the pre and post difficulty and/or threshold measures of an item overlapped, the item was considered to be stable over time. All items had stable difficulty measures and threshold structures from pre to post; therefore, no stacking or anchoring was performed. Item difficulty measures are reported in Table 19.

Table 19

Writing Self-efficacy Item Difficulty Statistics

Item	Pre difficulty measure (SE)	95% CI	Post difficulty measure (SE)	95% CI
D1	0.13 (.04)	0.05, 0.21	0.17 (.04)	0.09, 0.25
D2	0.11 (.04)	0.03, 0.19	0.05 (.04)	-0.03, 0.13
D3	-0.11 (.04)	-0.19, -0.03	-0.14 (.04)	-0.22, -0.06
D4	0.39 (.03)	0.33, 0.45	0.44 (.04)	0.36, 0.52
D5	-0.52 (.04)	-0.60, -0.44	-0.52 (.04)	-0.60, -0.44

Note. No item difficulties changed significantly pre to post.

Rasch person estimates for the writing self-efficacy pre scale had a range from -5.46 logits to 5.95 logits, with an overall mean score of 1.05 (SD = 1.95). The data had no significant skew (skewness = .07, SE = .06) and were significantly leptokurtic (kurtosis = .79, SE = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1872) = .08$, $p = .000$. Rasch person estimates for the writing self-efficacy post scale had a range from -5.33 logits to 5.95 logits, with an

overall mean score of 2.56 (*SD* = 2.07). The data had no significant skew (skewness = .03, *SE* = .06) and were significantly leptokurtic (kurtosis = .41, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1829) = .08$, $p = .000$.

Socio-science Self-efficacy

Observed score analysis. The socio-science self-efficacy pre scale had a range from 5 to 25 points, with an overall mean score of 20.77 (*SD* = 3.54). The data had a significant negative skew (skewness = ‑.92, *SE* = .06) and were significantly leptokurtic (kurtosis = .80, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1834) = .12$, $p = .000$. Cronbach's alpha for the 5-item scale was .84, and all items had a positive corrected item-total correlation. Mean scores for each item ranged from 4.37 (least difficult to endorse) to 3.85 (most difficult to endorse). Item mean scores are listed in Table 20. Data from 145 students (approximately 7%) was missing. Missing pre score data was associated with students in schools with higher percentages of free and reduced lunch. Missing data was more common among younger students, urban students, male students, and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally common among treatment and control groups, $\chi^2(1) = 1.51$, $p = .22$. Patterns of missing data were comparable for treatment and control groups.

The socio-science self-efficacy post scale had a range from 5 to 25 points, with an overall mean score of 20.19 (*SD* = 3.88). The data had a significant negative skew (skewness = ‑.80, *SE* = .06) and were significantly leptokurtic (kurtosis = .43, *SE* = .12).

The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1812) = .11$, $p = .000$. Cronbach's alpha for the 5-item scale was .84, and all items had a positive corrected item-total correlation. Mean scores for each item ranged from 4.24 (least difficult to endorse) to 3.71 (most difficult to endorse). Item mean scores are listed in Table 20. Data from 167 students (approximately 8%) was missing. Missing post score data was associated with students in schools with higher percentages of free and reduced lunch. Missing data was more common among urban students, male students, and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally common among treatment and control groups, $\chi^2(1) = 0.001$, $p = .97$.

Table 20

Mean Scores of Socio-science Self-efficacy Scale Items

Item	Difficulty in pre scale	Difficulty in post scale
D6	4.28	4.19
D7	4.37	4.24
D8	4.04	3.94
D9	4.22	4.10
D10	3.85	3.71

Pre and post score data were MCAR with respect to each other (Little's MCAR test $\chi^2(2) = 2.31$, $p = .31$). Additionally, individual item data was MCAR (Little's MCAR test $\chi^2(217) = 236.63$, $p = .17$).

Rasch analysis. The socio-science self-efficacy pre scale had person reliability of .77. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The pre scale first-contrast eigenvalue was 1.7, which met the requirement of being below 2.0. Item fit statistics are reported in Table 21. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. All categories had outfit mean-squares below 2.0 except for one: category 1 of item 7, "[How confident are you that you can…] get a good grade in social studies," which had an outfit mean-square of 2.4. Category statistics are reported in Table 22.

The socio-science self-efficacy post scale had person reliability of .76. All items had acceptable fit statistics (infit and outfit mean-squares below 1.5) and positive point-measure correlations. The post scale first-contrast eigenvalue was 1.7, which met the requirement of being below 2.0. Item fit statistics are reported in Table 21. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. Category structure met the requirements of 10 or more observations in each category, ordered average measures, and ordered thresholds. All categories had outfit mean-squares below 2.0. Category statistics are reported in Table 22.

Table 21

Socio-science Self-efficacy Item Fit Statistics

	Pre scale			Post scale		
Item	Infit mean square	Outfit mean square	Point-measure correlation	Infit mean square	Outfit mean square	Point-measure correlation
D6	0.96	0.93	.67	0.98	0.96	.64
D7	0.94	0.92	.66	0.94	0.93	.65
D8	0.91	0.90	.70	0.91	0.90	.68
D9	0.90	0.85	.69	0.85	0.80	.70
D10	1.28	1.29	.58	1.31	1.32	.55

Note. All items met specified fit criteria.

Table 22

Socio-science Self-efficacy Item Category Statistics

	Pre scale				Post scale			
Item	Observations	Average measures	Outfit mean square	Threshold (SE)	Observations	Average measures	Outfit mean square	Threshold (SE)
Item D6								
1	15	-1.15	1.2		33	-0.64	1.3	
2	59	-0.25	1.2	-2.36 (.31)	81	-0.08	1.1	-1.63 (.22)
3	247	0.68	0.9	-1.21 (.15)	295	0.52	0.9	-1.07 (.12)
4	626	2.20	0.8	0.54 (.08)	526	1.66	0.9	0.53 (.07)
5*	921	3.88	1.0	3.03 (.07)	897	3.10	1.0	2.17 (.07)
Item D7								
1	12	-0.67	2.4[a]		29	-0.54	1.6	
2	40	-0.22	1.0	-2.15 (.36)	60	-0.22	0.8	-1.47 (.24)
3	194	0.69	0.8	-1.28 (.17)	252	0.49	0.8	-1.20 (.13)
4	632	2.30	0.9	0.37 (.09)	596	1.76	0.9	0.28 (.08)
5*	988	4.05	1.0	3.06 (.07)	891	3.20	1.0	2.39 (.07)
Item D8								
1	27	-1.68	1.2		46	-1.16	1.1	
2	92	-0.83	0.9	-2.50 (.23)	115	-0.58	0.8	-1.91 (.18)
3	337	0.43	0.8	-1.41 (.12)	406	0.30	0.8	-1.33 (.10)
4	732	2.07	0.8	0.47 (.07)	614	1.63	0.9	0.54 (.07)
5*	676	3.55	1.0	3.44 (.07)	654	2.92	1.0	2.70 (.07)
Item D9								
1	18	-1.68	0.8		36	-1.15	0.9	
2	50	-0.54	1.0	-2.15 (.29)	69	-0.43	0.9	-1.54 (.21)
3	244	0.48	0.7	-1.52 (.15)	305	0.30	0.7	-1.43 (.12)
4	740	2.22	0.7	0.28 (.08)	684	1.70	0.7	0.23 (.07)
5*	802	3.85	1.0	3.39 (.07)	733	3.16	1.0	2.75 (.07)
Item D10								
1	52	-1.53	1.7		85	-1.21	1.5	
2	119	-0.93	1.0	-2.30 (.17)	156	-0.70	1.3	-1.86 (.13)
3	471	0.40	1.2	-1.69 (.10)	491	0.15	1.2	-1.52 (.09)
4	634	1.94	1.2	0.75 (.07)	593	1.49	1.2	0.53 (.07)
5*	592	2.68	1.4	3.24 (.08)	512	2.05	1.5	2.85 (.08)

*Difficulty changed significantly pre to post.

[a] Response category does not meet specified fit criteria.

Where the 95% confidence intervals around the pre and post difficulty and/or threshold measures of an item overlapped, the item was considered to be stable over time. Item 7 had a significantly different difficulty in the post data than in the pre data, as reported in Table 23. Therefore, a stacking and anchoring procedure was performed similar to the one recommended by Wolfe and Chiu (1999). First, a stacked dataset was used to create stable category structure measures. These measures were used as an anchor for the pre data, from which pre person measures and item difficulties were estimated. The item difficulties were used to anchor the invariant items in the post data, from which post person measures were estimated.

Table 23

Socio-science Self-efficacy Item Difficulty Statistics

	Pre scale		Post scale		Anchored
Item	Difficulty measure (SE)	95% CI	Difficulty measure (SE)	95% CI	difficulty measure
D6	-0.36 (.04)	-0.44, -0.28	-0.34 (.04)	-0.42, -0.26	-0.26
D7*	-0.63 (.05)	-0.73, -0.53	-0.46 (.04)	-0.54, -0.38	-0.60
D8	0.31 (.04)	0.23, 0.39	0.19 (.04)	0.11, 0.27	0.28
D9	-0.19 (.04)	-0.27, -0.11	-0.15 (.04)	-0.23, -0.07	-0.19
D10	0.86 (.04)	0.78, 0.94	0.76 (.04)	0.68, 0.84	0.77

*Difficulty changed significantly pre to post.

Rasch person estimates for the socio-science self-efficacy pre scale had a range from -4.80 logits to 5.88 logits, with an overall mean score of 2.56 (*SD* = 2.07). The data had no significant skew (skewness = .04, *SE* = .06) and were significantly platykurtic (kurtosis = ‑.73, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1872) = .10$, $p = .000$. Rasch person estimates for the socio-science self-efficacy post scale had a range from -4.52 logits to 5.56 logits, with an overall mean score of 2.13 (*SD* = 2.03). The data had a significant positive skew (skewness = .14, *SE* = .06) and were significantly platykurtic (kurtosis = ‑.55, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1837) = .10$, $p = .000$.

Writing

Observed score analysis. The writing pre scores had a range from 0 to 15.5 points (the scale could award up to 17 points, but no essays were awarded the full points), with an overall mean score of 6.10 (*SD* = 2.27). The data had a significant positive skew (skewness = .61, *SE* = .06) and were significantly leptokurtic (kurtosis = .36, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1897) = .09$, $p = .000$. Mean scores for each category are listed in Table 24.

Data from 82 students (approximately 4%) was missing. Missing pre score data was associated with students in schools with higher percentages of free and reduced lunch, and a lower score on the post assessment. Missing data was more common among urban and non-White students. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally

common among treatment and control groups, $\chi^2(1) = 3.81$, $p = .05$. Patterns of missing data were comparable for treatment and control groups.

The writing post scores had a range from 0 to 15.5 points, with an overall mean score of 6.88 (*SD* = 2.37). The data had a significant positive skew (skewness = .19, *SE* = .06) and no significant kurtosis (kurtosis = -.05, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1881) = .08$, $p = .000$. Data from 98 students (5%) was missing. Mean scores for each category are listed in Table 24.

Table 24

Mean Scores of Writing Categories

Item	Difficulty in pre scale	Difficulty in post scale
Claim (0 – 2)	1.26	1.30
Evidence (0 – 3)	1.12	1.26
Reasoning (0 – 2)	0.45	0.52
Addressing the Opposition (0 – 2)	0.20	0.24
Organization (0 – 2)	1.24	1.27
Science Content (0 – 3)	1.10	1.27
Social Studies Content (0 – 3)	0.73	1.02

Data from 98 students (5%) was missing. Missing pre score data was associated with students in schools with higher percentages of free and reduced lunch, and a lower score on the pre assessment. Missing data was more common among urban students, non-

White students, male students, and eighth graders. Although data were not missing completely at random (MCAR) with respect to several demographic variables, missing data were equally common among treatment and control groups, $\chi^2(1) = .09$, $p = .77$. Patterns of missing data were comparable for treatment and control groups.

Rasch analysis. The writing pre data had person reliability of .83. All traits had acceptable infit and outfit mean-squares between 0.5 and 1.5 with the exception of one, Opposition, which had an outfit mean-square of 2.91. Trait fit statistics are presented in Table 25. All raters had acceptable infit mean-squares between 0.5 and 1.5 (Englehard, 1994). Four raters had high outfit mean-squares between 1.5 and 2.0, and one, Rater 14, had an outfit mean-square of 2.88. Rater fit statistics are presented in Table 26. Rater severities on the pre ranged from -0.68 logits (*SE* = .03) (Rater 6, easiest rater) to 1.32 logits (*SE* = .06) (Rater 4, hardest rater). This is a significant difference in severity across raters, $\chi^2(13) = 1199.4$, $p < .001$. Rater severity statistics are presented in Table 27.

Category structures were estimated for each rater individually for a total of 98 separate rating scales; therefore, category structure statistics are not reported exhaustively. Restriction of range and central tendencies were evident for several raters in several traits, including Claim, where level 0 was used only 2% of the time; Evidence, where level 3 was used only 2% of the time; Opposition, where level 0 was used 85% of the time; and Science Content and Social Studies Content, where level 3 was used 4% and 3% of the time, respectively.

The writing post data had person reliability of .83. All traits had acceptable infit and outfit mean-squares between 0.5 and 1.5 with the exception of one, Opposition, which had an outfit mean-square of 2.15. Trait fit statistics are presented in Table 25. All

raters had acceptable infit mean-squares between 0.5 and 1.5. Two raters had high outfit mean-squares between 1.5 and 2.0. Rater fit statistics are presented in Table 26. Rater severities on the pre ranged from -0.82 logits (*SE* = .03) (Rater 6, easiest rater) to 1.45 logits (*SE* = .05) (Rater 4, hardest rater). This is a significant difference in severity across raters, $\chi^2(13) = 2243.8$, $p < .001$. Rater severity statistics are presented in Table 27.

Where the 95% confidence intervals around the pre and post difficulty/severity measures of a trait or rater overlapped, the trait or rater was considered to be stable over time. Two traits and seven raters had significantly different difficulty measures in the pre and post scales. Therefore, a stacking and anchoring procedure was performed similar to the one recommended by Wolfe and Chiu (1999). First, a stacked dataset including data from all participants at both pre and post time points was used to create stable category structure measures. These measures were used as an anchor for the pre data, from which pre person measures, rater severity, and trait difficulties were estimated. The rater severity and trait difficulties were used to anchor the invariant items in the post data, from which post person measures were estimated. Trait difficulties are presented in Table 28.

Rasch person estimates for the writing pre data had a range from -8.04 logits to 2.48 logits, with an overall mean score of -.48 (*SD* = 1.21). The data had a significant negative skew (skewness = ‑1.72, *SE* = .06) and were significantly leptokurtic (kurtosis = 7.29, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1873) = .08$, $p = .000$. Rasch person estimates for the writing post data had a range from -7.86 logits to 2.71 logits, with an overall mean score of -.72 (*SD* = 1.25). The data had a significant negative skew (skewness = ‑1.50,

SE = .06) and were significantly leptokurtic (kurtosis = 5.16, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1888) = .07$, $p = .000$.

Table 25

Writing Trait Item Fit Statistics

	Pre scale			Post scale		
Trait	Infit mean square	Outfit mean square	Point-measure correlation	Infit mean square	Outfit mean square	Point-measure correlation
Claim	1.18	1.27	.14	1.18	1.25	.19
Evidence	0.91	0.90	.30	0.97	0.97	.40
Reasoning	1.04	0.97	.32	1.02	1.00	.30
Addressing Opposition	1.29	2.91[a]	.42	1.29	2.15[a]	.15
Organization	1.09	1.02	.39	1.13	1.14	.35
Science Content	1.03	1.18	.37	0.95	0.93	.41
Social Studies Content	1.20	1.40	.22	1.24	1.30	.33

[a] Trait does not meet specified fit criteria

Table 26

Rater Fit Statistics

	Pre scale			Post scale		
Rater	Infit mean square	Outfit mean square	Point-measure correlation	Infit mean square	Outfit mean square	Point-measure correlation
1	1.22	1.26	.39	1.06	1.09	.40
2	0.93	1.03	.39	0.94	0.92	.39
3	1.11	1.18	.42	1.14	1.21	.42
4	1.30	1.33	.37	1.22	1.11	.39
5	0.99	1.07	.41	0.93	0.95	.42
6	1.22	1.65*	.34	1.25	1.76 [a]	.34
7	1.15	1.26	.41	1.15	1.28	.39
8	1.13	1.57 [a]	.41	1.10	1.15	.40
9	0.80	0.70	.46	1.02	1.23	.40
10	1.47	1.72 [a]	.47	1.05	1.07	.43
11	1.02	1.05	.44	0.98	0.95	.44
12	1.07	1.04	.42	1.26	1.43	.38
13	1.09	1.55 [a]	.38	1.08	1.20	.37
14	1.17	2.88 [a]	.37	1.21	1.78 [a]	.35

[a] Rater fit does not meet specified fit criteria

Table 27

Rater Severity Statistics

	Pre scale		Post scale		Anchored
Rater	Difficulty measure (SE)	95% CI	Difficulty measure (SE)	95% CI	rater difficulty measure
1	0.04 (.06)	-0.08, 0.16	0.05 (.06)	-0.07, 0.17	0.13
2*	-0.22 (.03)	-0.28, -0.16	-0.53 (.03)	-0.59, -0.47	-0.29
3*	-0.29 (.04)	-0.37, -0.21	0.10 (.04)	0.02, 0.18	-0.11
4	1.32 (.06)	1.20, 1.44	1.45 (.05)	1.35, 1.55	0.45
5	0.18 (.06)	0.06, 0.30	-0.08 (.05)	-0.18, 0.02	0.10
6*	-0.68 (.03)	-0.74, -0.62	-0.82 (.03)	-0.88, -0.76	-0.70
7	-0.58 (.04)	-0.66, -0.50	-0.55 (.04)	-0.63, -0.47	-0.60
8	-0.08 (.03)	-0.14, -0.02	-0.09 (.03)	-0.15, -0.03	-0.13
9*	0.12 (.06)	0.00, 0.24	0.39 (.05)	0.29, 0.49	0.43
10*	0.53 (.12)	0.29, 0.77	-0.14 (.12)	-0.38, 0.10	-0.29
11*	-0.12 (.05)	-0.22, -0.02	-0.29 (.04)	-0.37, -0.21	-0.10
12	-0.04 (.05)	-0.14, 0.06	0.13 (.05)	0.03, 0.23	0.09
13*	0.08 (.05)	-0.02, 0.18	0.45 (.05)	0.35, 0.55	0.20
14	-0.23 (.04)	-0.31, -0.15	-0.08 (.04)	-0.16, 0.00	-0.18

*Rater severity changed significantly pre to post

Table 28

Writing Trait Difficulty Statistics

	Pre scale		Post scale		Anchored
Trait	Difficulty measure (SE)	95% CI	Difficulty measure (SE)	95% CI	difficulty measure
Claim	-2.28 (.04)	-2.36, -2.20	-2.29 (.04)	-2.37, -2.21	-2.15
Evidence	0.43 (.03)	0.37, 0.49	0.45 (.03)	0.39, 0.51	0.49
Reasoning	0.92 (.03)	0.86, 0.98	0.98 (.03)	0.92, 1.04	0.74
Addressing Opposition	1.63 (.03)	1.57, 1.69	1.64 (.03)	1.58, 1.70	1.40
Organization	-2.03 (.04)	-2.11, -1.95	-2.08 (.04)	-2.16, -2.00	-1.90
Science Content*	0.24 (.03)	0.18, 0.30	0.47 (.03)	0.41, 0.53	0.35
Social Studies Content*	1.09 (.03)	1.03, 1.15	0.82 (.02)	0.78, 0.86	1.06

*Trait difficulty changed significantly pre to post

Revised Rasch analysis. In order to improve the model fit, several revised models were tested iteratively. The final revised model recoded the problematic Opposition category from a three-category rating scale into a two-category rating scale and removed Opposition scores from Raters 6 and 13. The revised model fit and difficulty statistics are provided below.

The revised writing pre data had person reliability of .83. All traits had acceptable infit and outfit mean-squares between 0.5 and 1.5 with the exception of one, Opposition, which had an outfit mean-square of 1.76. Trait fit statistics are presented in Table 29. All raters had acceptable infit mean-squares between 0.5 and 1.5. Two raters had high outfit mean-squares between 1.5 and 2.0. Rater fit statistics are presented in Table 30. Rater severities on the pre ranged from -0.73 logits (*SE* = .03) (Rater 6, easiest rater) to 1.34

logits (*SE* = .06) (Rater 4, hardest rater). This is a significant difference in severity across raters, $\chi^2(13) = 1182.6$, $p < .001$. Rater severity statistics are presented in Table 31.

The revised writing post data had person reliability of .83. All traits had acceptable infit and outfit mean-squares between 0.5 and 1.5 with the exception of one, Opposition, which had an outfit mean-square of 1.62. Trait fit statistics are presented in Table 28. All raters had acceptable infit and outfit mean-squares between 0.5 and 1.5. Rater fit statistics are presented in Table 29. Rater severities on the pre ranged from -0.87 logits (*SE* = .03) (Rater 6, easiest rater) to 1.51 logits (*SE* = .05) (Rater 4, hardest rater). This is a significant difference in severity across raters, $\chi^2(13) = 2265.2$, $p < .001$. Rater severity statistics are presented in Table 31.

Two traits and eight raters had significantly different difficulty measures in the pre and post scales. Therefore, a stacking and anchoring procedure was performed similar to the one recommended by Wolfe and Chiu (1999). First, a stacked dataset was used to create stable category structure measures. These measures were used as an anchor for the pre data, from which pre person measures, rater severity, and trait difficulties were estimated. The rater severity and trait difficulties were used to anchor the invariant items in the post data, from which post person measures were estimated. Trait difficulty statistics are presented in Table 32.

Rasch person estimates for the revised writing pre data had a range from -8.04 logits to 2.48 logits, with an overall mean score of -.48 (*SD* = 1.21). The data had a significant negative skew (skewness = ‑1.72, *SE* = .06) and were significantly leptokurtic (kurtosis = 7.29, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1873) = .08$, $p = .000$. Rasch person estimates for the

revised writing post data had a range from -7.86 logits to 2.71 logits, with an overall mean score of -.72 (*SD* = 1.25). The data had a significant negative skew (skewness = ‑1.50, *SE* = .06) and were significantly leptokurtic (kurtosis = 5.16, *SE* = .11). The Kolmogorov-Smirnov test confirmed the data did not follow a normal distribution, $D(1888) = .07$, $p = .000$.

Table 29

Writing Trait Item Fit Statistics—Revised Model

	Revised pre scale			Revised post scale		
Trait	Infit Mean Square	Outfit Mean Square	Point-measure Correlation	Infit Mean Square	Outfit Mean Square	Point-measure Correlation
Claim	1.19	1.30	.22	1.19	1.28	.19
Evidence	0.92	0.90	.41	0.97	0.97	.39
Reasoning	1.07	1.00	.33	1.04	1.03	.30
Addressing Opposition	1.16	1.76[a]	.10	1.16	1.62[a]	.07
Organization	1.10	1.02	.36	1.14	1.15	.34
Science Content	1.05	1.19	.39	0.98	0.95	.41
Social Studies Content	1.23	1.45	.32	1.27	1.33	.34

[a] Trait does not meet specified fit criteria

Table 30

Rater Fit Statistics—Revised Model

	Revised pre scale			Revised post scale		
Rater	Infit mean square	Outfit mean square	Point-measure correlation	Infit mean square	Outfit mean square	Point-measure correlation
1	1.28	1.38	.39	1.09	1.21	.40
2	0.93	0.91	.40	0.93	0.91	.40
3	1.16	1.19	.43	1.17	1.21	.43
4	1.32	1.53[a]	.37	1.25	1.50	.39
5	1.03	1.35	.41	0.95	1.10	.42
6	1.26	1.43	.30	1.27	1.38	.30
7	1.11	1.31	.43	1.15	1.31	.41
8	1.11	1.36	.42	1.11	1.12	.41
9	0.82	0.72	.46	1.06	1.27	.41
10	1.50	1.82[a]	.47	1.06	1.10	.44
11	1.04	1.10	.45	1.00	0.96	.46
12	1.00	.98	.44	1.18	1.18	.41
13	1.02	1.18	.37	1.02	1.04	.34
14	1.09	1.33	.40	1.11	1.28	.39

[a] Rater fit does not meet specified fit criteria

Table 31

Rater Severity Statistics—Revised Model

	Revised pre scale		Revised post scale		Anchored rater
Rater	Difficulty measure (SE)	95% CI	Difficulty measure (SE)	95% CI	difficulty measure
1	0.00 (.06)	-0.12, 0.12	-0.05 (.06)	-0.17, 0.07	0.10
2*	-0.18 (.03)	-0.24, -0.12	-0.55 (.03)	-0.61, -0.49	-0.27
3*	-0.33 (.04)	-0.41, -0.25	0.09 (.04)	0.01, 0.17	-0.13
4	1.34 (.06)	1.22, 1.46	1.51 (.05)	1.41, 1.61	1.48
5*	0.11 (.06)	-0.01, 0.23	-0.14 (.05)	-0.24, -0.04	0.03
6*	-0.73 (.03)	-0.79, -0.67	-0.87 (.03)	-0.93, -0.81	-0.75
7	-0.62 (.05)	-0.72, -0.52	-0.61 (.04)	-0.69, -0.53	-0.64
8	-0.06 (.04)	-0.14, 0.02	-0.07 (.03)	-0.13, -0.01	-0.12
9*	0.12 (.06)	0.00, 0.24	0.37 (.06)	0.25, 0.49	0.45
10*	0.49 (.12)	0.25, 0.73	-0.18 (.12)	-0.42, 0.06	-0.34
11	-0.17 (.05)	-0.27, -0.07	-0.30 (.05)	-0.40, -0.20	-0.12
12*	-0.01 (.05)	-0.11, 0.09	0.21 (.05)	0.11, 0.31	0.13
13*	0.18 (.06)	0.06, 0.30	0.57 (.05)	0.47, 0.67	0.28
14	-0.15 (.04)	-0.23, -0.07	0.01 (.04)	-0.07, 0.09	-0.11

*Rater severity changed significantly pre to post.

Table 32

Writing Trait Difficulty Statistics—Revised Model

	Revised pre scale		Revised post scale		Anchored
Trait	Difficulty measure (SE)	95% CI	Difficulty measure (SE)	95% CI	difficulty measure
Claim	-2.34 (.04)	-2.42, -2.26	-2.37 (.04)	-2.45, -2.29	-2.20
Evidence	0.43 (.03)	0.37, 0.49	0.43 (.03)	0.37, 0.49	0.47
Reasoning	0.92 (.03)	0.86, 0.98	0.97 (.03)	0.91, 1.03	0.73
Addressing Opposition	1.73 (.05)	1.63, 1.83	1.84 (.05)	1.74, 1.94	1.56
Organization	-2.09 (.04)	-2.17, -2.01	-2.15 (.04)	-2.23, -2.07	-1.96
Science Content*	0.24 (.03)	0.18, 0.30	0.46 (.03)	0.40, 0.52	0.33
Social Studies Content*	1.11 (.03)	1.05, 1.17	0.82 (.02)	0.78, 0.86	1.06

*Trait difficulty changed significantly pre to post.

Comparisons of Measures and Conclusions

Science knowledge. The science knowledge observed pre scores and Rasch pre measures had a very strong linear relationship and a perfect monotonic relationship ($r[1924] = .99$, $p = .000$; $r_s[1924] = 1.00$, $p = .000$), as did observed post scores and Rasch post measures ($r[1879] = .99$, $p = .000$; $r_s[1879] = 1.00$, $p = .000$).

Levene's test confirmed equality of error variances for the treatment and control groups for observed scores, $F(1) = 2.14$, $p = .14$ and Rasch measures, $F(1) = 0.10$, $p = .76$. There was no significant interaction between the covariate and treatment condition for observed scores, $F(1) = 0.00$, $p = .99$ or Rasch measures, $F(1) = 0.23$, $p = .63$, confirming homogeneity of regression slopes. ANCOVA using observed scores

indicated a significant impact of treatment on post scores after controlling for pre scores, $F(1, 1833) = 37.05$, $p = .000$, $\eta_p^2 = .02$. ANCOVA using Rasch measures indicated the same significant effect, $F(1, 1833) = 37.98$, $p = .000$, $\eta_p^2 = .02$. The WWC procedure for cluster randomization indicated ICCs of 0.35 (observed scores) and 0.32 (Rasch measures) and Hedges' *g* effect sizes of 0.22 (observed scores) and 0.22 (Rasch scores). The corrected *t* statistics found the effect to be significant using observed scores ($p = .01$) and Rasch measures ($p = .01$).

As a follow-up comparison, classroom-level observed scores and Rasch measures were created using the mean of the individual scores and measures in each classroom. A two-way repeated measures ANOVA (time x treatment) using the observed scores revealed a significant time x treatment effect such that classrooms that participated in GE2 had greater gains over the time period than control classrooms, $F(1) = 8.25$, $p = .01$, $\eta_p^2 = .14$. The same analysis using the Rasch measures revealed the same effect, $F(1) = 9.15$, $p = .004$, $\eta_p^2 = .15$.

Science interest. The science interest observed pre scores and Rasch pre measures had a very strong linear relationship and a perfect monotonic relationship ($r[1792] = .99$, $p = .000$; $r_s[1792] = 1.00$, $p = .000$), as did observed post scores and Rasch post measures ($r[1816] = .99$, $p = .000$; $r_s[1816] = 1.00$, $p = .000$).

Levene's test confirmed equality of error variances for the treatment and control groups for observed scores, $F(1) = 0.53$, $p = .47$ and Rasch measures, $F(1) = 0.16$, $p = .69$. There was no significant interaction between the covariate and treatment condition for observed scores, $F(1) = 0.33$, $p = .56$ or Rasch measures, $F(1) = 0.95$, $p = .33$, confirming homogeneity of regression slopes. ANCOVA using observed scores

indicated no impact of treatment on post scores after controlling for pre scores, $F(1, 1662) = 1.57$, $p = .21$. ANCOVA using Rasch measures indicated the same non-effect, $F(1, 1763) = 1.58$, $p = .21$. The WWC procedure for cluster randomization indicated ICCs of 0.32 (observed scores) and 0.31 (Rasch measures) and Hedges' g effect sizes of 0.04 (observed scores) and 0.05 (Rasch scores). The corrected t statistics found the effect to be non-significant using observed scores ($p = .52$) and Rasch measures ($p = .56$).

As a follow-up comparison, classroom-level observed scores and Rasch measures were created using the mean of the individual scores and measures in each classroom. A two-way repeated measures ANOVA (time x treatment) using the observed scores revealed no significant time x treatment effect, $F(1) = 0.31$, $p = .58$. The same analysis using the Rasch measures revealed the same non-effect, $F(1) = 0.25$, $p = .62$.

Science career interest. The science career observed pre scores and Rasch pre measures had a very strong linear relationship and a perfect monotonic relationship ($r[1818] = .98$, $p = .000$; $r_s[1818] = 1.00$, $p = .000$), as did observed post scores and Rasch post measures ($r[1823] = .99$, $p = .000$; $r_s[1823] = 1.00$, $p = .000$).

Levene's test suggested inequality of error variances for the treatment and control groups for observed scores, $F(1) = 9.56$, $p = .002$ and Rasch measures, $F(1) = 6.73$, $p = .01$. There was also a significant interaction between the covariate and treatment condition for observed scores, $F(1) = 6.25$, $p = .01$ and Rasch measures, $F(1) = 6.20$, $p = .01$, indicating heterogeneity of regression slopes. Based on these violations of assumptions, findings should be interpreted with caution. ANCOVA using observed scores indicated no significant impact of treatment on post scores after controlling for pre scores, $F(1, 1682) = 0.02$, $p = .88$. ANCOVA using Rasch measures indicated the same

non-significant effect, $F(1, 1763) = 0.03$, $p = .86$. The WWC procedure for cluster randomization indicated ICCs of 0.48 (observed scores) and 0.42 (Rasch measures) and Hedges' *g* effect sizes of -0.01 (observed scores) and 0.01 (Rasch scores). The corrected *t* statistics found the effect to be non-significant using observed scores ($p = .96$) and Rasch measures ($p = .95$).

As a follow-up comparison, classroom-level observed scores and Rasch measures were created using the mean of the individual scores and measures in each classroom. A two-way repeated measures ANOVA (time x treatment) using the observed scores revealed no significant time x treatment effect, $F(1) = 1.37$, $p = .25$. The same analysis using the Rasch measures revealed the same non-effect, $F(1) = 1.10$, $p = .30$.

Social studies interest. The social studies interest observed pre scores and Rasch pre measures had a very strong linear relationship and a perfect monotonic relationship ($r[1842] = .97$, $p = .000$; $r_s[1842] = 1.00$, $p = .00$), as did observed post scores and Rasch post measures ($r[1829] = .97$, $p = .000$; $r_s[1829] = 1.00$, $p = .000$).

Levene's test suggested inequality of error variances for the treatment and control groups for observed scores, $F(1) = 4.67$, $p = .03$ but not for Rasch measures, $F(1) = 3.40$, $p = .07$. There was no significant interaction between the covariate and treatment condition for observed scores, $F(1) = 0.11$, $p = .74$ or Rasch measures, $F(1) = 0.95$, $p = 0.33$, indicating homogeneity of regression slopes. Based on the violation of the assumption of equal variances, findings should be interpreted with caution. ANCOVA using observed scores indicated no impact of treatment on post scores after controlling for pre scores, $F(1, 1714) = 0.19$, $p = .67$. ANCOVA using Rasch measures indicated the same non-effect, $F(1, 1761) = 0.10$, $p = .75$. The WWC procedure for cluster

randomization indicated ICCs of 0.22 (observed scores) and 0.23 (Rasch measures) and Hedges' *g* effect sizes of 0.02 (observed scores) and 0.01 (Rasch scores). The corrected *t* statistics found the effect to be non-significant using observed scores ($p = .81$) and Rasch measures ($p = .89$).

As a follow-up comparison, classroom-level observed scores and Rasch measures were created using the mean of the individual scores and measures in each classroom. A two-way repeated measures ANOVA (time x treatment) using the observed scores revealed no significant time x treatment effect, $F(1) = 0.85$, $p = .36$. The same analysis using the Rasch measures revealed the same non-effect, $F(1) = 1.06$, $p = .31$.

Writing self-efficacy. The writing self-efficacy observed pre scores and Rasch pre measures had a very strong linear relationship and a perfect monotonic relationship ($r[1809] = .98$, $p = .000$; $r_s[1809] = 1.00$, $p = .000$), as did observed post scores and Rasch post measures ($r[1797] = .98$, $p = .000$; $r_s[1797] = 1.00$, $p = .000$).

Levene's test confirmed equality of error variances for the treatment and control groups for observed scores, $F(1) = 0.78$, $p = .38$ and Rasch measures, $F(1) = 0.00$, $p = .97$. There was no significant interaction between the covariate and treatment condition for observed scores, $F(1) = 2.23$, $p = .14$ or Rasch measures, $F(1) = 2.40$, $p = .12$, confirming homogeneity of regression slopes. ANCOVA using observed scores indicated a significant impact of treatment on post scores after controlling for pre scores, $F(1, 1656) = 4.15$, $p = .04$, $\eta_p^2 = .002$. ANCOVA using Rasch measures indicated the same significant effect, $F(1, 1752) = 5.52$, $p = .02$, $\eta_p^2 = .003$. The WWC procedure for cluster randomization indicated ICCs of 0.28 (observed scores) and 0.29 (Rasch measures) and Hedges' *g* effect sizes of 0.08 (observed scores) and 0.09 (Rasch scores).

The corrected *t* statistics found the effect to be non-significant using observed scores ($p = .32$) and Rasch measures ($p = .26$).

As a follow-up comparison, classroom-level observed scores and Rasch measures were created using the mean of the individual scores and measures in each classroom. A two-way repeated measures ANOVA (time x treatment) using the observed scores revealed no significant time x treatment effect, $F(1) = 2.09$, $p = .15$. The same analysis using the Rasch measures revealed the same non-effect, $F(1) = 3.12$, $p = .08$.

Socio-science self-efficacy. The socio-science self-efficacy observed pre scores and Rasch pre measures had a very strong linear relationship and a perfect monotonic relationship ($r[1834] = .97$, $p = .000$; $r_s[1834] = 1.00$, $p = .000$), as did observed post scores and Rasch post measures ($r[1812] = .96$, $p = .000$; $r_s[1812] = 1.00$, $p = .000$).

Levene's test confirmed equality of error variances for the treatment and control groups for observed scores, $F(1) = 0.43$, $p = .51$ and Rasch measures, $F(1) = 0.22$, $p = .64$. There was no significant interaction between the covariate and treatment condition for observed scores, $F(1) = 0.22$, $p = .64$ or Rasch measures, $F(1) = 0.09$, $p = .77$, confirming homogeneity of regression slopes. ANCOVA using observed scores indicated no significant impact of treatment on post scores after controlling for pre scores, $F(1, 1689) = 1.76$, $p = .19$. ANCOVA using Rasch measures indicated the same non-significant effect, $F(1, 1749) = 2.07$, $p = .15$. The WWC procedure for cluster randomization indicated ICCs of 0.20 (observed scores) and 0.20 (Rasch measures) and Hedges' *g* effect sizes of 0.05 (observed scores) and 0.06 (Rasch scores). The corrected *t* statistics found the effect to be non-significant using observed scores ($p = .43$) and Rasch measures ($p = .40$).

As a follow-up comparison, classroom-level observed scores and Rasch measures were created using the mean of the individual scores and measures in each classroom. A two-way repeated measures ANOVA (time x treatment) using the observed scores revealed no significant time x treatment effect, $F(1) = 2.18$, $p = .15$. The same analysis using the Rasch measures revealed the same non-effect, $F(1) = 1.30$, $p = .26$.

Writing (original model). The writing observed pre scores and Rasch pre measures had very strong linear and monotonic relationships ($r[1888] = .87$, $p = .000$; $r_s[1888] = .91$, $p = .000$), as did observed post scores and Rasch post measures ($r[1873] = .86$, $p = .000$; $r_s[1873] = .90$, $p = .000$).

Levene's test suggested inequality of error variances of error variances for the treatment and control groups for observed scores, $F(1) = 5.09$, $p = .02$ and Rasch measures, $F(1) = 9.25$, $p = .00$. There was a significant interaction between the covariate and treatment condition for observed scores, $F(1) = 12.86$, $p = .00$ and Rasch measures, $F(1) = 18.78$, $p = .00$, suggesting heterogeneity of regression slopes. Based on these violations of assumptions, findings should be interpreted with caution. ANCOVA using observed scores indicated a significant impact of treatment on post scores after controlling for pre scores, $F(1, 1805) = 129.09$, $p = .000$, $\eta_p^2 = .07$. ANCOVA using Rasch measures indicated the same significant effect, $F(1, 1805) = 178.60$, $p = .000$, $\eta_p^2 = .09$. The WWC procedure for cluster randomization indicated ICCs of 0.07 (observed scores) and 0.11 (Rasch measures) and Hedges' g effect sizes of 0.49 (observed scores) and 0.56 (Rasch scores). The corrected t statistics found the effect to be significant using observed scores ($p = .000$) and Rasch measures ($p = .000$).

As a follow-up comparison, classroom-level observed scores and Rasch measures were created using the mean of the individual scores and measures in each classroom. A two-way repeated measures ANOVA (time x treatment) using the observed scores revealed a significant time x treatment effect such that classrooms that participated in GE2 had greater gains over the time period than control classrooms, $F(1) = 45.30$, $p = .000$, $\eta_p^2 = .47$. The same analysis using the Rasch measures revealed the same effect, $F(1) = 54.68$, $p = .000$, $\eta_p^2 = .52$.

Writing (revised Rasch model). After the writing model was revised to address some elements of misfit, the observed pre scores and Rasch pre measures had very strong linear and monotonic relationships ($r[1888] = .87$, $p = .000$; $r_s[1888] = .90$, $p = .000$), as did observed post scores and Rasch post measures ($r[1873] = .85$, $p = .000$; $r_s[1873] = .88$, $p = .000$).

Levene's test suggested inequality of error variances of error variances for the treatment and control groups for the revised model Rasch measures, $F(1) = 6.35$, $p = .01$. There was a significant interaction between the covariate and treatment condition for revised Rasch measures, $F(1) = 16.65$, $p = .00$, suggesting heterogeneity of regression slopes. Based on these violations of assumptions, findings should be interpreted with caution. ANCOVA using the revised Rasch measures indicated a significant effect, $F(1, 1805) = 181.92$, $p = .000$, $\eta_p^2 = .09$. The WWC procedure for cluster randomization indicated ICC of 0.11 and Hedges' *g* effect of 0.57. The corrected *t* statistics found the effect to be significant ($p = .000$).

As a follow-up comparison, classroom-level Rasch measures were created using the mean of the individual measures in each classroom. A two-way repeated measures

ANOVA (time x treatment) using the observed scores revealed a significant time x treatment effect such that classrooms that participated in GE2 had greater gains over the time period than control classrooms, $F(1) = 62.37$, $p = .000$, $\eta_p^2 = .55$.

All conclusion findings are summarized in Table 33.

Table 33
Summary of findings

Instrument	Pre		Post		Student level ANCOVA	Student-level effect size	Cluster-corrected *p*	Classroom-level ANOVA	Classroom-level effect size
	Treatment *M* (*SD*)	Control *M* (*SD*)	Treatment *M* (*SD*)	Control *M* (*SD*)					
Knowledge									
Rasch measures	.29 (.99) *n* = 972	.18 (1.04) *n* = 954	.62 (1.10) *n* = 944	.32 (1.10) *n* = 937	$F(1, 1833) = 37.98$	$\eta_p^2 = .02$	.01	$F(1) = 9.15, p = .004$	$\eta_p^2 = .15$
Observed scores	10.05 (3.11) *n* = 972	9.68 (3.22) *n* = 954	11.10 (3.26) *n* = 944	10.11 (3.36) *n* = 937	$F(1, 1833) = 37.05, p = .000$	$\eta_p^2 = .02$	.01	$F(1) = 8.25, p = .01$	$\eta_p^2 = .14$
Science interest									
Rasch measures	1.05 (2.27) *n* = 948	.90 (2.19) *n* = 925	.81 (2.43) *n* = 934	.56 (2.41) *n* = 921	$F(1, 1763) = 1.58, p = .21$	n/a	.56	$F(1) = 0.25, p = .62$	n/a
Observed scores	17.20 (4.42) *n* = 913	16.91 (4.36) *n* = 881	16.78 (4.81) *n* = 915	16.24 (4.84) *n* = 903	$F(1, 1662) = 1.57, p = .21$	n/a	.52	$F(1) = 0.31, p = .58$	n/a
Science career interest*									
Rasch measures	-1.93 (3.14) *n* = 948	-2.32 (3.00) *n* = 925	-1.82 (3.22) *n* = 934	-2.1 (3.14) *n* = 921	$F(1, 1763) = 0.03, p = .86$	n/a	.95	$F(1) = 1.10, p = .30$	n/a
Observed scores	11.36 (5.65) *n* = 924	10.66 (5.25) *n* = 894	11.52 (5.76) *n* = 920	11.03 (5.53) *n* = 903	$F(1, 1682) = 0.02, p = .88$	n/a	.96	$F(1) = 1.37, p = .25$	n/a
Social studies interest*									
Rasch measures	1.46 (1.88) *n* = 947	1.44 (1.82) *n* = 925	1.31 (1.96) *n* = 934	1.24 (2.05) *n* = 920	$F(1, 1761) = 0.10, p = .75$	n/a	.89	$F(1) = 1.06, p = .31$	n/a
Observed scores	15.11 (3.35) *n* = 930	15.03 (3.32) *n* = 912	14.72 (3.57) *n* = 920	14.56 (3.70) *n* = 909	$F(1, 1714) = 0.19, p = .67$	n/a	.81	$F(1) = 0.85, p = .36$	n/a
Writing Self-efficacy									
Rasch measures	1.06 (2.03) *n* = 949	1.03 (1.86) *n* = 924	1.24 (2.24) *n* = 926	.98 (2.12) *n* = 914	$F(1, 1752) = 5.52, p = .02$	$\eta_p^2 = .003$	.26	$F(1) = 3.12, p = .08$	n/a
Observed scores	17.61 (4.22) *n* = 924	17.58 (3.96) *n* = 885	17.91 (4.51) *n* = 906	17.43 (4.39) *n* = 891	$F(1, 1656) = 4.15, p = .04$	$\eta_p^2 = .002$	.32	$F(1) = 2.09, p = .15$	n/a
Socio-scientific Self-efficacy									
Rasch measures	2.59 (2.08) *n* = 949	2.53 (2.07) *n* = 924	2.22 (2.01) *n* = 926	2.05 (2.04) *n* = 912	$F(1, 1749) = 2.07, p = .15$	n/a	.40	$F(1) = 1.30, p = .26$	n/a
Observed scores	20.79 (3.56) *n* = 932	20.74 (3.52) *n* = 902	20.34 (3.77) *n* = 914	20.03 (3.98) *n* = 898	$F(1, 1689) = 1.76, p = .19$	n/a	.43	$F(1) = 2.18, p = .15$	n/a
Writing*									
Rasch original measures	-.68 (1.23) *n* = 944	-.76 (1.27) *n* = 945	-.13 (.98) *n* = 907	-.78 (1.26) *n* = 901	$F(1, 1805) = 178.60, p = .000$	$\eta_p^2 = .09$	.00	$F(1) = 54.68, p = .000$	$\eta_p^2 = .52$
Rasch revised measures	-.71 (1.27) *n* = 944	-.80 (1.31) *n* = 945	-.02 (1.13) *n* = 907	-.78 (1.42) *n* = 901	$F(1, 1805) = 181.92, p = .000$	$\eta_p^2 = .09$	.00	$F(1) = 62.37, p = .000$	$\eta_p^2 = .55$
Observed scores	6.21 (2.25) *n* = 944	6.00 (2.28) *n* = 945	7.52 (2.26) *n* = 907	6.32 (2.31) *n* = 901	$F(1, 1805) = 129.09, p = .000$	$\eta_p^2 = .07$	.00	$F(1) = 45.30, p = .000$	$\eta_p^2 = .47$

Note. Sample sizes vary among instruments due to incomplete data from some participants. Sample sizes vary within interest and self-efficacy scales for observed score and Rasch analyses because observed score analyses excluded any participant that did not complete all items; Rasch analyses do not have this requirement and include participants who partially completed the scales.
*Data did not meet all assumptions of ANOVA; findings should be interpreted with caution.

Chapter 5: Conclusion

Public and private funders devote hundreds of millions of dollars each year to fund theoretically sound educational interventions and research on these interventions. However, the minority of these interventions (10-15%) have been established as clearly effective; most have been shown to be weak or ineffectual compared with normal educational practice (Coalition for Evidence-based Policy, 2013). This study focused on one pedagogical approach, PBL, as an example of an intervention that has yielded mixed findings, in order to illuminate some of the reasons for the unclear conclusions provided by empirical research.

The purpose of this study was to explore one possible reason for mixed findings in educational intervention studies: the use of observed scores in parametric statistics, where interval measures should be used. The study compared the statistical conclusions derived from observed scores and Rasch measures of several outcomes, using real data from a large, grant-funded, PBL intervention study. The dataset was intended as an exemplar; results are meant to be applicable to a wide variety of educational interventions; therefore, discussion of the impact of GE2 on the outcome variables is not relevant to this study.

The purpose of the first research question was to establish that corresponding sets of observed scores and Rasch measures were closely correlated for all of the outcome assessments in the study (objective test, attitude measures, and rater-scored essay). The first step in answering this research question was to evaluate how well data fit the Rasch model. For the science knowledge scale, interest scales, and self-efficacy scales, deviations from model fit criteria were minor. The science knowledge scale had a lower

person reliability than is desirable (.68 pre and .70 post), although all items had acceptable fit statistics; conclusions drawn from the science knowledge scale should be interpreted with caution. The interest and self-efficacy scales all had person reliabilities above .80, except for social studies interest (.75 pre and .78 post) and socio-science self-efficacy (.77 pre and .76 post). Reliabilities between .65 and .80 are still "minimally acceptable" (.65 to .70) or "respectable" (.70 to .80) and usable for research purposes at the group level (DeVellis, 2003, p. 95). Certain individual categories and items displayed slightly high outfit mean-squares; closer examination suggests these items might be too easy or too difficult to endorse, or as in the case of two items in the science career interest scale, might be confusing to certain students.

The rater-scored writing assessment did not meet all model fit criteria. One trait (Opposition) and four raters had unacceptably high outfit mean-squares. In order to improve model fit, the Opposition category was collapsed from three rating scale points to two rating scale points, and Opposition ratings from two raters were removed. This improved outfit mean-squares; however, Opposition and two raters still had high levels of misfit. Additionally, collapsing the Opposition category created an imbalance in the scoring rubric where the multiple traits, which are ostensibly meant to contribute equally to a final score, have different scoring structures and do not all contribute equally. However, this was also an issue in the original rubric and in the observed score model (e.g. Claim is worth two points but Evidence is worth three points). The appropriate remedy for these outstanding issues (the scoring structure, Opposition misfit, and misfit for two raters) is to revise the rubric, retrain raters, and collect additional data.

Results showed that for the objective test and attitude measures, the linear relationships were very strong ($r > .97$), and the monotonic relationships were perfect ($r_s = 1.00$). This result is consistent with prior research that shows high correlations between observed scores and Rasch measures. Further, it is implied by a property of Rasch measures, which is that observed scores are sufficient statistics for Rasch measures (Wright & Linacre, 1989). When observed scores and Rasch measures are based on identical data, they will always have a perfect monotonic relationship and a high linear relationship; more extreme high or low scores will distort the linear relationship slightly.

The correlation between Rasch measures and observed scores for the rater-scored essay was slightly lower ($r > .86$, $r_s > .88$), because the datasets for the observed scores and Rasch measures were *not* identical. Observed scores were the average scores awarded by two or three raters; Rasch scores were estimated based on the scores awarded by only two raters. After the writing model was revised to improve model fit, the correlations remained almost the same. The less-than-perfect linear relationship suggests that for this dataset and scoring method, there are some differences, potentially substantive, between the observed scores and Rasch measures. The second research question further explores potentially substantive differences.

Having established that Rasch measures and observed scores are closely correlated, especially when identical data is used, the second research question probed for substantive differences between observed scores and Rasch measures, in particular, different conclusions about the impact of the intervention on science knowledge, attitudes, or writing quality. If observed scores and Rasch measures led to different

statistical conclusions about the impact of the intervention, this would provide evidence that the use of observed scores in research may be one source of mixed findings.

The impact of the intervention on each outcome was first analyzed at the student level, controlling for pre-scores. Observed scores and Rasch measures agreed on all statistical conclusions: GE2 had a significant impact on science knowledge, no impact on science interest, no impact on science career interest, no impact on social studies interest, a significant impact on writing self-efficacy, and no impact on socio-science self-efficacy. Observed scores for writing quality and the two different sets of Rasch measures (original model and revised model) all agreed that GE2 had a significant impact on writing quality.

The intervention impact was next corrected for classroom-level cluster randomization using the WWC procedure. After correcting for ICC, results were the same as they were in the original analysis, with the exception of writing self-efficacy. The original student-level analysis found a significant impact of GE2 on writing self-efficacy; after correcting for ICC, no significant effect was found. Observed scores and Rasch measures agreed on this finding.

Finally, classroom-level observed scores and Rasch measures were created and used in a repeated-measures analysis. The repeated-measures analyses agreed with the ICC-corrected student-level analyses for all outcomes, including finding no significant impact of GE2 on writing self-efficacy. Observed scores and Rasch measures agreed on all findings.

In summary, for all seven outcomes tested and using three analyses, Rasch measures and observed scores consistently agreed on the statistical significance of

findings. This is in contrast with certain earlier findings. Lynn, Yukhymenko, and Lawless (under review), which used some of the same scales as the present dissertation (with a different sample of students), found that Rasch measures and observed scores produced divergent conclusions regarding the writing self-efficacy scale; according to Rasch measures, there was a significant effect ($p < .05$), which observed scores did not detect. In this prior study and in the present one, the writing self-efficacy scale had high reliability and good item and category fit to the Rasch model. The general trend in both studies was the same, with Rasch measures finding a slightly lower *p*-value than observed scores did. In general, the writing self-efficacy scale can be seen as a "borderline" case, where different measurement and analysis methods (Rasch vs. observed scores in Lynn, Yukhymenko, and Lawless [under review] and student-level vs. classroom-level in the present study) led to *p*-values on opposite sides of .05 and thus different conclusions.

The finding that observed scores and Rasch measures agreed on a statistical conclusion for writing self-efficacy ($p = .04$ and $p = .02$, respectively) is far from sufficient evidence that there is no difference between observed scores and Rasch measures. Had the critical *p* value been set differently, for example, at .025 to correct for familywise error if writing self-efficacy and socioscientific self-efficacy were considered together, then observed scores and Rasch measures would provide different statistical conclusions. Similarly, if the effect of the intervention were slightly smaller or within-groups variance slightly larger, or if the sample size were slightly smaller, these values could have fallen on opposite sides of critical *p* value. The issue of sample size is one possible explanation for the disagreement in findings between Lynn, Yukhymenko, and Lawless (under review). The earlier study had a sample size of approximately 600

participants, while the present sample size is approximately 1,700 participants. Both studies agreed that the effect size was small ($\eta_p^2 < .01$); therefore, a large sample would be required to detect the effect. However, it is impossible to determine empirically whether the small effect size is due to the intervention having a weak effect on writing self-efficacy or whether the writing self-efficacy scale is insufficiently sensitive to detect a strong, true effect.

Limitations and Future Directions

The difficulty explaining the writing self-efficacy findings highlights one of the major limitations of this study: because it uses real data, it is impossible to know the "true" underlying effect and determine empirically which analysis is closer to providing the correct conclusion. By creating datasets with known parameters, simulated data would provide more precision in exploring the various circumstances under which observed scores and Rasch measures produce different conclusions. Borderline cases, where the *p*-value is near .05, may be a fertile ground for exploration, as these are the cases where a slight difference ($p = .04$ versus $p = .06$) could be considered important. A combination of real data, which is powerful for demonstration purposes, and simulated data that controls dataset parameters, might be an ideal combination for conducting a robust study of Rasch measures and observed scores *and* for communicating findings to a general research audience.

Regardless of what type of data was used, this dissertation is firmly rooted in null hypothesis significance testing, which has its drawbacks. The *p*-value of .05 as a cutoff for statistical significance makes results easy to communicate to a general research audience; however, it is far from the final word on substantive evidence of practical

importance. Introductory statistics textbooks explain that .05 as a critical value is arbitrary and recommend using effect size as an indicator of the practical significance along with a *p*-value as an indicator of statistical significance (e.g. Field, 2005; APA, 2009). More sophisticated criticism of null hypothesis significance testing as a practice that is misunderstood and misused dates back at least to the 1960's (Cohen, 1994; Krantz, 1999). Gelman and Stern (2006) specifically recommend against the type of comparison conducted in the present dissertation ("*X* is statistically significant by *Y* is not"), explaining that the difference between a dataset with a significant result and one with a non-significant result may not itself be significant.

Even if the difference between the two sets of results is considered important, favoring one measurement model over another may not be an ideal solution. Thissen (2016) questions the importance of the distinction between ordinal and interval measures on the same basis as this dissertation, but with an opposite conclusion: Thissen argues that if "an important system of educational evaluation [is] so sensitive it [gives] different answers as a function of an issue so minor as the relative spacing of high and low scores," the system itself should be reconsidered entirely (p. 84). In other words, if a measure is not strong enough to capture a clear effect or non-effect, the measure itself may be flawed.

Future research should continue to explore various ways to evaluate the difference between observed score and Rasch measure datasets using both qualitative and quantitative methods. Quantitative alternatives to comparing statistical significance could include comparing effect sizes or model fit (see Cheema [2013], which used R^2). In the case of rater-scored essays, expert qualitative analysis could help establish whether Rasch

measures or average observed scores better reflect the construct(s) intended by the rubric and scoring system.

Finally, this study did not use Rasch modeling in a theoretically ideal way. Crucial to the Rasch paradigm is the idea that the Rasch model is inherently sound, and the model should be used to create and validate scales in order to collect data that fits the model, which then is able to estimate true interval measures. Forcing data into the Rasch model to create interval measures does not create high-quality measures or "fix" data collected using low-quality measures (e.g., Salzberger 2010). This study found that, with some exceptions, collected data fit the Rasch model. It should be noted that model fit may be overestimated in this study due to a large sample size. Where needed, as in the rater-scored essay data, some modifications were made to improve model fit. Ideally, this step would have been part of an iterative process in which instruments were revised to better fit the Rasch model and the population of students in the intervention. The purpose of this demonstration is to establish the importance of using Rasch person measure estimates after scales have been revised and validated, not to diminish the importance of the early steps of developing, revising, and validating scales.

Besides scale validation, Rasch modeling offers several other advantages that were not explored in this study, including identification of rater bias and differential item functioning (DIF). Future studies could focus on the specific consequences of ignoring poor model fit, rater bias, and/or DIF, in terms of their impacts on statistical conclusions. Numerous extant studies demonstrate how Rasch can be used to identify these issues in collected data (and sometimes remedy them); additionally, it could be useful for researchers to see how ignoring these issues can lead to incorrect statistical decisions.

This study explored the use of observed scores and Rasch measures for the same analyses of outcomes associated with an educational intervention, with the goal of determining whether the use of observed scores, which are ordinal data, in parametric statistics, which require interval data, may be a source of mixed findings in PBL intervention literature. Based on the dataset and analyses used in this dissertation, there is no evidence to conclude that the use of observed/ordinal scores is a likely culprit for null or mixed findings. Findings suggest that, when sample size is adequately large, and when measures are already determined (i.e. Rasch analysis is not being used for scale development) and meet the general standards of reliability (in Classical Test Theory and Rasch analyses) and model fit (in Rasch analysis), there is little to be gained by using Rasch modeling to convert ordinal observed scores into interval Rasch person measures for this data. In this situation, "what goes in is what comes out" of Rasch analysis, or, as Salzberger said, the Rasch model cannot "travel faster than the speed of light" to make measures more accurate or sensitive than they are (2010). There is no basis to recommend that researchers use Rasch modeling in this manner to improve the quality of their interpretation of efficacy or conclusion of the impact of an intervention, especially given that producing Rasch measures has a higher cost (in terms of specialized knowledge and software tools) compared to the very simple calculations involved with observed scores.

A contradictory prior finding, in which Rasch person measures and observed scores led to opposite conclusions about efficacy, may have been the result of a particular set of circumstances in which the sample size was almost inadequate to detect a small but statistically significant effect (Lynn, Yukhymenko, & Lawless, under review). Taken

together, the present dissertation and this prior finding may suggest that in limited cases, where statistical power (as a function of sample size and effect size) is low, the transformation from ordinal observed scores to interval Rasch person measures may offer a small advantage in evaluating the efficacy or impact of an intervention.

This conclusion does not rule out measurement issues as a possible source of mixed findings, nor does it suggest that there is no benefit to using Rasch measurement. The Rasch model provides numerous advantages to researchers developing their own measures and using measures to evaluate the impact of educational interventions: it provides detailed information regarding how students respond to items (for instance, that the way some items broke across lines led students to misread them) and in the case of MFRM, how raters use a rubric (for instance, several raters struggled with the "opposition" trait) that would not be available through Classical Test Theory analysis. The Rasch model provides numerous means of assessing the fit between the model and the collected data, and, when the data fits the model, provides estimates of true interval measures of a construct. It is an important tool for creating high-quality measurement tools, which, along with well-designed, well-implemented studies, can illuminate the benefits of PBL and other types of educational interventions.

Cited Literature

Airasian, P. W., & Russell, M. K. (2007). *Classroom assessment: Concepts and applications.* (6th ed.). New York: McGraw-Hill.

Akınoğlu, O., & Tandoğan, R. Ö. (2007). The effects of problem-based active learning in science education on students' academic achievement, attitude and concept learning. *Eurasia Journal of Mathematics, Science & Technology Education*, *3*(1), 71-81.

Albanese, M. A., & Mitchell, S. (1993). Problem-based learning: a review of literature on its outcomes and implementation issues. *Academic medicine*, *68*(1), 52-81. doi:10.1097/00001888-199301000-00012

American Psychological Association (APA) (2009). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Pyschometrika*, *43*(4), 561-573. doi:10.1007/BF02293814

Andrich, D. (2002). Implications and applications of modern test theory in the context of outcomes based education. *Studies in Educational Evaluation*, *28*(2), 103-121. doi:10.1016/S0191-491X(02)00015-9

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*(1), I7-I16. doi:10.1098/01.mlr.0000103528.48582.7c

Barrows, H. S., & Tamblyn, R. S. (1980). *Problem-based learning: An approach to medical education*. New York: Springer Publishing Company, Inc.

Belland, B. R., French, B. F., & Ertmer, P. A. (2009). Validity and problem-based learning research: A review of instruments used to assess intended learning outcomes. *Interdisciplinary Journal of Problem-based Learning*, *3*(1), 5. doi:10.7771/1541-5015.1059

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Mahway, NJ: Lawrence Erlbaum Associates, Inc.

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. doi:10.1002/sce.20106

Boone, W. J., Townsend, J. S., & Staver, J. (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education*, *95*(2), 258-280. doi:10.1002/sce.20413

Boud, D., & Feletti, G. (1998). *The challenge of problem-based learning*. Psychology Press.

Brown, S. W., Lawless, K. A., & Boyer, M. A. (2013). Promoting positive academic dispositions using a web-based PBL environment: The GlobalEd 2 project. *Interdisciplinary Journal of Problem-Based Learning*, *7*(1), 7. doi:10.7771/1541-5015.1389

Brown, S. W., & Lawless, K. A. (2014). Promoting students' writing skills in science through an educational simulation: The GlobalEd 2 project. In P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies* (pp. 371-379). Springer.

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical education*, *42*(12), 1150-1152. doi:10.1111/j.1365-2923.2008.03172.x

Chaney, B. (2015). Reconsidering findings of "no effects" in randomized control trials: Modeling differences in treatment impacts. *American Journal of Evaluation*. doi:10.1177/1098214015573788

Chang, C. Y. (2001). Comparing the impacts of a problem-based computer-assisted instruction and the direct-interactive teaching method on student science achievement. *Journal of Science Education and Technology*, *10*(2), 147-153. doi:10.1023/A:1009469014218

Chang, C. Y., & Barufaldi, J. P. (1999). The use of a problem-solving-based instructional model in initiating change in students' achievement and alternative frameworks. *International Journal of Science Education*, *21*(4), 373-388. doi:10.1080/095006999290606

Cheema, J. R. (2013). Does it matter how you measure it? The case of self-efficacy in mathematics. *Issues in Educational Research*, *23*(3), 345-356. Retrieved from http://www.iier.org.au/iier23/cheema.html

Coalition for Evidence Based Policy, (2013). Randomized controlled trials commissioned by the Institute of Education Sciences since 2002: How many found positive versus weak or no effects. Retrieved June 31, 2015 from http://coalition4evidence.org/wp-content/uploads/2013/06/IES-Commissioned-RCTs-positive-vs-weak-or-null-findings-7-2013.pdf

Cohen, J. (1994). The Earth is round (p <.05). *American Psychologist*, *49*(12), 997-1003.

Cohen, L. G., & Spenciner, L. J. (2007). *Assessment of Children & Youth with Special Needs* (3rd edition). New York: Pearson.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Orlando, FL: Holt, Rinehart, & Winston.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137-163. doi:10.1111/j.2044-8317.1963.tb00206.x

Davison, M. L., & Sharma, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin*, *107*(3), 394. doi:10.1037/0033-2909.107.3.394

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist*, *26*(3-4), 325-346. doi:10.1080/00461520.1991.9653137

DeVellis, R. F. (2003). *Scale Development: Theory and Applications* (Vol. 26). Thousand Oaks: Sage.

Dochy, F., Segers, M., Van, D. B., Piet, & Gijbels, D. (2003). Effects of problem-based learning: a meta-analysis. *Learning and Instruction*, *13*(5), 533-568. doi:http://dx.doi.org/10.1016/S0959-4752(02)00025-7

Dods, R. F., Segers, M., Van den Bossche, P., & Gijbels, D. (1997). An action research study of the effectiveness of problem-based learning in promoting the acquisition and retention of knowledge. *Journal for the Education of the Gifted*, *20*(4), 423-437.

Drake, K. N., & Long, D. (2009). Rebecca's in the dark: A comparative study of problem-based learning and direct instruction/experiential learning in two 4th-grade classrooms. *Journal of Elementary Science Education*, *21*(1), 1-16. doi:10.1007/BF03174712

Eggert, S., & Bögeholz, S. (2010). Students' use of decision-making strategies with regard to socio-scientific issues: An application of the Rasch Partial Credit Model. *Science Education*, *94*(2), 230-258. doi:10.1002/sce.20358

Embretson, S. E. (1996). Item Response Theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201-212. doi:10.1177/014662169602000302

Englehard, G. J. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191. doi:10.1207/s15324818ame0503_1

Erhart, M., Hagquist, C., Auquier, P., Rajmil, L., Power, M., & Ravens-Sieberer, U. (2010). A comparison of Rasch item-fit and Cronbach's alpha item reduction analysis for the development of a Quality of Life scale for children and adolescents. *Child: Care, Health, & Development*, *36*(4), 473-484. doi:10.1111/j.1365-2214.2009.00998.x

Fan, X. (1998). Item Response Theory and Classical Test Theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357-381. doi:10.1177/0013164498058003001

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage Publications Ltd.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines* (3rd ed.). Boston, MA: Allyn and Bacon.

Gallagher, S. A., & Stepien, W. J. (1996). Content acquisition in problem-based learning: Depth versus breadth in American studies. *Journal for the Education of the Gifted*, *19*(3), 257-275. doi:10.1177/016235329601900302

Gallagher, S. A., Stepien, W. J., & Rosenthal, H. (1992). The effects of problem-based learning on problem solving. *Gifted Child Quarterly*, *36*(4), 195-200. doi:10.1177/001698629203600405

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328-331. doi:10.1198/000313006X152649

Germann, P. J. (1988). Development of the attitude toward science in school assessment and its use to investigate the relationship between science achievement and attitude toward science in school. *Journal of research in science teaching*, *25*(8), 689-703. doi:10.1002/tea.3660250807

Gijbels, D., Dochy, F., Van, D. B., Piet, & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research*, *75*(1), 27-61. doi:10.3102/00346543075001027

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects of analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237-288.

Granberg-Rademacker, J. S. (2009). An algorithm for converting ordinal scale measurement data to interval/ratio scale. *Educational and Psychological Measurement*. doi:10.1177/0013164409344532

Granger, R. C., & Maynard, R. (2015). Unlocking the potential of the "what works" approach to policymaking and practice: improving impact evaluations. *American Journal of Evaluation*, 1-12. doi:10.1177/1098214015594420

Haiyang, S. (2010). An application of classical test theory and Many Facet Rasch Measurement in analyzing the reliability of an English rest for non-English major graduates. *Chinese Journal of Applied Linguistics*, *33*(2), 87-102.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*(1), 105-131. doi:10.3102/00346543071001105

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*(2), 107-128.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87. doi:10.3102/0162373707299706

Hernández-Ramos, P., & De La Paz, S. (2009). Learning history in middle school by designing multimedia in a project-based learning experience. *Journal of Research on Technology in Education*, *42*(2), 151-173. doi:10.1080/15391523.2009.10782545

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, *42*(2), 99-107. doi:10.1080/00461520701263368

Hmelo-Silver, C. E. (2004). Problem-Based Learning: What and How Do Students Learn? *Educational Psychology Review*, *16*(3), 235-266. doi:10.1023/B:EDPR.0000034022.16470.f3

Hmelo, C. E. (1998). Problem-based learning: Effects on the early acquisition of cognitive skill in medicine. *The journal of the learning sciences*, *7*(2), 173-208. doi:10.1207/s15327809jls0702_2

Hmelo, C. E., Holton, D. L., & Kolodner, J. L. (2000). Designing to learn about complex systems. *Journal of the Learning Sciences*, *9*(3), 247-298. doi:10.1207/S15327809JLS0903_2

Hung, W., Jonassen, D. H., & Liu, R. (2008). Chapter 38: Problem-based learning. In *Handbook of research on educational communications and technology* (3rd ed., pp. 485-506). New York: Lawrence Erlbaum Associates.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237-263. doi:10.3102/00346543060002237

IBM Corp. (2015). *IBM SPSS Statistics for Macintosh, Version 23.0.* Armonk, NY: IBM Corp.

Institute of Education Science (IES) (n.d.). About IES: Connecting Research, Policy and Practice. Retrieved 6/15/15, from http://ies.ed.gov/aboutus/

Institute of Education Science (IES) (n.d.). About the WWC. Retrieved 6/18/15, from http://ies.ed.gov/ncee/wwc/aboutus.aspx

Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical education*, *38*(12), 1217-1218. doi:10.1111/j.1365-2929.2004.02012.x

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *The Annals of Family Medicine*, *2*(3), 204-208. doi:10.1370/afm.141

King, B. M., & Minium, E. W. (2008). *Statistical reasoning in the behavioral sciences.* (5th ed.). New York: John Wiley & Sons, Inc.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75-86. doi:10.1207/s15326985ep4102_1

Kohli, N., Koran, J., & Henn, L. (2014). Relationships among Classical Test Theory and Item Response Theory frameworks via factor analytic models. *Educational and Psychological Measurement*. doi:10.1177/0013164414559071

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *94*(448), 1372-1381. doi:10.1080/01621459.1999.10473888

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115-129. doi:10.1007/BF02289694

Kuzon, J., William M, Urbanchek, M. G., & McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of plastic surgery*, *37*(3), 265-272.

Lawless, K. A., & Brown, S. W. (2012). GlobalEd2: Efficacy and replication, goal 3, educational technology. *Proposal submitted to Institute of Education Sciences (IES)*.

Lawless, K. A., & Brown, S. W. (2015). Developing scientific literacy skills through interdisciplinary, technology-based global simulations: GlobalEd 2. *Curriculum Journal*, ahead-of-print), 1-22.

Leary, H., Walker, A., Shelton, B. E., & Fitt, M. H. (2013). Exploring the relationships between tutor background, tutor training, and student learning: a problem-based learning meta-analysis. *Interdisciplinary Journal of Problem-based Learning*, *7*(1), 6. doi:10.7771/1541-5015.1331

Likert, R. (1974). A method of constructing an attitude scale. *Scaling: A sourcebook for behavioural scientists*, 233-243.

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. (1992). Why fuss about statistical sufficiency? *Rasch Measurement Transactions*, *6*(3), 230.

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

Linacre, J. M. (2015). *Facets computer program for many-facet Rasch measurement, version 3.71.4*. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2015). *Winsteps (R) Rasch measurement computer program, version 3.90.0*. Beaverton, OR: Winsteps.com.

Liu, M., Hsieh, P., Cho, Y., & Schallert, D. L. (2006). Middle school students' self-efficacy, attitudes, and achievement in a computer-enhanced problem-based learning environment. *Journal of Interactive Learning Research*, *17*(3), 225-242.

Lynn, L., & Lawless, K. A. (2015). Observed score versus Rasch score analysis of efficacy data: A case study comparison. In A. M. Columbus (Ed.), *Advances in Psychology Research Volume 109*. Hauppage NY: Nova Science Publishers.

Lynn, L. J., Yukhymenko, M. A., & Lawless, K. A. (under review). A Comparison of Rasch and CTT Methods for Scoring Motivation Scales. *Journal of Experimental Education*.

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement*, *62*(6), 921-943. doi:10.1177/0013164402238082

Macmillan, P. D. (2000). Classical, Generalizability, and Multifaceted Rasch detection of interrater variability in large, sparse data sets. *The Journal of Experimental Education*, *68*(2), 167-190. doi:10.1080/00220970009598501

Mallinson, T. (2011). Rasch analysis of repeated measures. *Rasch Measurement Transactions*, *251*(1), 1317. Retrieved from http://www.rasch.org/rmt/rmt251b.htm

Markowski, C. A., & Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. *The American Statistician*, *44*(4), 322-326. doi:10.1080/00031305.1990.10475752

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/BF02296272

McNeill, K. L., & Krajcik, J. (2008). Inquiry and scientific explanations: Helping students use evidence and reasoning. *Science as inquiry in the secondary setting*, 121-134.

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil*, *70*(4), 308-312. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2535599

Mergendoller, J. R., Maxwell, N. L., & Bellisimo, Y. (2006). The effectiveness of problem-based instruction: A comparative study of instructional methods and student characteristics. *Interdisciplinary Journal of Problem-based Learning*, *1*(2), 5. doi:10.7771/1541-5015.1026

Mergendoller, J. R., Maxwell, N. L., & Bellisimo, Y. (2000). Comparing problem-based learning and traditional instruction in high school economics. *The Journal of Educational Research*, *93*(6), 374-382. doi:10.1080/00220670009598732

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166. doi:10.1037/0033-2909.105.1.156

Myford, C. M., & Wolfe, E. W. (2004). Chapter 20: Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part I. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 460-517). Maple Grove, MN: JAM Press.

Neufeld, V. R., & Barrows, H. S. (1974). The "McMaster Philosophy": an approach to medical education. *Academic Medicine*, *49*(11), 1040-1050.

Nijsten, T., Unaeze, J., & Stern, R. S. (2006). Refinement and reduction of the Impact of Psoriasis Questionnaire: Classical Test Theory vs. Rasch analysis. *British Journal of Dermatology*, *154*(4), 692-700. doi:10.1111/j.1365-2133.2005.07066.x

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, *15*(5), 625-632. doi:10.1007/s10459-010-9222-y

Norman, G. R., & Schmidt, H. G. (1992). The psychological basis of problem-based learning: A review of the evidence. *Academic medicine*, *67*(9), 557-565. doi:10.1097/00001888-199209000-00002

Pell, G. (2005). Use and misuse of Likert scales. *Medical Education*, *39*(9), 970.

Pintrich, P. R., Smith, D., Garcia, T., & McKeachie, W. (1991). The motivational strategies for learning questionnaire (MSLQ). *Ann Arbor: University of Michigan*.

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes*, *1*(1), 27. doi:10.1186/1477-7525-1-27

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.).

Romanoski, J., & Douglas, G. (2002). Rasch-Transformed Raw Scores and Two-Way ANOVA: A Simulation Analysis. *Journal of Applied Measurement*, *3*(1), 420-428.

Salzberger, T. (2010). Does the Rasch model convert an ordinal scale into an interval scale? *Rasch Measurement Transactions*, *24*(2). Retrieved from http://www.rasch.org/rmt/rmt242a.htm

Savery, J. R., & Duffy, T. M. (1995). Chapter 11: Problem based learning: An instructional model and its constructivist framework. In B. G. Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design* (pp. 135-148). Englewood Cliffs, NJ: Educational Technology Publications, Inc.

Saye, J. W., & Brush, T. (1999). Student engagement with social issues in a multimedia-supported learning environment. *Theory & Research in Social Education*, *27*(4), 472-504. doi:10.1080/00933104.1999.10505891

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company. Retrieved from http://impact.cgiar.org/pdf/147.pdf

Smith, E. V., Jr. (2002). Understanding Rasch measurement: Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205-231.

Smith, E. V., Jr., & Kulikowich, J. M. (2004). An application of Generalizability Theory and Many-Facet Rasch Measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, *64*(4), 617-639. doi:10.1177/0013164404263876

Smith, E. V., Jr., & Smith, R. M. (2004). *Introduction to Rasch measurement: theory, models and applications*. Maple Grove, MN: JAM Press.

Smith, E. V., Jr., Wakely, M. B., de Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing rating scales for self-efficacy (and other) research. *Educational and Psychological Measurement*, *63*(3). doi:10.1177/0013164403063003002

Smith, E. V., Jr. (2004). Evidence of the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In J. E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications*. Maple Grove, MN: JAM Press.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, *10*(3), 516-517. Retrieved from http://www.rasch.org/rmt/rmt103a.htm

Son, B., & VanSickle, R. L. (1993). Problem-solving instruction and students' acquisition, retention and structuring of economics knowledge.

Song, M., & Herman, R. (2010). Critical issues and common pitfalls in designing and conducting impact studies in education: Lessons learned from the What Works Clearinghouse (Phase I). *Educational Evaluation and Policy Analysis*, *32*(3), 351-371. doi:10.3102/0162373710373389

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Stewart, J. (2012). Does IRT provide more sensitive measures of latent traits in statistical tests? An empirical examination. *Shiken Research Bulletin*, *16*(1), 15-22.

Strobel, J., & van Barneveld, A. (2009). When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. *Interdisciplinary Journal of Problem-based Learning*, *3*(1), 44-58. doi:10.7771/1541-5015.1046

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*(3), 239-261. doi:10.1016/j.asw.2004.11.001

Sungur, S., & Tekkaya, C. (2006). Effects of problem-based learning and traditional instruction on self-regulated learning. *The Journal of Educational Research*, *99*(5), 307-320. doi:10.3200/JOER.99.5.307-320

Sungur, S., Tekkaya, C., & Geban, Ö. (2006). Improving achievement through problem-based learning. *Journal of Biological Education Journal of Biological Education*, *40*(4), 155-160. doi:10.1080/00219266.2006.9656037

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston, MA: Allyn & Bacon/Pearson Education.

Tarhan, L., Ayar-Kayali, H., Urek, R., & Acar, B. (2008). Problem-based learning in 9th grade chemistry class: "Intermolecular forces". *Research in Science Education*, *38*(3), 285-300. doi:10.1007/s11165-007-9050-0

Thissen, D. (2016). Bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics*, *41*(1), 81-89. doi:10.3102/1076998615621300

Toulmin, S. E. (1958). *The philosophy of science* (14). Genesis Publishing Pvt Ltd.

Vernon, D. T., & Blake, R. L. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic medicine*, *68*(7), 550-563.

Viadero, D. (2009). "No effects" studies raising eyebrows. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2009/04/01/27rct_ep.h28.html

Walker, A., & Leary, H. (2009). A problem based learning meta analysis: Differences across problem types, implementation types, disciplines, and assessment levels. *Interdisciplinary Journal of Problem-based Learning*, *3*(1), 6.

What Works Clearinghouse (WWC) (2014). *Procedures and standards handbook Version 3.0*. Washington, DC: What Works Clearinghouse.

Wirkala, C., & Kuhn, D. (2011). Problem-based learning in K-12 education: Is it effective and how does it achieve its effects? *American Educational Research Journal*, *48*(5), 1157-1186. doi:10.3102/0002831211419491

Wolfe, E. W., & Chiu, C. W. T. (1999). Measuring pretest-posttest change with a Rasch rating scale model. *Journal of Outcome Measurement*, *3*(2), 134-161. Retrieved from http://jampress.org/JOM_V3N2.pdf

Wright, B. D. (1992). Point-biserial correlations and item fits. *Rasch Measurement Transactions*, *5*(4), 174. Retrieved from http://www.rasch.org/rmt/rmt54a.htm

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know*. Hillsdale, NJ: Lawrence Erlbaum Associates. Retrieved from http://www.rasch.org/memo64.htm

Wright, B. D. (2003). Rack and stack: Time 1 vs. time 2 or pre-test vs. post-test. *Rasch Measurement Transactions*, *17*(1), 905-906. Retrieved from http://www.rasch.org/rmt/rmt171a.htm

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of physical medicine and rehabilitation*, *70*(12), 857-860. Retrieved from http://www.rasch.org/memo44.htm

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370. Retrieved from http://www.rasch.org/rmt/rmt83b.htm

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.

Appendix A

Knowledge

Instructions

Identify the letter of the choice that best completes the statement or answers the question.

_____ 1. Most groundwater withdrawn worldwide is used for _____ .
A. Industry
B. Irrigation
C. Personal hygiene
D. Swimming pools

_____ 2. Water that is good enough to drink is called _____.
A. Potable water
B. Lake water
C. Sparkling water
D. Artesian water

_____ 3. What word means the change of state from liquid to a gas?
A. evaporation
B. condensation
C. transpiration
D. precipitation

_____ 4. How much of Earth's water is fresh water?
A. 97%
B. 50%
C. 3%
D. 1%

_____ 5. Transpiration is a process where water vapor
A. enters the atmosphere when animals breathe.
B. forms clouds.
C. exits a plant through holes in the leaves.
D. enters the atmosphere as water evaporates from the ground.

_____ 6. The process of evaporation, condensation, and precipitation is called the _____.
A. hydrologic cycle
B. life cycle
C. cellular cycle
D. geological cycle

_____ 7. What is the source of energy for the water cycle?
A. large clouds
B. rain droplets
C. the sun
D. the earth

_____ 8. Which of the following is a possible solution to water shortages?
A. reclamation of sewage water
B. desalination
C. developing drought-resistant crops
D. all of the above are possible solutions

_____9. Countries that are most likely to suffer from water stress would be located in
A. South America
B. Western Europe
C. The Middle East
D. North America

_____10. Desalination is the process of removing what from ocean water?
A. pollution
B. micro-organisms
C. chlorine
D. salt

_____11. Most of the water on the Earth is found in
A. the oceans
B. river and lakes
C. underground reservoirs
D. the polar icecaps

_____12. The scientific name for all the water on the Earth is
A. the atmosphere
B. the ocean
C. the lithosphere
D. the hydrosphere

_____13. How much of the Earth's surface is covered by water?
A. 97%
B. 75%
C. 3%
D. 25%

_____ 14. During the water cycle, when water vapor changes to liquid water, it is called
A. evaporation.
B. precipitation.
C. desalination
D. condensation.

_____ 15. The amount of water on the earth
A. is always increasing because of rain
B. is increasing because polar ice caps are melting
C. changes from year to year because of the weather
D. stays the same

_____ 16. Why is it important not to have a well for drinking water near a landfill?
A. pollution from the dump might leach down into the aquifer
B. it is alright to dig a well near a landfill
C. it might get hit by a garbage truck
D. the well might flood the garbage dump

_____ 17. Which of the following best describes the "hydrologic cycle"?
A. the interconnected, underground water movement system
B. the interconnected, endless movement of water in the Earth system
C. the interconnected, endless movement of streams in the Earth system
D. the interconnected, finite movement of water vapor in the Earth system

_____ 18. Water soaks into the ground through a process known as
A. infiltration
B. evaporation
C. equilibrium
D. transpiration

Appendix B

Interest

Strongly Disagree (1), Disagree (2), Undecided (3), Agree (4), Strongly Agree (5). Circle the response that best represents your opinion.

	Strongly Disagree				Strongly Agree
1. I enjoy going to science class.	**1**	**2**	**3**	**4**	**5**
2. I like learning about science in my free time.	**1**	**2**	**3**	**4**	**5**
3. Learning about science topics interests me.	**1**	**2**	**3**	**4**	**5**
4. Science inquiry is interesting to me.	**1**	**2**	**3**	**4**	**5**
5. Knowing about science is important to me.	**1**	**2**	**3**	**4**	**5**
6. I plan to become a scientist when I graduate.	**1**	**2**	**3**	**4**	**5**
7. When I graduate, I would like to work with people who make discoveries in science.	**1**	**2**	**3**	**4**	**5**
8. A career in science interests me.	**1**	**2**	**3**	**4**	**5**
9. I am interested in pursuing a science career in the future.	**1**	**2**	**3**	**4**	**5**
10. I am interested in pursuing a college degree in science.	**1**	**2**	**3**	**4**	**5**
11. I like my social studies class.	**1**	**2**	**3**	**4**	**5**
12. I am interested in other countries/cultures.	**1**	**2**	**3**	**4**	**5**
13. I think current events in the news are interesting.	**1**	**2**	**3**	**4**	**5**
14. Learning about international politics is interesting.	**1**	**2**	**3**	**4**	**5**

Appendix C

Self-efficacy

How confident are you that you can…

	Not Confident				Extremely Confident
1. write a well organized essay on a given topic?	**1**	**2**	**3**	**4**	**5**
2. draft a persuasive position on a given topic?	**1**	**2**	**3**	**4**	**5**
3. incorporate data into your essays?	**1**	**2**	**3**	**4**	**5**
4. write about science topics?	**1**	**2**	**3**	**4**	**5**
5. write a convincing argument?	**1**	**2**	**3**	**4**	**5**
6. get a good grade in science?	**1**	**2**	**3**	**4**	**5**
7. get a good grade in social studies?	**1**	**2**	**3**	**4**	**5**
8. learn science well?	**1**	**2**	**3**	**4**	**5**
9. learn social studies well?	**1**	**2**	**3**	**4**	**5**
10. learn how science and social studies are related?	**1**	**2**	**3**	**4**	**5**

Appendix D

Persuasive Essay on Social Studies and Science

Prompt: The world is in danger of running out of fresh water. Do you think this is true? Do you agree or disagree with this statement? Why?

Assignment: Write a **persuasive** essay stating your point of view on the prompt above. Give evidence to support your answer and provide your reasoning why this evidence supports your claim. Use your knowledge about water, science, world geography and cultures to help you write your response. You will have a total of 30 minutes to complete your essay.

Directions

Take a few minutes to plan your paper. Make notes on the other side of this page. An outline may help you plan well.

1. Decide if you **agree** or **disagree** that the world is in danger of running out of fresh water. Take **one** position on this issue.

2. Think of evidence that supports your position.

3. Think of reasons why this evidence supports your position.

4. Organize your ideas carefully.

5. Manage your time to allow for **writing** a closing statement.

After you have planned the paper, begin to write. Finally, proofread your finished paper to check for correct sentences, punctuation, and spelling.

Appendix E

GlobalEd 2 Writing Rubric: Claim-Evidence-Reasoning

This Rubric is designed for the pre- and post- GE2 essay prompts.

Raters: This is about CHAINS of logic that need to tie together. Quickly review the essay before scoring. If there is more than one chain, Identify the BEST single logic chain in the essay and score only that one. If there are multiple chains that are all the same quality, pick the first of these as the chain to code.
Of high Importance – DO NOT BE SWAYED by the "look" or length of an essay. Read it carefully! Neatness and quantity are NOT proxies for a well-formed essay! There are MANY examples of neat essays free of spelling and grammatical errors that contain a lot of content, but that are not advancing CER chains in a systematic way… You have to really concentrate on what CERs the student is trying to advance and divorce that from aesthetics!

What is a C-E-R logic chain? There are 3 parts to a CER logic chain: (1) The Claim; (2) The Evidence; and (3) The Reasoning. The claim is an assertion or conclusion that addresses the original inquiry question. The evidence is scientific data that supports the student's claim. This data can come from an experiment that students conduct or from another information source such as a journal or news article, a textbook, or a data archive. The data needs to be relevant to, and sufficiently support, the proposed claim. The reasoning provides a justification that **links** the claim and evidence and illustrates why the data counts as evidence to support the claim by using the appropriate scientific principles.
For a complete C-E-R all the components must be linked together, though the actual ORDER of the 3 components may vary form the standard C-E-R. It is NOT your job as a rater to re-structure the chain for the student. It must be a logical chain. You are NOT to "cherry pick" across the essay and "find" CER components that are present somewhere in the essay, but not intentionally linked by the student. Do NOT select unrelated links in the chain to increase the rating of the essay. The chain must be the Intent of the student BUT a chain CAN be inferred by the rater.

Example of a CER chain:

Claim	Evidence	Reasoning
"It is the position of the environmental committee that carbon sinks are an effective measure to reduce the amount of greenhouse gases in the atmosphere."	*"there have been successful uses of carbon sinks documented. For instance, forests planted in China 20 years ago have proved to be massive absorbers of carbon dioxide gas. "*	*"...by increasing the number of carbon sinks, we can reduce the amount of carbon dioxide as reduce the greenhouse effect."*

Student Essay Assignment

Prompt: The world is in danger of running out of fresh water. Do you think this is true? Do you agree or disagree with this statement? Why?

Assignment: Write a **persuasive** essay stating your point of view on the prompt above. Give evidence to support your answer and provide your reasoning why this evidence supports your claim. Use your knowledge about water, science, world geography and cultures to help you write your response.

RUBRIC Key

Essay Position (0-2)

What is the student's position in the essay?

0 = absent position or no position

Student has written about something unrelated to the topic of water shortages

Student has no position, for or against…. Agree or disagree.

1 = Support the statement – True; "We are running out of fresh water."

2 = Rejects the position; Not True; "We are **not** running out of fresh water."

3 - Student presents both positions (For and Against) BUT takes no position of what he/she believes; "The world may or may not be running out of fresh water."

Claim (0-2)

The Claim is NOT a restatement of the prompt. The Claim is the "*causal connector*" – the "*because*" of the essay. It is the statement addressing WHY they believe the world is or is not running out of fresh water. It does NOT have to contain a Direct Causal Connector. You as the rater can infer the missing the word "because".

0 = absent

Cannot discern a claim or claim DOES NOT RESPOND TO PROMPT.

1= Partial

The Claim can be inferred.

EX1: "Freshwater is being polluted…" – Here, you can infer that they are stating that the freshwater problem is due to pollution, although they have not expressly made that connection for you.

2 = Well developed

The Claim is clearly identifiable

EX1: "Water shortages world-wide are being caused by increased population."

EX2: "The world is running out of fresh water because there are more people than ever before."

Evidence – For the presented claim (0-3).

If there is no Claim, THEN there can be no Evidence or Reasoning

0 = Absent

None provided for the presented Claim

The evidence is missing or unrelated to the Claim.

1 = Partial

Provides some Evidence, but it is either weak or incomplete or the Evidence is related to the Claim <u>but</u> it requires an inference, rather than being clearly stated. The Evidence does <u>not</u> have to be specific data.

EX1: "People like me waste a lot of water"
EX2: "People waste water"
EX3: "Australia has a huge drought"

2 = Well developed

The evidence is related to the claim <u>and does not require an inference</u>; <u>clearly stated</u>.

EX1: "Since the turn of the century the population of the world has doubled."
EX2: "Due to climate change powerful storms are changing the fresh water distributions"

3 = Data Included for a 2 response

This is reserved for the highest level of evidence.
Clearly stated ***and*** *includes stated Reasonable Data.*

Reasoning (0-2)

If there is no Claim, THEN there can be no Evidence or Reasoning. Reasoning **<u>MUST</u>** provide the **<u>LINK</u>** between the Evidence and the Claim. The essay tells how the Claim and Evidence are linked together. This section must address the WHY portion of the prompt.

0 = Absent

Provides no reasoning LINKED to Claim and Evidence

1 = Partial

Reasoning LINK is incomplete or weak or clearly incorrect

EX: "This is why the population of the world matters."

2 = Well Developed

Reasoning is well thought out and clearly **LINKS** Claim and Evidence.

EX: "More people on earth means more people using freshwater – more people needing water is depleting our freshwater supply."

Holistic Section

Scores in this section are for the overall essay. Score the following sections based on the holistic nature of the essay on Addressing the Opposition, Organization, Science Content and Social Studies Content.

Addressing the Opposition (0-2)

This can occur anywhere in the essay, and does not need to be directly attached to the CER chain being coded above.

0 = Absent

No attempt to address the opposition OR opposing positions.

1 = Partial

The opposition is addressed, but it is done in an incomplete manner. No counter argument to the opposition presented.

Ex: *"... of course there are people who disagree..."*

2 = Well Developed

Addresses the opposition AND provide counter arguments.

Rationale – you recognize there are alternative views out there and you provide counter arguments refuting the opposition.

Ex: "Other people believe we have plenty of water, many of these individuals look at the ocean and see plenty of water, however the water in the oceans is salt water, not fresh water and cannot be used for human consumption."

Organization (0-2)

Score the organization of the essay holistically. You are judging the organization of the entire essay, overall.

0 = Disorganized, difficult for rater to follow coherent flow

1 = Clear attempt at organization but not optimized, thoughts not clearly flowing, *may or may not have a conclusion.*

2 = Coherent structure, including a conclusion

Science Content (0-3)

Score the science content of the essay. You are judging the science content of the entire essay, overall (climate, water cycle, pollution, earth science topics – ground water, etc.). Did the student mention science concepts?

0 = Absent
1 = Mentions a low level science content; colloquial terms
Examples: pollution, wasted water, lakes, streams oceans, environment
2 = Partially present but not complete OR using multiple scientific terms;
Examples: Point and non-point pollution; agricultural run-off; desalination; desertification, etc.
3 = Complete and Strong
Either an elaborated, accurate discussion of 1 science topic **OR** at least THREE (3) scientific terms (see 2 above).

Social Studies Content (0-3)

Score the social studies content of the essay holistically. You are judging the social studies /social systems content of the entire essay, overall (geography, politics, economics, culture, human rights …). Did the student mention social issues in their essay?

0 = Absent
1 = Mentions a low level social content; colloquial terms
EX: People working together; around the world; community
Generally mention the environment/landforms – lakes, rivers, …
All we need is money and we can fix the problem.

2 = Partially present but not complete OR using multiple social terms;
EX: Help from leaders around the world; regulation, laws/policies
They tie the environment to geography, such as tying deserts, rain forest, etc. to water resources in different parts of the world.
Some countries have a lot of money and others have very little money.

3 = Complete and Strong
Either an elaborative discussion of 1 theme of social science **OR THREE** different social systems **at a level 2** – economics, geography, politics…

Curriculum Vita

Lisa J. Lynn

University Address
Department of Educational Psychology
University of Illinois at Chicago
1040 West Harrison St. M/C 147
Chicago, IL 60607

Home Address
1017 South Belmont Avenue
Arlington Heights, IL 60005
(224) 764-1017
lsteve5@uic.edu

Education

2010 M.Ed., Measurement, Evaluation, Statistics, and Assessment
University of Illinois at Chicago, Chicago, Illinois

2009 B.A., Psychology
University of Illinois at Chicago, Chicago, Illinois

Publications

Under review

Lynn, L. J., Yukhymenko, M. A., & Lawless, K. A. (under review). A comparison of Rasch and CTT methods for scoring motivation scales.

In preparation

Lynn, L. J., Shea, K. & Lawless, K. A. (in preparation). Conceptual change in an online socio-scientific learning simulation.

2015 **Lynn, L. J.**, & Lawless, K. A. (2015). Chapter 1. Observed score versus Rasch score analysis of efficacy data: A case study comparison. In Columbus, A (Ed.) *Advances in Psychology Research. Volume 109.* Hauppauge NY: Nova.

2015 Riel, J., Lawless, K. A., Brown, S. W., & **Lynn, L. J.** (2015). Teacher participation in ongoing online professional development to support curriculum implementation: Effects of the GlobalEd 2 PD program on student affective learning outcomes. *Proceedings of the 2015 Annual Meeting of the Society for Information Technology and Teacher Education, Las Vegas, NV.* Waynesville, NC: SITE.

2014 **Lynn, L. J.**, Lawless, K. A., Brown, S. W., Brodowinska, K., Mullin, G., Powell, N., Richards, K. A., & Boyer, M. A. (2014). Science

vocabulary development in a problem-based learning simulation. *National Teacher Education Journal 7*(3), 5-12.

2012 Lawless, K. A., Brodowinska, K. B., **Lynn, L. J.**, Khodos, G. A., Brown, S. W., Boyer, M. A., Yukhymenko, M., Mullin, G., & Le, L. (2012). Chapter 4. The GlobalEd 2 game: developing scientific literacy skills through interdisciplinary, technology-based simulations. In Baek, Y (Ed.) *Psychology of Gaming*. Hauppauge NY: Nova.

2011 Lawless, K. A., Brown, S.W., Boyer, M.A., Brodowinska, K., Mullin, G.P., Yukhymenko, M., Khodos, G., **Lynn, L.**, Cutter, A., Powell, N. & Fernada Enriquez, M., (2011). Expanding the science and writing curricular space: The GlobalEd2 Project. In D. Sampson, J. M. Spector, D. Ifenthaler and P. Isaias (Eds.) *Proceedings of The IADIS International Conference Cognition and Exploratory Learning in Digital Age (CELDA)*, p. 154-160. Rio de Janeiro, Brazil: International Association for Development of the Information Society.

Awards and Honors

2016 Excellence in Undergraduate Mentoring Award, Honors College and Graduate College, University of Illinois at Chicago

2016 Chancellor's Student Service Award (CSSA), University of Illinois at Chicago

2015 – 2016 100-Hour Undergraduate Research Internship (HURI), College of Education, University of Illinois at Chicago

Conference Presentations

2017 (accepted)
Lynn, L. J. & Lawless, K. A. (2017, April). Assessing Argumentative Writing: A Facets Analysis of Rater Use of a Claim-Evidence-Reasoning Rubric. Paper presentation at American Educational Research Association Annual Meeting, San Antonio, TX.

2017 (accepted)

Oren, J. B., Lawless, K. A., Brown, S. W., Riel J., Lynn, L. J., Bruscianelli, K. B. (2017, April). Exploring Teachers' Facilitative Behaviors Across Implementation Years. Roundtable presentation at American Educational Research Association Annual Meeting, San Antonio, TX.

2016 Wang, M.S., Morassini, M., Newton, S.D., Song, S., Zhao, A., Brown, S.W., Lawless, K.A., **Lynn, L.**, Riel, J., Bruscianelli, K., & Oren, J. (2016, May). *Prior knowledge and scientific literacy: An inspection on the contemporary knowledge sources and effects on scientific writing.* Paper presented at The 6th International Symposium on Society, Education and Psychology (ISSTEP 2016), Beijing, China.

2016 Wang, M.S., Morassini, M., Newton, S.D., Song, S., Zhao, A., Brown, S.W., Lawless, K.A., **Lynn, L.**, Riel, J., Bruscianelli, K., & Oren, J. (2016, May). *Capitalizing on students' prior knowledge to improve scientific writing proficiency.* Poster presented at 2016 APS Convention, Chicago, IL.

2016 Song, S., Newton, S.D., Wang, M.S., Morassini, M., Zhao, A., Brown, S.W., Lawless, K.A., **Lynn, L.**, Riel, J., Bruscianelli, K., & Oren, J. (2016, May). *The influence of gender and internet access on academic self-efficacy in middle school students.* Poster presented at 2016 APS Convention, Chicago, IL.

2016 Brown, S.W., Lawless, K.A., Rhoads, C., Newton, S.D., & **Lynn, L.** (2016, Oct.). Increasing students' science writing skills through a PBL simulation. In D. Sampson, J.M. Spector, D. Ifenthaler & P. Isaias (Eds.) *Proceedings of The 13th IADIS International Conference Cognition and Exploratory Learning in Digital Age (CELDA),* p. 86-94. Mannhiem, Germany: International Association for Development of the Information Society.

2016 **Lynn, L. J.**, Yukhymenko, M. A., & Lawless, K. A. (2016, May). *Similar scores but differential conclusions produced by Rasch and Classical Test Theory (CTT) analyses.* Poster presentation at APS conference, Chicago IL.

2016 Bruscianelli, K. B., Lawless, K. A., Brown, S. W., Zhao, A., Oren, J. B., Riel, J., **Lynn, L. J.** (2016, April). *Examining Differences Among Teacher Implementation Patterns.* Paper presentation at American Educational Research Association Annual Meeting, Washington, DC.

2016 Lawless, K. A., Brown, S. W., **Lynn, L. J.**, Bruscianelli, K. B., Riel, J., Oren, J. (2016, April*). Improving Written Argumentation Through Web-Based, Interdisciplinary Simulations: The GlobalEd2 Project.* Paper presentation at American Educational Research Association Annual Meeting, Washington, DC.

2016 Oren, J. B., Lawless, K. A., Brown, S. W., Bruscianelli, K. B., Riel, J., **Lynn, L. J.** (2016, April*). Improving Written Argumentation Through Web-Based, Interdisciplinary Simulations: The GlobalEd2 Project.* Paper presentation at American Educational Research Association Annual Meeting, Washington, DC.

2016 Brown, S.W., Lawless, K.A., Riel, J., Wang, M., **Lynn, L.**, Newton, S., Bruscianelli, K., Zhao, A., Song, S. & Oren, J. (2016, January). *Promoting STEM Literacies and Attitudes in Students through an Online Problem-Based Learning Simulation of International Negotiation.* Paper presented at the 2016 Hawaiian International Conference on Education, Honolulu, HI.

2016 **Lynn, L. J.**, Shea, K. M, & Lawless, K. A. (2016, January). *Science conceptual change in an online, socio-scientific problem-based learning simulation.* Poster presentation at UIC College of Education Annual Research Day, Chicago IL.

2015 **Lynn, L. J.**, Lawless, K. A., Brown, S. W., Brodowinska Bruscianelli, K., & Riel, J. (2015, April). *Socioscientific conceptual change in an online problem-based learning scenario.* Roundtable presentation at AERA conference, Chicago IL.

2015 Shea, K. M., **Lynn, L. J.**, Bruscianelli, K. B., Riel, J., & Lawless, K. A., (2015, April). *Teachers' reported use of assessment in GlobalEd 2.* Poster presentation at UIC Student Research Forum, Chicago IL.

2015 Brown, S. W., Lawless, K. A., **Lynn, L. J.**, Brodowinska Bruscianelli, K., & Riel, J. (2015, April). *Promoting science knowledge through an educational game: Global Ed 2*. Presentation at AERA conference, Chicago IL.

2015 Lawless, K. A., Brown, S. W., **Lynn, L. J.**, Bruscianelli, K.B., Riel, J., Field, K., Dye, C., Le-Gervais, L., Lin-Steadman, P., & Alanzi, R. (2015, April). *Global Ed 2: A problem-based, interdisciplinary simulation targeted at written argumentation*. Presentation at AERA conference, Chicago IL.

2015 Lawless, K.A., Brown, S.W. & The GlobalEd 2 Team (Alanazi, R., Brodowinska, K., Dye, C., Field, K., Le-Gervais, L., Lin-Steadman, P., **Lynn, L.**, Riel, J.) (2015, March). An Interdisciplinary Approach for the Teaching of Written Argumentation: GlobalEd 2. Presented at the International Convention of Psychological Science (Teaching Institute), Amsterdam, NL.

2015 Lawless, K.A., Brown, S.W. & The GlobalEd 2 Team (**Lynn, L.**, Brodowinska, K., Field, K., Riel, J., Dye, C., Le-Gervais, L., Alanazi, R.) (2015, February). *Promoting science knowledge and interest: The GlobalEd 2 Project.* Paper presented at the Eastern Educational Research Association Conference, Sarasota, FL.

2014 **Lynn, L. J.**, Lawless, K., & Brown, S. (2014, October). Comparing Rasch and classical test theory methods for assessment validation in GlobalEd 2. Poster presentation. UIC College of Education Annual Research Day, Chicago, IL.

2014 Brodowinska, K., **Lynn, L. J.**, Lawless, K. A., Brown, S. W., Boyer, M. A., O'Brien, D. W., Cutter, A., Enriquez, M., Khodos, G. A., Maneggia, D., Mullin, G. P., Powell, N., & Williams, S. G. (2014, October). *Teachers' approaches to implementing a problem-based learning simulation.* Poster presentation. University of Illinois at Chicago College of Education Annual Research Day.

2014 Riel, J., Lawless, K., and **Lynn, L.** (2014, October). *Teacher participation in ongoing professional development using email newsletters and teacher journaling: The GlobalEd 2 online professional development program*. Poster presentation. University of Illinois at Chicago College of Education Annual Research Day.

2014 Brown, S.W., Lawless, K.A., Boyer, M.A., Yukhymenko, M.A., Brodowinska Bruscianelli, K., **Lynn. L.**, & Mullin, G.P. (2014, May). *Promoting students' writing about science skills through PBL: GlobalEd 2.* Poster presented at the 2014 APS Conference, San Francisco, CA.

2014 Brown, S.W., Lawless, K.A., & the GlobalEd 2 Team (Alanazi, R., Brodowinska, K., Dye, C., Field, K., Le-Gervais, L., Lin-Steadman, P., **Lynn, L.**, Mullin, G., Powell, N., & Riel. J.) (2014, April). *Simulating the negotiations of international science advisers: The GlobalEd 2 Project.* Paper presented at the 2014 AERA Conference, Philadelphia, PA.

2014 Lawless, K.A., Brown, S.W., & The GlobalEd 2 Team (Alanazi, R., Brodowinska, K., Dye, C., Field, K., Le-Gervais, L., Lin-Steadman, P., **Lynn, L.**, Riel. J.) (2014, April). *Developing a scientifically literate citizenry.* Invited paper presented at the 2014 AERA conference, Philadelphia, PA.

2014 **Lynn, L. J.**, Brodowinksa Bruscianelli, K., Brown, S. W., Mullin, G. P., Yukhymenko, M. A., & Boyer, M. A. (2014, April). Development of science conceptual knowledge in an online learning simulation. Roundtable presentation at AERA conference, Philadelphia, PA.

2014 Lawless, K.A., Brown, S.W., Brodowinska, K., Field, K., **Lynn, L.**, Riel, J., Le-Gervais, L., Dye, C. & Alanazi, R. (2014, February). *Expanding the science and literacy curricular space: The GlobalEd 2 Project.* Paper presented at Eastern Educational Research Association Annual Meeting, Jacksonville, FL.

2014 Brown, S.W., Boyer, M.A., Lawless, K.A., Yukhymenko, M.A., Mullin, G.P., Gervais, L.L., **Lynn, L.**, Brodowinska, K., & Khodos, G. (2014, January). *The impact of an international simulation game on students' academic self-efficacy and social perspective taking.* Paper presented at the 2014 Hawaii International Conference on Education, Honolulu, HI.

2013 Lawless, K. A., Brown, S. W., Brodowinska, K., **Lynn, L.**, Riel, J., Lee, L., Mullin, G., & Managgia, D. (2013, October). *Using digital communications to create a scientifically literate citizenry – The GlobalEd 2 Project.* University of Illinois at Chicago, College of Education Research Day.

2013 Lawless, K.A., Brown, S.W., Boyer, M.A., Brodowinska, K., Khodos, G., **Lynn, L.**, Yukhymenko, M.A., & Mullin, G.P. (2013, June.). *The Differential Impact of Implementation Fidelity in a Technology Enhanced PBL Environment: The GlobalEd2 Project.* Paper presented at the EdMedia 2013 Conference, Victoria, BC, Canada.

2013 Brown, S.W., Lawless, K.A. Boyer, M.A. & The GlobalEd 2 Team. (2013, May). The impact of a PBL simulation on college students' self-efficacy. Poster presented at the Annual Association of Psychological Science Convention on May 26, 2013, Washington, DC.

2013 Lawless, K.A., Brown, S.W., Boyer, M.A., Brodowinska, K., Khodos, G., **Lynn, L.**, Yukhymenko, M.A., Mullin, G.P., and Gervais, L.L. (2013, March). *The GlobalEd 2 game: Developing scientific literacy skills through interdisciplinary, technology-based simulations.* Paper presented at the 2013 Society for Information Technology and Teacher Education Conference, New Orleans, LA.

2013 Brown, S.W., Lawless, K.A., Boyer, M.A., & The GlobalEd 2 Team (Yukhymenko, M.A., Mullin, G.P., Brodowinska, K., Khodos, G., **Lynn, L.** & Gervais, L.L.). (2013, February). *Promoting middle school writing skills in science through an educational game: GlobalEd 2*. Paper presented at the 2013 Eastern Educational Research Association Conference, Sarasota, FL.

2012 Lawless, K.A., Brown, S.W., Brodowinska, K., **Lynn, L.**, Khodos, G., Mullin, G.P., Yukhymenko, M., & Boyer, M.A. (2012, June). Distributed learning environments: Cooperative/collaborative learning. Paper presented at EdMedia 2012, Denver, CO.

2012 Brown, S.W., Lawless, K.A., Boyer, M.A., Yukhymenko, M.A., Mullin, G.P., Brodowinska, K., Khodos, G., & **Lynn, L.** (2012, May). The impact of simulation games on science knowledge: The GlobalEd 2 Project. Poster presented at the 2012 Association for Psychology Science: Annual Convention, Chicago, IL.

2012 **Lynn, L. J.**, Lawless, K. A., Brown, S. W., Bruscianelli, K. B., Mullin, G., Powell, N., Richards, K. A., Boyer, M. A. (2012, May). Development of science vocabulary in an online learning environment. Poster presentation at APS conference, Chicago, IL.

2012 Bruscianelli, K. B., **Lynn, L. J.**, Lawless, K. A., Brown, S. W., Boyer, M. A., O'Brien, D. W., Cutter, A., Enriquez, M. F., Khodos, G. A., Maneggia, D., Mullin, G., Powell, N., and Williams, G. (2012, April). *Teachers' Varied Approaches to Implementing a PBL, GlobalEd 2 Simulation: An Evolved Analysis.* Roundtable presentation at AERA conference, Vancouver, B.C.

2012 **Lynn, L. J.**, Lawless, K. A., Brown, S. W., Bruscianelli, K. B., Powell, N., Mullin, G., Richards, K. A., Boyer, M. A. (2012, April). Science vocabulary development in a problem-based learning simulation. Paper presentation at AERA conference, Vancouver, B.C.

2012 Brown, S.W., Lawless, K. A., Boyer, M.A., Yukhymenko, M., Mullin, G.P., Brodowinska, Khodos, G. K., Powell, N., & **Lynn, L.** (2012, February). *Increasing technology and writing self-efficacy through a PBL simulation: GlobalEd 2.* Paper presented at the Eastern Educational Research Association Conference; Hilton Head, SC.

2012 Lawless, K. A., Brown, S.W., Boyer, M.A., Brodowinska, K., **Lynn, L.**, Mullin, G.P., & Yukhymenko, M., (2012, February). *Developing scientific literacy skills through interdisciplinary, technology-based global simulations: GlobalEd 2.* Paper presented at the Eastern Educational Research Association Conference; Hilton Head, SC.

2012 Yukhymenko, M., Brown, S.W., Lawless, K. A., Boyer, M.A., Mullin, G.P., Brodowinska, Khodos, G. K., Powell, N., & **Lynn, L.** (2012, February). *The GlobalEd Project in middle school: Effectiveness of ICT use for writing social studies PBL classrooms.* Paper presented at the Eastern Educational Research Association Conference; Hilton Head, SC.

2011 Lawless, K. A., Brown, S.W., Boyer, M.A., Brodowinska, K., Mullin, G.P., Yukhymenko, M., Khodos, G., **Lynn, L.**, Cutter, A., Powell, N. & Fernada Enriquez, M., (2011, November). Expanding the science and writing curricular space: The GlobalEd2 Project. Paper presented at the IADIS International Conference Cognition and Exploratory Learning in Digital Age, Rio de Janeiro, Brazil.

2011 Bruscianelli, K. B., **Lynn, L. J.**, Lawless, K. A., Brown, S. W., Boyer, M. A., O'Brien, D. W., Cutter, A., Enriquez, M. F., Khodos, G. A., Maneggia, D., Powell, N., and Williams, G. (2011, April). *Teachers' Varied Approaches to Implementing a Problem-Based Learning (PBL) Simulation: GlobalEd 2 Project*. Paper presentation at AERA conference, New Orleans, LA.

Teaching Experience

Graduate

Introduction to Quantitative Analysis of Educational Data Summers 2012, 2013
Primary instructor, online course (overall evaluation 4.1/5.0)
Primary instructor, classroom course (overall evaluation 4.2/5.0)

Undergraduate

Laboratory in Developmental Psychology Spring 2013
Teaching assistant, classroom course (overall evaluation 3.8/4.0)

Assessment in the Urban Elementary Classroom I & II
Spring 2010 – Spring 2013
Primary instructor, online course (overall evaluation 4.3/5.0)
Primary instructor, classroom course (overall evaluation 4.9/5.0)

Integrating Technology in the Elementary Classroom I & II
Fall 2010 – Spring 2011
Teaching assistant, blended course (evaluation not available)

Research Experience

UIC INCLUDES Fall 2016 – Spring 2017
Department of Educational Psychology, University of Illinois at Chicago
Visiting Research Specialist

GlobalEd 2 Fall 2013 – Spring 2017
Spring 2010 – Fall 2012
Department of Educational Psychology, University of Illinois at Chicago
Data manager and researcher on a multi-year, IES grant-funded educational intervention efficacy trial.

Service to Profession

2016 AACE E-Learn 2016 Program Committee
2015 CAURS Research Symposium Judge
2013 AERA Review Panel: Studying and Self-regulated Learning
2013 AERA Review Panel: Technology Research
2013 APA Review Panel: Division 15 Educational Psychology
2013 APS Student Caucus RISE Award Review Panel

University Service

2016 – 2017 Alumni Volunteer, UIC Honors College
2014 – 2016 Executive Editor, Interdisciplinary Undergraduate Research Journal
2012 – 2013 UIC Graduate Student Council

Other Work

2013 Evaluation Specialist, PIE (Program Implementation and Evaluation) Consulting, Chicago, IL

2012 – 2013 Statistical Consultant, Bilingualism Research Laboratory, College of Liberal Arts and Sciences, University of Illinois at Chicago

Professional Memberships

2009 American Educational Research Association (AERA)
2012 Association for Psychological Science (APS)