Measurement invariance techniques to enhance measurement sensitivity

Brian F. French*
Department of Educational Leadership and Counseling Psychology,
Washington State University,
Pullman, Washington, 99164, USA
Email: frenchb@wsu.edu

W. Holmes Finch*
Department of Educational Psychology,
Ball State University,
Muncie, IN 47306, USA
Email: whfinch@bsu.edu
*Corresponding authors

Bruce Randel
Century Analytics, Inc.,
7224 South Garfield Street, Centennial,
CO 80122, USA
Email: bruce.randel@centuryanalytics.com

Brian Hand
College of Education,
University of Iowa,
Teaching and Learning,
N259 Lindquist Center,
Iowa City, IA, 52240, USA
Email: brian-hand@uiowa.edu

Chad M. Gotch
Department of Educational Leadership and Counseling Psychology,
Washington State University,
Pullman, Washington, 99164, USA
Email: cgotch@wsu.edu

Abstract: Rigorous evidence supporting the effectiveness of interventions is needed to inform teaching practice and improve educational outcomes. In many instances, gathering such evidence includes cluster randomised control trials estimating the effectiveness of educational treatments. Such studies often require the collection of data from large samples in order to accurately detect
intervention effects. A failure to detect these effects could be due to the inability of the intervention to produce effects or due to a lack of measurement sensitivity to the intervention itself. The current study outlines a two-stage method for evaluating measurement sensitivity by first conducting content analysis to align items with hypothesised intervention effects, followed by the use of differential item functioning analyses to detect intervention effects more precisely, and thereby test for measurement sensitivity. Increasing measurement sensitivity could lead to increased effect sizes, increased statistical power, reduced sample sizes and reduced costs.

**Keywords:** measurement sensitivity; differential item functioning; randomised control trials; measurement invariance; differential item functioning; DIF; educational research.


**Biographical notes:** Brian F. French is a Professor in the College of Education at Washington State University. His research focuses of applied and methodological issues related to test score validity.

W. Holmes Finch is a Professor in the College of Education at Ball State University. His research interests involve issues in psychometrics, including dimensionality assessment, differential item functioning, generalisability theory and unfolding models.

Bruce Randel is President and Consultant at Century Analytics, Inc., a small business providing comprehensive technical assistance and analytic services in research methods and design, instrument development, psychometrics, statistics and data analysis, and reporting.

Brian Hand is a Professor in the College of Education at the University of Iowa. His research focuses on how we can use language as a learning tool to improve students’ understanding of science. His work has focused on using writing as a learning tool and is now moving to focus on the use of multi-modal representation within science classrooms. He also focuses on the development of scientific argument through the use of the science writing heuristic.

Chad M. Gotch’s interests centre on maximising effective and proper use of educational and psychological measurements. To this end, he studies assessment/measurement literacy among teachers, score reporting, and building validity arguments from both technical and non-technical evidence. These complimentary lines of research inform the life cycle of assessment, from development to use and policy.

This paper is a revised and expanded version of a paper entitled ‘Measurement invariance techniques to enhance measurement sensitivity’ presented at the National Council on Measurement in Education Conference, San Francisco, CA, 17–19 April 2013.
1 Introduction

When conducting educational and psychological research, a key choice that researchers must make, and one that will directly impact the observed effect of interest in an intervention study, is the measurement tool to be used for assessing the outcome of interest. Indeed, research suggests that the measure operationalising the outcome of interest appears to be associated with at least as much variance in observed effects as do other features of the design (e.g., comparison group, random versus non-random assignment; Wilson and Lipsey, 2001). In other words, the instrument that is selected for measuring the outcome of interest can, in some cases, have as much impact on the final assessment of the intervention as the impact of the intervention itself. Measurement, therefore, is a non-trivial aspect of efficacy and effectiveness research, regardless of the field of study, including assessments of teacher effectiveness. This impact of the instrument on the size of the observed effect comes from its sensitivity, or its ability to detect the expected change in the construct of interest.

In educational research, many efficacy and effectiveness studies use standardised tests, in part for their policy relevance, but also for their demonstrated validity, standardisation, and relative ease of administration to large groups. However, most of these standardised tests are designed to measure very general constructs (e.g., reading proficiency), often for accountability purposes, and therefore may not be particularly sensitive to specific abilities that are influenced by fairly discrete changes in instruction (Popham, 2007a; Popham and Ryan, 2012). In other words, the purpose of most standardised tests is to provide stable estimates of individuals’ general abilities so that educators can ensure that they meet some basic skill requirements. As such, they are not optimised to detect hypothesised changes in achievement that result from a specific intervention (e.g., Hevey et al., 2004) such as a new instructional technique, which may be designed to influence a small number of very specific skills, rather than the more global abilities measured by standardised instruments. Unfortunately, there is no agreement on how best to evaluate measurement sensitivity (Polikoff, 2010). Thus, the goal of this research is:

a to outline an innovative solution to this problem involving new methods to evaluate measurement sensitivity

b to provide a new paradigm for the development of more sensitive measures for assessing important but relatively specific outcomes that are a function of a particular intervention.

1.1 Measurement sensitivity

Measurement sensitivity is paramount for research directed at the evaluation of the impacts of programs; treatments in the social, behavioural and clinical sciences; and quality of instruction, among other areas. In fact, embedded in accountability systems is the implicit and sometimes explicit assumption that the measures used are sensitive enough to capture change or growth on the target skills. Such sensitivity is especially relevant in applied settings where program or treatment effects are often not particularly strong, and measurement conditions can be quite variable (Lipsey, 1983). The statistical noise created by this high degree of variability, coupled with the relatively small effects
of interventions in fields such as education, psychology, and medicine, mean that small effects are often difficult to detect with the desired degree of statistical precision. As a result, large samples are frequently needed to achieve adequate statistical power in order to detect the effect of these interventions. It is important to keep in mind however, that the measured size of the intervention effect is not merely the product of the intervention itself. Many of the decisions made by the researchers in planning their study – design, sample selection and choice of outcome measures – may have as much of an influence on the measured size of the effect as the intervention itself (Wilson and Lipsey, 2001). Of particular interest in this study is the case where the selected outcome measures are in fact insensitive to the construct(s) that are impacted by the intervention. Such insensitivity may in turn lead to false conclusions about intervention or program effects, and result in increased research costs particularly due to the need for large samples, particularly in cluster randomised trials. Unfortunately, there is no agreement on how best to evaluate measurement sensitivity, and every current approach has limitations (Polikoff, 2010). To tackle this problem we propose the use of measurement invariance (MI) methods in a two-step process to evaluate measurement sensitivity, in accord with suggestions of such use (Millsap, 2011; Popham, 2007a; Popham and Ryan, 2012). With accountability on the forefront of so many fields, including education, healthcare and psychology, the time is appropriate to propose an innovative method that can be used across many fields to evaluate measurement sensitivity and support the development of more sensitive measures.

As noted above, standardised measures are popular tools given their policy relevance and technical merits, but may not be optimised to detect changes in very specific skill sets (Popham, 2007b). Similarly, in clinical health settings, research instruments that are designed to measure specific traits or conditions are typically more sensitive to the effects of interventions than are generic health status measures (Wright and Young, 1997). Standardised assessments that measure global traits, therefore, may not be sufficiently sensitive to intervention or program effects. Some standardised measures are even specifically designed to be insensitive to instruction. Aptitude tests, for example, use items and item formats that are dissimilar to those of achievement tests so that examinees must apply reasoning skills that are not taught in the classroom through instruction. At the same time, because of their ubiquity, convenience and technical strengths, such general measures are quite often used as the focal outcomes in many intervention studies.

### 1.2 Impact of sensitivity on effect sizes

The sensitivity of an outcome measure to an intervention is strongly related to the size of the intervention effect. In general, effect sizes will vary depending on the degree to which the outcome measure aligns with the intervention and focuses on the knowledge and skills targeted by the intervention, as well as to the actual impact of the intervention itself. Measures that are more closely aligned with and focused on the content and intent of the intervention will be associated with larger observed effect sizes (Bloom, 2007). For example, in a study in which

- broad standardised tests associated with no child left behind (NCLB)

- standardised tests with content aligned to an intervention
specialised tests that target the focus of an intervention were compared on educational interventions, mean effect sizes of 0.07, 0.23, and 0.44, respectively, were reported (Bloom, 2007).

In other words, the more specialised tests that were focused on the actual outcome intended to be influenced by the intervention provided larger effect sizes, when compared with more insensitive measures designed to assess broad outcomes. Bloom’s data support previous research (e.g., Tobias, 1982) identifying the importance of measurement sensitivity to detecting the effects of interventions. Such examples demonstrate that measurement sensitivity has implications for detecting the influence of interventions, particularly demonstrating that the minimally detectable effect (MDE) is likely to be small for an insensitive measure (Bloom, 2007). However, an associated caveat is that specialised measures designed for a specific intervention require time and resources to develop and may not have the necessary psychometric evidence supporting them (i.e., reliability and validity), when compared to commercially developed measures. Such specialised instruments also may lack measurement sensitivity because the current approaches available to evaluate measurement sensitivity have several limitations and may not be easily implemented.

1.3 Instructional sensitivity

Evaluating the sensitivity of outcome measures has been an important part of research in a variety of fields. In educational research, the sensitivity of instruments is important to accountability assessment, instructional improvement, and evaluation of program effectiveness, with the greatest emphasis being on instructional sensitivity. Instructional sensitivity refers to the degree to which students’ performance on a test accurately reflects the quality of instruction that was provided to promote student mastery of the skill(s) being assessed. Instructionally sensitive tests should be capable of distinguishing between strong and weak instruction such that high performance on the test is attributable to good instruction (Polikoff, 2010; Popham, 2007a). Instructional sensitivity also has been labelled instructional validity, and much of the emphasis in the area of instructional sensitivity focuses on the validity of scores from instruments designed to assess instructional outcomes. Sensitivity is a concern when considering test score validation because scores from an insensitive test cannot appropriately capture changes in student learning. In turn, this failure to capture the impact on student learning that results from changes in instruction has implications for effectiveness research in that truly effective interventions may not be shown so by insensitive measurement tools.

Instructional sensitivity has its roots in and borrows much from research in the area of opportunity to learn (OTL). OTL research broadly addresses what students are taught and tested on as well as the equity of instructional opportunities, whereas instructional sensitivity research more specifically addresses the extent to which a test can detect the impact of instruction. Examining the relationship between instruction, as captured by measures of OTL, and student achievement is a common approach to examining instructional sensitivity and can include both expert judgement and empirical methods (Popham, 2007a; Popham and Ryan, 2012). A variety of approaches are used to measure OTL, including direct observation of instruction, student self-report surveys, artefacts and assignment sampling. The disadvantage of the OTL approach to assessing instructional sensitivity is that data collection and measurement of classroom practice can be
expensive and time consuming. To appropriately carry out an OTL study, these measurements of classroom practice need to be made precisely and in detail for each specific context (Muthén et al., 1995). Although many methods and measures are available, the heterogeneity of instruction from one intervention study to another makes capturing their impact accurately a continuing challenge.

OTL studies of sensitivity also should include background characteristics of students in the analyses (Burstein et al., 1989; Muthén et al., 1995). Various methods have been proposed to account for the relationships between OTL, prior performance, student background, and achievement. However, regardless of the method used, limitations due to the availability and accuracy of the collected data remain. Thus, given the need to create measures designed for assessing specific interventions for each new study, OTL research may not be the most effective and efficient way to evaluate measurement sensitivity.

1.4 Measuring sensitivity at the item level

Although much prior work in the assessment of program effects has utilised total test scores, evaluation of sensitivity can be approached at the item and/or the item bundle level as well. Much of the research to identify items that are sensitive to intervention effects has focused on statistical indices (see Haladyna and Roid, 1981) and models (e.g., item response theory) to examine the item characteristics. This item level approach to sensitivity assessment compares specific item characteristics (e.g., item difficulty) for the same subjects before and after instruction between different groups of students who received and did not receive instruction.

In this context, the ultimate sensitivity of the entire test to instruction is generally not of interest, nor is the sensitivity of the instrument further examined for evidence of general score validity. Rather, specific items that should theoretically be sensitive to the instruction are of primary interest. Examination of the sensitivity of items to specific interventions, therefore, continues to be recommended (Millsap, 2011; Popham, 2007a; Popham and Ryan, 2012).

Schmidt et al. (1998) argue for the item level approach to sensitivity evaluation for four reasons. First, the use of composite scores in effectiveness studies assumes the measurement of a unidimensional trait, which may not be true for broad measures of achievement. Second, a composite score from a test including items not relevant to the curricula or instruction will introduce extra variation, through measurement insensitivity, to the assessment of student ability in the specific instructional target. As noted above, a composite score is typically a measure of general ability in some domain, while curricula and instruction typically target specific skills that may be related to only a small part of the global, underlying trait (Airasian and Madaus, 1983). The third reason in favour of item level sensitivity evaluation, which is closely allied to the second, is that instruments across various fields of inquiry are typically composed of multiple items selected in order to maximise internal consistency reliability. Many of the items on these scales are of average difficulty (Liang et al., 2002) with only a few being relatively difficult or easy. An instrument dominated by such items will not be equally discriminating across the entire ability distribution. Thus, a meaningful change for a person who is of high or low ability may not be detected by analyses at the total score level, because so few items are
Included that would be sensitive at these extremes. Finally, current methods for analyses at the item level may provide more relevant, fine-grained information regarding instruction as compared to general analyses of the composite score.

1.5 The need for new methods to evaluate sensitivity

Sensitive measures are generally not available in many areas in which interventions are routinely used and studied. Standardised measures developed to provide reliable and valid score estimates of individual abilities are typically too broad to be sensitive to the unique aspects of particular interventions. In addition, educational interventions vary largely in their approach to increasing students’ performance, as measured by a test score, such that it may be very difficult for a single standardised test to accurately detect the effects of the various foci of many interventions. Even a state-wide achievement test aligned with state standards may cover too broad a construct area to be sensitive to the changes resulting from a method of instruction targeting a small number of skill areas.

On the other hand, while researcher-developed measures for specific interventions are typically well-aligned and more sensitive to an intervention’s hypothesised effects, they can suffer from poor psychometric quality (e.g., Belland et al., 2009). Low internal consistency reliability, for instance, will attenuate correlations and mask potentially informative results. In addition, the development of specific well-aligned and psychometrically sound instruments for each intervention is time consuming, costly, and, most importantly, risky because current methods to evaluate measurement sensitivity may not assure a researcher that valuable resources placed into developing a measure will render it sensitive to the intervention under investigation.

Given the range of issues described above, a new method to evaluate measurement sensitivity of items would have numerous benefits at the national level. First, evaluating the sensitivity of specific items would help identify those that were truly associated with the intervention so that instruments could be developed from only those items with demonstrated sensitivity. Second, the ability to select for inclusion during the instrument development process items from existing measures that should detect effects of an intervention would reduce the cost (both in money and time) of developing new intervention-specific measures. Third, the identification of sensitive items could provide detailed descriptions of the sub-domains where the intervention was effective. Fourth, identification of multiple sensitive items (i.e., bundles) would offer the possibility of informing the future development of instruments that would have higher score reliability than attainable with a single item. Fifth, a cost effective method coupled with cost effective approaches would provide researchers with a heretofore unavailable process to evaluate sensitivity. Sixth, the evaluation and subsequent identification of sensitive items and bundles would ultimately lead to instruments with increased sensitivity that could be used in subsequent efficacy research.

1.6 Estimating sensitivity using differential item functioning

There is no universally agreed upon technique for estimating the responsiveness or sensitivity of an instrument to a particular intervention (Liang et al., 2002; Polikoff, 2010). However, all current practices assess responsiveness in some fashion at the total score level. Most instruments are composed of multiple items to maximise internal
consistency reliability. Some of these items may be sensitive to the treatment while others may not be due to variations in difficulty or content. In such cases, an analytic technique that examines responsiveness at the item level can help develop instruments that are more sensitive to educational interventions and instruction.

Differential item functioning (DIF) – the difference in the probabilities of a correct response (for achievement measures) or endorsement of an item (for psychological measures) for examinees matched on the measured trait – is an analytic technique that has the potential to detect intervention effects at the item level. Although traditionally the presence of DIF in an instrument is viewed as a threat to score validity and is therefore undesirable, in the context of measurement sensitivity assessment, DIF would indicate a difference in the way experimental and control group individuals, matched on an overall proficiency in the area of interest, respond to an item that has been hypothesised to directly measure the treatment effect. In other words, in this context DIF would not represent problematic item bias against a group of examinees, but rather hypothesised treatment effects (e.g., Millsap, 2011). That said, one would expect a mean difference across groups but it certainly is possible that measures are not able to detect intervention effects that may target certain skills that are aligned to specific items.

Strength of the measurement sensitivity approach outlined here is that a priori hypotheses are specified for specific items that should exhibit DIF due to content alignment and expectations regarding intervention effects. Indeed, it needs to be clear that we are not suggesting that the method outlined below be used post-facto to identify and use only those items that are sensitive to an intervention’s effect (i.e., demonstrate DIF between a treatment and control group) after conducting an efficacy or effectiveness trial. Such an approach could lead to erroneous conclusions regarding an intervention’s effectiveness because apparent group differences may not be tied to hypotheses grounded in the literature, but rather be the result of type I error, and/or random sampling variation. The post-facto approach must, therefore, be shunned. Instead, we demonstrate the following method only to evaluate sensitivity during the instrument development and item selection stage of a project. Our goal is to develop and provide a cost effective and efficient method for evaluating sensitivity such that researchers and test developers can easily implement it during the instrument development phase of their research, in conjunction with the development of a priori hypotheses regarding expected effects.

2 Method

The current study consisted of two distinct steps. Step 1 involved engaging experts with a teacher-focused professional development intervention to conduct a content analysis with a goal of aligning specific items on a general assessment with the intervention, based on expectations about what impact the intervention should have on learning. Step 2 employed DIF analysis to test the sensitivity of items identified in Step 1 for detecting effects related to the intervention. In other words, in Step 2 the items that were initially identified as being potentially sensitive to the intervention were then assessed for the presence of DIF with respect to the group of students receiving the intervention versus those who did not. As a check, we also assessed other items as well to see if non-aligned items showed no DIF. With content analysis providing a priori hypotheses regarding each item’s sensitivity to the intervention, the use of DIF detection in this study is in accord
with the effort to test explicit DIF hypotheses in order to identify causes of DIF (see Roussos and Stout, 2004) and new uses of such analyses for answering other group difference questions (Millsap, 2011). The two-step process is also in accord with suggestions to use judgemental and empirical strategies to evaluate measurement and instructional sensitivity (Popham, 2007a; Popham and Ryan, 2012) as well as treatment effects (Millsap, 2011).

2.1 Data source

Data from a current randomised cluster field trial that implements the science writing heuristic (SWH) approach to learning science was used. SWH is an immersive approach to teaching on the scientific argument and is examined in a randomised control trial study in the Midwest with comparison and treatment schools. SWH takes a language-based approach to science, requiring students to pose questions, gather data, make claims, generate evidence, compare their answers to scientists’ answers, and reflect on the activity as a whole. A critical element of this process requires students to continually defend and debate their ideas, both publicly and privately. Students must both construct their own ideas and critique their own and others’ ideas as a necessary component of learning. The schools not assigned to the SWH condition continued to operate under a business-as-usual model.

Participants were a representative sample of students taken from 48 schools in a Midwestern USA location in grades 3 through 5. We had access to 2,181 treatment and 1,004 control students who completed the Cornell critical thinking test. A co-author involved in the trial and graduate students assisted with item-level content review and hypothesis development and post hoc review of results.

2.2 Instrument

The Cornell critical thinking tests (CCTT; Ennis et al., 2005) assess general critical thinking ability across four factors including induction, deduction, observation/credibility and assumptions. This study examines Form X, for grades 4 through the sophomore year of post-secondary education. The CCTT-Form X requires approximately 50 minutes for 71 items. There are three response options per item in a multiple-choice format, and the items are scored dichotomously (i.e., correct or incorrect). These questions mostly involve higher-order thinking beyond factual recall. The technical manual provides score reliability and validity information. Internal consistency reliability estimates via Cronbach’s alpha range from 0.67 to 0.90 for Form X. Internal consistency for the scores based on data employed in this study was 0.83.

Correlations between the CCTT and other critical thinking tests range from 0.60 with the critical reading in social studies assessment to 0.50 and 0.41 with the Logical Reasoning Test and the Watson-Glaser assessment, respectively (Ennis et al., 2005). The correlations between the CCTT and other constructed tests range from 0.74 with the Otis-Lemon assessment, 0.53 with the Houghton-Mifflin cognitive abilities verbal assessment, and 0.52 with the SAT total score. These values together support validity of the inferences from the measure. CCTT items appear to function similarly across girls and boys as well (French et al., 2012).
3 Analysis

3.1 Alignment

The alignment process involved a review of items by content experts who were trained to match items to specific skills to be targeted by the intervention. The process for identifying the relevant items was similar to the manner in which content experts are used in item bias reviews (Tittle, 1982). The review process sought to identify items on the CCTT that were theoretically aligned to the intervention content and therefore likely candidates for detecting the intervention’s effects. The content experts (i.e., developer of the intervention and two trained graduate assistants) matched items with the targeted intervention skills, rating items on a scale of 1 to 5 with 5 representing an exact match of the item to the targeted skill and 1 representing no match. The items with a rating of 4 or higher were those we hypothesised would be potentially sensitive to the treatment effects and thus show significant DIF. This process is in accord with recommendations for review procedures for the judgemental process (e.g., Popham, 2007a; Popham and Ryan, 2012). Judges rated items individually and then compared ratings and resolved any discrepancies in ratings in a second round of ratings.

3.2 Multilevel Mantel-Haenszel for estimating sensitivity

The data were collected in a multilevel framework (students nested in classrooms). When the clustering inherent in multilevel data is ignored, standard errors and hypothesis tests associated with most statistical methods, including those for DIF detection, will be incorrectly estimated (Raudenbush, 1997; Raudenbush and Bryk, 2002; French and Finch, 2010). Therefore, methods appropriate for such multilevel data must be employed to ensure that the results are statistically valid. To this end, we used a multilevel version of the Mantel-Haenszel (MH) procedure.

Holland and Thayer (1988) applied the MH procedure, first used in medical research to match patients for comparison (Mantel and Haenszel, 1959), to DIF detection. This procedure is an extension of chi-square methods, allowing for comparison of item responses between the focal and reference groups’ conditioning on a matching subtest score. The MH procedure matches individuals in group 1 and group 2 on an observed score estimate of ability, and then compares the probabilities of a correct response on the target item between the two groups, conditioned on this matching score. A set of \( k \) such comparisons are made, where \( k \) is the number of different scores on the matching score. The MH \( \chi^2 \) value is calculated as:

\[
\chi^2_{MH} = \left[ \frac{\sum A_k - \sum E(A_k)}{\sum \text{Var}(A_k)} \right]^{.5} \]

where

\[
A_k = \text{number in reference group answering target item correctly at score level } k
\]

\[
E(A_k) = \frac{N_{Rk}N_{k}}{N_k}
\]
Measurement invariance techniques to enhance measurement sensitivity

\[ \text{Var}(A_k) = \frac{N_{Rk}N_{Fk}N_{1k}N_{0k}}{(N_k)^2(N_k - 1)} \]

- \(N_{Rk}\) number of reference group individuals at score level \(k\)
- \(N_{Fk}\) number of focal group individuals at score level \(k\)
- \(N_{1k}\) number of correct responses at score level \(k\)
- \(N_{0k}\) number of incorrect responses at score level \(k\)
- \(N_k\) number of individuals at score level \(k\).

A statistically significant result indicates the presence of uniform DIF for the target item. This method is effective at detecting uniform DIF in a variety of conditions (e.g., Narayanan and Swaminathan, 1994) The MMH method is based on MH, and includes an adjustment that estimates the variance in the MH statistic due to clustering of examinees and the naïve variance assuming no such clustering (Begg, 1999; BMH). This BMH approach, adjusts the MH test statistic as in equation (1):

\[ f = \frac{\sigma_{GEE}^2}{\sigma_{\text{Naive}}^2} \]  

where

\(\sigma_{GEE}^2\) GEE adjusted variance of the score statistic accounting for clustering

\(\sigma_{\text{Naive}}^2\) naïve variance of the score statistic ignoring clustering; proportional to the variance of MH

The adjusted MH statistic takes the form:

\[ MH_B = \frac{MH}{f} \]  

As an example, when there is no correlation in scores among examinees from the same school, \(f = 1\) and \(MH_B = MH\). This method has been shown to work with detecting DIF (French and Finch, 2013). We employed the delta statistic as an effect size measure and ETS classification (e.g., A, B, C) to judge the magnitude of DIF and supplement the chi-square significance test.

We report the percentage of total items analysed that were identified by

- a both content and DIF analysis
- b only content analysis
- c only DIF analysis.

These results provide a sense of how often one can expect these methods to agree. Results will assist to determine the degree to which the DIF analyses are capable of detecting treatment effects in items less aligned to the intervention.
4 Results

4.1 Content analysis

Several items within the four domains were rated as matching to the intervention. However, no items were rated as a perfect match (e.g., 5). In the domain of induction, 100% of the items received a score of 4 across the raters. Likewise, in the domain of deduction, 100% of the items also were scored a 4 across raters. In the domain of observation and credibility, 0% of items were rated as aligned with the intervention, that is, all items were rated a 0 across raters. Finally, in the domain of assumptions, 100% of the items were given a score of 4 across raters. From the content ratings, it would appear treatment and comparison groups should differ on scores on the induction, deduction, and assumptions scores.

Initially, we approached the problem of assessing intervention effects in the traditional manner by comparing mean total scores on the assessments described above between the intervention and control groups. In order to test for any group differences at the total score level for each domain, independent t-tests were used. Results of the t-tests revealed that there were no statistically significant differences between the groups on the Induction domain, $t(2,263) = 0.42, p = 0.67, d = 0.02$; Deduction domain, $t(2,263) = 0.59, p = 0.55, d = 0.02$; or the Observation domain, $t(2,263) = 0.52, p = 0.60, d = 0.02$. However, there was a significant difference on the scores between groups in the Assumptions domain, $t(2,263) = 4.89, p =< 0.01, d = 0.21$.

4.2 DIF analysis

In contrast to the standard approach using the $t$-test, the alternative measurement sensitivity method based on content alignment and DIF analysis was also used. In the domain of Induction, there were four items that were identified statistically as DIF items. The associated effect sizes ranged from 1.80 to 3.34 indicating C DIF items (i.e., large magnitude of DIF) based on the ETS classification. This outcome reflects 16% of the items rated as aligned with the intervention being identified as DIF items. We also note that although not significant, the remaining items had delta values ranging from 1.8 to 2.9, and favoured the treatment group.

In the deduction domain, there were three items (20%) that were identified statistically as DIF items out of the items that were rated as aligned with the intervention. The associated effect size of these items ranged from 1.8 to 3.2, again reflecting C DIF items on the ETS classification scale and thus reflecting a large magnitude of difference. Similarly, the non-significant DIF items also had delta values above 1.8, and favoured the treatment group.

In the observation/credibility domain, no items were identified as DIF items. This outcome was in accord with the content analysis where no items were rated as being aligned with the intervention.

In the assumptions domain, there were two items (20%) that were identified statistically as DIF items out of the items that were rated as aligned with the intervention. The associated effect size of these items ranged from 2.5 to 2.6, again reflecting C DIF items on the ETS classification scale and thus a large magnitude of difference. Similarly, the non-significant DIF items had delta values above 1.9, and favoured the treatment group.
5 Discussion

This study proposes an innovative solution to a common challenge in educational effectiveness research. A new paradigm is illustrated through the implementation of an instructionally sensitive invariance framework. The lack of mean differences scores on three of the domains is in accord with prior work in this area, showing that there may be effects of the intervention that are not visible or detectable at the overall mean level. Furthermore, our study results suggest that although mean differences were not observed across all domains between treatment and comparison groups, there were intervention effects associated with specific items on the assessments from among those that were rated as being aligned with the intervention. In each of three domains that were identified a priori as being targeted by the intervention there was a rate of DIF that ranged from 16% to 20%. In fact, the differences were large in magnitude, based on the commonly used DIF effect size measure employed here.

What is unknown in this work is the extent to which DIF hit rates in terms of percentages would be expected and at what magnitude. That is, raters expected that 100% of items were identified as being aligned with the intervention and thus the teacher instructional techniques. The statistical results suggest that only about a fifth of the items show performance differences in terms of difficulty levels across the treatment and comparison groups. This does not mean that all items are not aligned but rather only a fifth of the items showed large enough differences (e.g., easier for the intervention students) to be detected statistically. In addition, we did not control for other factors that may moderate the differences noted. That is, one could consider a different DIF model (e.g., logistic regression) to include variables such as treatment fidelity to see if that influences results. In that analysis, the classrooms with teachers with higher fidelity may show greater amounts of DIF. Also, such models would allow for other types of DIF to be explored (e.g., non-uniform DIF).

This research is motivated by evaluation in the social, behavioural, and clinical sciences where the lack of sensitive outcome measures renders limited assessment of intervention effects. Increasing instrument sensitivity is relevant to the research community because it improves accuracy of conclusions regarding effectiveness of interventions or instruction. Detecting sensitive items also is relevant because it can provide detailed evidence regarding the specific sub-domains and skills where the intervention is most effective in producing change. Developing measures that are sensitive to program effects (e.g., instruction, intervention) at the sub-domain level may reveal benefits that go undetected when only total scores are used. Policy, funding, and clinical treatment decisions all rely on accurate program evaluation. New methods to evaluate instrument responsiveness can help researchers provide methods to evaluate item sensitivity in order to develop more sensitive instruments. More sensitive instruments have a many potential benefits, ranging from detecting small effects, lower costs for conducting effectiveness research, and even improved indications of teacher effectiveness that can be used in the growing need to evaluate teachers against student achievement gains. We hope this work is a start in a sustainable line of work with such techniques that allows the field to move forward in addressing past concerns about sensitivity in measurement.
Author note

Part of this research was supported by the Institute of Education Sciences, US Department of Education, through Grant R305D110014. The opinions expressed are those of the authors and do not represent views of the institute or the US Department of Education.

References


Measuremen
tinvariance techniques to enhance measurement sensitivity


