Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 23 Number 4, April 2018 ISSN 1531-7714

Exploring Differences in Measurement and Reporting of Classroom Observation Inter-Rater Reliability

Anne Garrison Wilhelm, *Southern Methodist University*
Amy Gillespie Rouse, *Southern Methodist University*
Francesca Jones, *Southern Methodist University*

Although inter-rater reliability is an important aspect of using observational instruments, it has received little theoretical attention. In this article, we offer some guidance for practitioners and consumers of classroom observations so that they can make decisions about inter-rater reliability, both for study design and in the reporting of data and results. We reviewed articles in two major journals in the fields of reading and mathematics to understand how researchers have measured and reported inter-rater reliability in a recent decade. We found that researchers have tended to report measures of inter-rater agreement above the .80 threshold with little attention to the magnitude of score differences between raters. Then, we conducted simulations to understand both how different indices for classroom observation reliability are related to each other and the impact of reliability decisions on study results. Results from the simulation studies suggest that mean correlations with an outcome are slightly lower at lower levels of percentage of exact agreement but that the magnitude of score differences has a more dramatic effect on correlations. Therefore, adhering to strict thresholds for inter-rater agreement is less helpful than reporting exact point estimates and also examining measures of rater consistency.

With an increased focus on evidence-based instruction, assessment, and teacher accountability, it is critical that educators, administrators and researchers have valid and reliable ways of recording what is occurring in K-12 classrooms (Hill, Charalambous, & Kraft, 2012; MET project, 2013). Systematic classroom observation is one such method for identifying and quantifying teacher and student behaviors in the classroom (Kelcey & Carlisle, 2013; Pianta & Hamre, 2009; Vaughn & Briggs, 2003). Using an observational instrument, researchers or trained raters systematically record and categorize the occurrence of teacher and student behaviors of interest. Such tools allow researchers and evaluators to observe education in action as well as to document the frequency and type of behaviors that occur. In this way, observational instruments provide a direct means of examining the content and complexity of teacher instruction and student learning (Kelcey & Carlisle, 2013; Kennedy, 1999; Vaughn & Briggs, 2003).

Using systematic classroom observation has become especially important for teacher evaluations, with classroom observations being the most widely adopted teacher evaluation method (Cash, Hamre, Pianta, & Myers, 2012; Cohen & Goldhaber, 2016; MET project, 2013; Strong, 2001; Van der Lans, van de Grift, van Veen, & Marjon, 2016). Beyond evaluating teacher quality, classroom observations are often used in educational research to understand teacher and student behavior, determine the impact of interventions, and examine the fidelity of interventions. In fact, the primary source of recording activities and interactions in the classroom is observation research (Swanson, Solis, Ciullo, & McKenna, 2012). Data from quantitative observational instruments can inform research on

teaching effectiveness, too, but researchers must ensure that the tools they use are valid and reliable. Further, they must look beyond the observational instrument to also ensure they are employing well-trained raters and robust scoring designs to produce reliable teacher scores (Cash et al., 2012; Hill et al., 2012; Kelsey & Carlisle, 2013).

One important dimension of the technical adequacy of observational measurements is inter-rater reliability. Inter-rater reliability is a critical piece of ensuring that classroom observations are accurate and meaningful (Ho & Kane, 2013; Semmelroth & Johnson, 2014). Without demonstrating that two independent judges can be reliably trained to similarly rate a particular behavior, the possibility of achieving objective measurement of educational phenomena is diminished (Krippendorff, 2016). Unfortunately, the concept of inter-rater reliability has received far less theoretical attention than it warrants (Stemler, 2004).

In this article, we address this inattention and offer empirically-based guidance about the concept of inter-rater reliability. We examined articles in two major journals in the fields of reading and mathematics to understand current practice in how researchers measure and report inter-rater reliability. Then, we completed a statistical simulation in which we examined scoring differences and their effects on different reliability indices, and whether different levels of reliability affect the relations to simulated outcomes (e.g., student achievement averages). In what follows, we review the literature pertaining to assessing validity and reliability of classroom observations.

Observational Systems

Given the increasing use of observational instruments in both research and in teacher evaluation, information on best practices for the use of these instruments is critical (Cash et al., 2012; Hill et al., 2012). Observational systems vary widely in demonstrated validity and in the level of training provided to people interested in using them, ranging from researcher-developed strategies to commercial observational systems complete with manuals and trainings (Cash et al., 2012). There are numerous methodological approaches to classroom observation and virtually no standard practices in the field (Kennedy, 1999). In addition, there is very little research on how best to train raters to use the observational instruments consistently (Cash et al., 2012). As we expect to see an increase in the number and use of observational instruments, researchers need to more carefully examine the sources of variation in observational scores and to consider implications for how these ratings are used (Hill et al., 2012; Semmelroth & Johnson, 2014).

Sources of Variation in Observations

Recent generalizability studies of popular instruments (e.g., Framework for Teaching [FFT], Mathematical Quality of Instruction [MQI], Recognizing Effective Special Education Teachers [RESET]) have examined several potential sources of variation in classroom observations (Hill et al., 2012; Ho & Kane, 2013; Kane & Steiger, 2012). Using *Generalizability Theory (G-Theory)* as a statistical method for evaluating the dependability (or reliability) of behavioral measurements (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), the studies provide a comprehensive framework for sampling observations (Hill et al., 2012). In particular, they provide information about the optimal number of raters and the number of lessons required to produce desired reliabilities (Hill et al., 2012). In general, findings from these studies are that multiple observations and multiple highly-trained raters are critical for achieving high levels of measurement score reliability (Hill et al., 2012; Ho & Kane, 2013, Kane & Steiger, 2012). Unfortunately, for many instruments, thorough validation studies and generalizability studies have not been carried out (Hill et al., 2012).

Even when a generalizability study has been conducted to recommend the number of raters, the number of observations, and the level of training required of raters, the use of a validated observational system does not ensure that the data produced will be reliable (van der Lans, et al., 2016). The critical final piece is ensuring that the rating process has not produced irrelevant variation. Demonstrating agreement between replications by different raters "allows us to infer the extent to which data can be considered as reliable surrogates for phenomena of analytical interest," (Krippendorff, 2016, p.139). While there is agreement about the need for reliability (AERA, APA, and NCME, 2014), there is little empirically verified guidance with respect to collecting data about and reporting reliability of observational ratings (Swanson et al., 2012).

Calculating Inter-Rater Reliability

There are a number of approaches used in the field to assess inter-rater reliability. In fact, people often use the terms inter-rater agreement and inter-rater reliability

(IRR) interchangeably. However, these two terms are not interchangeable and represent different notions of reliability (e.g., Kottner et al., 2011; Liao, Hunt, Chen, 2010; Stolarova, Wolf, Rinker, Brielmann, 2014). Inter-rater agreement is a dimension of reliability and assesses the degree of agreement or consensus between raters (Hintze & Matthews, 2004; Kazdin, 1982; Stemler, 2004); it provides no information about the alignment of those ratings with the phenomena of interest (e.g., the occurrence of the behavior) (Hintze & Matthews, 2004). IRR, on the other hand, is more general and attends to the accuracy of the rating process. For example, the agreement between two raters contributes to the overall accuracy of the rating process, but IRR also includes the degree to which the rating process consistently differentiates objects of measurement (e.g., teachers or classrooms). In what follows, we make the distinction between indices of consensus (i.e., agreement) and indices of consistency as they both contribute to our understanding of IRR.

Starting with an empirically validated observational system makes it more likely that the observers' ratings will be closer to representing the phenomena of interest (Hill et al., 2012). However, it is still very important to attend to the reliability of the measure as it is implemented. Further, many studies use incorrect statistical indices to compute consensus or consistency, misinterpret the results from reliability analyses, or fail to consider the implications that reliability estimates have on subsequent analyses (Hallgren, 2012; Krippendorff, 2016).

Statistics and interpretation guidelines. Historically, different statistics have been used to estimate IRR (Stemler, 2004). To trace the history of these statistics (i.e., percent agreement, Cohen's kappa, intra-class correlations, Cronbach's alpha, and correlation coefficients) and locate seminal papers that provide guidelines for their interpretation, we had to consult fields outside of education, including content analysis, medicine and psychology.

Traditionally, researchers have approached estimates of consensus using percentage of exact agreement (Hintze & Matthews, 2004). This method involves dividing the number of exact agreements in observations by the total number of observations. Hartmann, Barrios and Wood (2004) reported guidelines for interpreting percentages of exact agreement, suggesting that exact agreement between raters of 80 to 90 percent is sufficient but that for more complex instruments, exact agreement between raters of 70 percent may suffice.

In 1960, Cohen questioned the use of percent agreement, noting that this index does not account for chance agreements between raters. He proposed the use of Cohen's kappa (κ), another measure of consensus, which accounts for chance agreement, and provides a standardized value for consensus that can be interpreted across studies. Landis and Koch (1977) provided guidelines for interpreting κ, suggesting that κ values of 0.41 to 0.60 were moderate, 0.61 to 0.80 were substantial, and 0.81 to 1.00 indicated almost perfect agreement. This field has continued to expand with a number of methodologists proposing new indices that similarly attempt to adjust for chance or expected agreement including Krippendorff's α (2013), π (Scott, 1955; Fleiss, 1971), and AC1 (Gwet, 2002). All of these indices of consensus attempt to improve on Cohen's κ but none of them have taken hold in the education research community.

Other reliability indices have been more focused on consistency rather than consensus (Stemler, 2014). One seminal paper in this tradition was produced by Shrout and Fleiss (1979) and described the use of intra-class correlations (ICCs) to measure IRR. Whereas percent agreement, κ, and the other new indices focus on agreement between raters, ICCs attempt to index the extent to which the instrument is able to consistently differentiate between participants with different scores (Kottner et al., 2011; Liao, Hunt, Chen, 2010; Stolarova, Wolf, Rinker, Brielmann, 2014). Cichetti (1994) provided guidelines for interpreting ICC values for IRR, stating that ICCs less than 0.40 were poor, between 0.40 and 0.59 were fair, between 0.60 and 0.74 were good, and between 0.75 and 1.00 were excellent.

Several other approaches that are more focused on consistency than consensus are Cronbach's alpha (α) or correlation coefficients (Liao et al., 2010; Stolarova et al., 2004). A measure of internal consistency, Cronbach's α is sometimes considered reasonable because "items" that are internally consistent are comparable to raters agreeing about the "true" value of the construct (Gwet, 2012). Common interpretation guidelines for α come from the assessment field, as this statistic is typically used for calculating the internal consistency of test items. Bland and Altman (1997) noted that α values of .70 to .80 are satisfactory but that for clinical application (i.e.,

in the medical field), α values should be much higher (0.90 to 0.95). Additionally, Cohen (1988) suggested guidelines for interpreting correlation coefficients as effect sizes, with 0.10 as small, 0.30 as medium, and 0.50 and above as large. Yet, while focusing on the consistency dimension of reliability, alpha and correlation coefficients provide little information about the exact agreement between raters. For example, the Spearman rank-order correlation will be 1 if two raters place classroom observations into the same rank order, even if their scores consistently differ by 1 score category.

In sum, to measure the consensus and consistency of raters, researchers have historically conflated agreement and reliability and utilized a number of different indices. Reasons for using different indices include alignment with the research question, familiarity with or accessibility of different computational procedures, and logistics of instrument use (e.g., Boston, Bostic, Lesseig, & Sherman, 2015; Stuhlman, Hamre, Downer, & Pianta, 2014). Yet, despite the different foci on consensus or consistency via the different indices, there seems to be a standard practice of taking a reliability (or agreement) level of .7 or .8 as sufficient for research (Lance, Butts, & Michels, 2006; McHugh, 2012). After asking numerous methodologists, performing several literature searches, and hand searching popular statistics textbooks with chapters on inter-rater agreement and IRR (e.g., Fleiss, 1981; von Eye & Mun, 2005), we found no consensus about the history of the .7 or .8 cutoff criterion. We reason, as Lance and colleagues (2006) did, that this cutoff may have become part of research history by error or misinterpretation, by overgeneralizing cutoffs from classical test theory for internal consistency reliability within indices of IRR.

Decisions Points in Using Observational Instruments

In the case where an observation system is not fully specified, there are several decision points, beyond indices for calculating IRR, that can affect reliability of observational instrument implementation. First, there are multiple ways to train raters and determine sufficient reliability between them. One way to determine reliability is to designate an expert and then train your raters to meet the "gold standard" set by that expert's "correct" coding of events (Gwet, 2012). Often, training continues until a rater passes the calibration assessment with a pre-determined criterion, such as 60%-80% agreement with expert scores. Once raters are trained, there are a number of other decisions that arise in the use of observational instruments. One such decision is whether and how to calculate reliability of raters during data collection. Any rating process that extends over time necessitates the calculation of ongoing data collection reliability because of the potential for rater drift (Kazdin, 1997). Further, decisions about the appropriate level of ongoing (i.e., data collection) reliability arise as well as decisions about how to create scores for each observation. For example, when two raters observe and rate a classroom, they can come to consensus to determine what is "correct" or their scores can be averaged.

With so many decision points there are multiple opportunities for error (Hintze & Matthews, 2004). Therefore, it is critical that we determine the extent to which these decisions affect IRR and study outcomes. This paper aims to offer empirically based guidance about the concept of IRR for classroom observations. In particular, we address the following research questions.

1. How do researchers report classroom observation IRR?
2. How are different indices of classroom observation IRR related to each other?
3. What is the impact of differences in classroom observation scoring and IRR?

We first describe the method and results for research question 1 and then describe the method and results for research questions 2 and 3.

Method: Research Question 1

To answer research question 1, we systematically reviewed articles published from 2007-2016 in the *Journal for Research in Mathematics Education (JRME)* and *Reading Research Quarterly (RRQ)*. Our starting point was driven by a research commentary in *JRME* (Hill & Shih, 2009); in it, the authors examined the quality of articles published in the journal from 1997 to 2006 and made several recommendations for future work, including asking researchers to report reliability and validity of the measures they used. We then proceeded to examine articles published in the decade after this commentary (i.e., 2007 to 2016), under the assumption that editors of *JRME* and authors would respond to Hill and Shih's (2009) recommendations. Subsequently, we chose to

examine *RRQ* for the same date range to compare the trends of reporting classroom observation reliability in mathematics education research to trends in another prominent area of education research (i.e., reading). We chose *RRQ* as our journal of interest by consulting experts in the field of reading research to determine what they considered the top journal in their field. We examined 2015 Journal Citation Reports® (Thomson Reuters, 2016) to confirm the experts' recommendation. *RRQ* had an impact factor of 2 and was ranked in the top 25 journals in education and education research (*JRME* was also ranked in the top 25 and had an impact factor of about 2).

In addition to being published in the 10 year span we chose in *JRME* or *RRQ*, articles had to meet the following criteria to be included in our review: (a) included a classroom observational instrument used to measure student and/or teachers' behaviors, (b) used quantitative data analysis, and (c) examined classroom observation data as an independent variable using statistical data analysis (studies using classroom observation data for treatment fidelity purposes were only included if treatment fidelity data were used to predict study outcomes in statistical analyses). Our inclusion criteria were consistent with those employed by Hill and Shih (2009) (i.e., including only quantitative studies that used statistical data analysis) and with our own research approaches (i.e., two of the authors routinely conduct classroom observations in their research and all authors have expertise in quantitative methods).

The three authors each independently examined approximately one-third of the 272 articles published in *JRME* from 2007 to 2016 and approximately one-third of the 230 articles published in *RRQ* during the same data range to determine if they met inclusion criteria for the review. For inclusion reliability purposes, the authors rotated, each reexamining another author's recommendations for studies that met inclusion criteria. All differences regarding studies' acceptability for inclusion in the review were resolved by discussion and consensus. In the end, 17 studies met our inclusion criteria, with 8 from *JRME* and 9 from *RRQ*.

The authors each independently coded approximately one-third of the qualifying studies for: (a) year of publication; (b) number of teachers in the study; (c) grade level of students in the study; (d) observational instrument(s) used; (e) number of observations; (e) type of observation (in-person, video, transcripts of teacher and student talk); (f) training reliability on the observational instrument (type of reliability reported and statistic) and (g) reliability during data collection (type of reliability reported and statistic). All studies were double coded by a second author, and authors resolved any disagreements by discussion and consensus.

Results: Research Question 1

Study Characteristics

In Table 1, we present all coded features for each of the 17 studies included in the review.

Reporting of Inter-Rater Reliability

To answer research question 1, we looked at both the training reliability reported when observers learned to use classroom observational instruments as well as the reliability reported when observers used classroom observational instruments for study data collection. Across the two journals, we found a range, both in terms of *whether* reliability was reported in studies and *how* reliability was reported in studies. Additionally, it is important to note, as previously mentioned, that inter-rater agreement and IRR are routinely conflated in research publications. In this paper, and across the majority of the 17 studies, the term "reliability" (or IRR) is used to represent the general category of information. As we describe results from the studies we reviewed, we focus on the indices that were used and the point estimates (or thresholds reported) rather than the general terms used to describe them.

Out of 8 studies in *JRME*, only 3 reported training reliability. Of these three studies, two reported agreement of 80% (Jackson et al., 2013; Wilhelm, 2014) and one reported an ICC of 0.80 (Copur-Gencturk, 2015). In *RRQ*, a majority of studies reported training reliability information (n = 7). Four of the seven RRQ studies reported Cohen's kappa values for IRR on training observations; two studies reported exact kappa values (Connor et al., 2011, kappa = 0.73; Silverman & Crandell, 2010, kappa = 0.82), while the other two studies reported kappa greater than or equal to 0.80 (Sailors et al., 2014; Silverman et al., 2014). Three other *RRQ* studies included percentages of exact agreement on training observations ranging from 88% to greater than 90% (Kelcey & Carlisle, 2013; Rodgers et al., 2016; Vaughn et al., 2013).

Table 1. Individual Study Descriptions

Study	Teachers and Grade Level	Observation instrument	*N* observations	Type of observation	Training	Reliability Data Collection
			JRME			
Boston & Smith (2009)	11, middle and high school	IQA	3 per teacher	L	NR	EA lesson observations (tasks and implementation) = 100%
Brown et al. (2009)	14, grades 1 and 2	RD	33 total	V	NR	NR
Clements et al. (2011)	106, pre-k	COEMET, RD	2 per teacher	L	NR	EA = 0.80
Copur-Gencturk (2015)	21, grades 1-7	combined LSC & OMLI	2-3 per teacher	V for training IP for data collection	ICC = 0.80	NR
Grouws et al. (2013)	33, high school	CVP, RD	3 per class (43 classes)	L	NR	EA = 94%
Jackson et al. (2013)	165 middle school	Expanded IQA	1-2 per teacher	V	EA = 80%	EA = 70.5% (kappa = 0.48)
Tarr et al. (2013)	64, grades 9-12	CVP, RD	3 per teacher	L	NR	EA = 94%
Wilhelm (2014)	213, middle school	IQA	2 per teacher	V	EA = 80%	EA task potential = 62.6% (kappa = 0.41) task implementation = 77.4% (kappa = 0.48)
			RRQ			
Connor et al. (2011)	33, grade 3	RD	3 per teacher	V	kappa = 0.73	kappa = 0.73
Guthrie et al. (2013)	20, grade 7	RD	2 per teacher	L	NR	NR
Kelcey & Carlisle (2013)	87, grades 2 and 3	ACOS-R, RD	4 per teacher	V for training L for data collection	EA = 88%	EA = 87%
Rodgers et al. (2016)	10, grade 1	RD	2 per teacher	T	EA > 90%	EA = 85% (when new codes were established)
Sailors et al. (2014)	162, grades 1-3	RD	2 per teacher	L	kappa ≥ 0.80	NR
Silverman & Crandell (2010)	16, pre-k and k	RD	3 per teacher	L	kappa = 0.82	kappa = 0.97
Silverman et al. (2014)	33, grades 3-5	RD	3 per teacher	L and AR	kappa > 0.80	kappa > 0.80
Vaughn et al. (2013)	5, grade 8	RD	2 per teacher	V for training L for data collection	EA > 90%	NR
White et al. (2014)	81, grade 3	RD	2 per teacher	V	NR	EA > 80%

IQA = Instructional Quality Assessment; L = live; EA = exact agreement; NR = not reported; RD = researcher-developed instrument; V = video; COEMET = Classroom Observation of Early Mathematics Environment and Teaching; LSC = Local Systemic Change; OMLI = Oregon Mathematics Leadership Institute; CVP = Classroom Visits Protocol; T = transcripts of student and teacher talk; AR = audio recording

For data collection, a majority of studies in JRME reported reliability information (n = 6), with four studies reporting percentages of exact agreement ranging from 80 to 100% (Boston & Smith, 2009; Clements et al., 2011; Grouws et al., 2013; Tarr et al., 2013) and two studies reporting both percentages of exact agreement (range = 62.6 to 77.4%) and kappa values (range = 0.41 to 0.48) (Jackson et al., 2013; Wilhelm, 2014). In RRQ, a majority of studies also reported reliability information for data collection (n = 6). Three studies reported percentages of exact agreement; two studies reported exact percentages (Rodgers et al., 2016, EA = 85%; Kelcey & Carlisle, 2013, EA = 87%) and one reported agreement greater than 80% (White et al., 2014). The three remaining *RRQ* studies reported kappa values for reliability during data collection, with two of these studies reporting exact kappa values (Connor et al., 2011, kappa = 0.73; Silverman & Crandell, 2010, kappa = 0.97) and one reporting kappa greater than 0.80 (Silverman et al., 2014).

Method: Research Questions 2 and 3

We set out to answer the second and third research questions by using Monte Carlo simulation (Robert & Casella, 2004). We wanted to understand both how different indices for classroom observation reliability are related to each other and the impact of scoring decisions on study results. For both purposes, we used simulated data designed around a hypothetical classroom observational instrument. We designed the instrument to look similar to frequently used classroom observational instruments (e.g., CLASS, IQA) but not to be exactly like any one of them. Our simulated instrument consisted of ten rubrics with scores ranging from 0 to 4. For ease of interpretation, we assume that 0 represents "Low", 2 represents "Medium", and 4 represents "High" incidence of the behavior of interest. We assumed that the scores on individual rubrics were normally distributed with a mean of 2.5. We simulated data for a sample of 100 observations (representing 100 different classrooms) by randomly generating normally distributed ordinal scores for each of the ten rubrics across the 100 classrooms. For both research questions, we repeated the simulations 100 times to examine the trends across the simulations.

To address research question 2, we generated data with specific percentages of exact agreement and then calculated several other reliability indices as well, to understand how the different indices are related to each other at the different levels of exact agreement (which was the simplest to model). In particular, we generated data that exactly agreed with the original, and then modified the data for 40, 30, 20, and 10 percent of the observations, respectively. For example, for the 60% exact agreement case, we randomly selected 60 of the 100 observations to agree exactly with the original scores, and then created new scores for the other 40 observations, based on two different score characteristics: we had to decide how far off the scores would be from the original score so we modeled it two different ways, with scores off by 1 (in either direction), or off by 2 (again, in either direction). For example, if the original score (i.e., rater 1) was a 2 ("Medium"), then the rater 2 score off by 1 was randomly assigned to either a 1 ("Medium Low") or a 3 ("Medium High"), and the rater 2 score off by 2 was randomly assigned to either a 0 ("Low") or a 4 ("High"). We made this distinction because we expected that the magnitude of the disagreement might have an impact on some of the reliability indices or on the variation in the relationships with other variables. Further, given the expected mean between 2 and 3, and dramatic qualitative differences between categories, we would expect few score differences greater than 2 for this hypothetical instrument. Therefore, scores that differ by 2 represent a realistic, yet worst-case, scenario of inter-rater agreement. We simulated the data for 70, 80, and 90 percent exact agreement in the same way. Once this data was simulated, we then calculated several additional reliability indices that have historically been used to characterize reliability including Cohen's kappa, Spearman's rank order correlation coefficient rho, and Cronbach's alpha. As described above, we repeated this simulation 100 times to examine the trends across the simulations.

To address research question 3, seeking to understand the impact of reliability decisions, we created classroom observation scores from the data sets generated for research question 2 and then correlated those scores with a randomly generated variable representing an outcome of interest. In this case, we decided that a study outcome of interest that would be easy to interpret was classroom average student achievement scores. We can interpret differences in correlations between classroom observation scores and student achievement data as simple effect sizes and then explore differences in those effect sizes as representing the impact of different decisions. For each of the 100

simulated data sets, we generated student achievement averages to have a mean correlation of .4 between the original score and the student achievement average. We chose .4 because a correlation of that magnitude would constitute a moderately significant relation between instruction and student achievement. To examine the impact of variation in classroom observation IRR on the outcome, we focused on variation in percentage of exact agreement as well as the magnitude of score differences. In particular, we examined the four different percentages of exact agreement (60, 70, 80, and 90), and then, for each of those levels of agreement, we examined two different magnitudes of score differences (off by 1 or off by 2).

Because we created a hypothetical classroom observational instrument, we had to make some additional decisions about how to aggregate scores across raters and across rubrics. We opted to aggregate scores across raters by averaging scores[1] . For example, in the case of scores that differed by 2, if rater 1's score on a particular rubric was 2 and rater 2's scores was 4, then the final rubric score was recorded as a 3. Therefore, at each level of percent exact agreement (60, 70, 80, and 90), we generated 2 different final rubric scores (off by 1 or off by 2), resulting in a total of 8 different sets of classroom observation scores (with final scores for 10 rubrics across 100 classrooms). In addition, we decided to average across all ten rubrics to create an overall score for each observation, resulting in 8 different overall scores corresponding to the two different magnitudes of disagreement for each of the four levels of percent exact agreement. We also created an average classroom score for each of the 100 original scores, used to generate the student achievement averages. These original scores are the scores for rater 1, and allow for comparison under the different inter-rater reliability scenarios. In sum, for each of the 100 hypothetical classrooms, we generated 9 different classroom observation scores to be correlated with the simulated student achievement averages. We repeated this simulation a total of 100 times to explore trends in variation in relations between classroom observation scores and student achievement averages.

In the results section, we describe the results from our 100 simulations of a classroom observation study to describe the impact of IRR decisions within a hypothetical research scenario. In particular, we examined the differences between the correlations that resulted from the different decisions. For example, if the expected correlation was .4 and the correlation between the classroom observation score at 80% exact agreement and off by 1 and the simulated student achievement data was .3, then the difference between the two correlations was .1. We interpret the difference between the two correlations as the impact of that decision

Results: Research Questions 2 and 3

Research Question 2: Comparing Different Reliability Indices

Perhaps our most significant finding is that the magnitude of disagreement (i.e., off by 1 or off by 2) matters, both when comparing different reliability indices and, as we describe below, when examining the impact of inter-rater reliability decisions on an outcome measure. The mean kappa, rho, and alpha for each simulated percentage of exact agreement are given in Table 2 and represented below in Figures 1 and 2. Figures 1 and 2 demonstrate that, in general, as percentage of exact agreement increases, so do the other reliability indices. In the case of Figure 1, displaying the reliability indices when the amount of disagreement is "off by 1," the graph demonstrates that the indices of consistency, rho and alpha, were always above .8 and were always greater than the indices of consensus, percentage of exact agreement and kappa. In particular, even at 60 percent exact agreement, rho was .83 and alpha was .90. This is a clear example of variation in magnitude between the different reliability indices. The relative ranking of the indices was different in the case of disagreement by two score points, displayed in Figure 2. In this case, the indices of consistency, rho and alpha, were always lower than percent agreement. Therefore, in general, indices of consensus are correlated and indices of consistency are correlated, but the former are not sensitive to the magnitude of disagreement between scores and the latter are very sensitive to the magnitude of disagreement between scores, as might be expected based on what they are purported to measure.

[1] In the case of aggregating scores across raters we modeled decision in two different ways, coming to consensus or averaging scores, and found that the difference between the two was small. So for simplicity, we present results from the cases with averaged scores.

Table 2. Measures of IRR for two different magnitudes of score disagreement

% Exact Agreement	kappa		rho		alpha	
	Off by 1	Off by 2	Off by 1	Off by 2	Off by 1	Off by 2
60	0.459	0.470	0.830	0.312	0.900	0.502
70	0.592	0.598	0.869	0.465	0.925	0.649
80	0.726	0.729	0.912	0.631	0.950	0.779
90	0.862	0.863	0.954	0.809	0.975	0.896

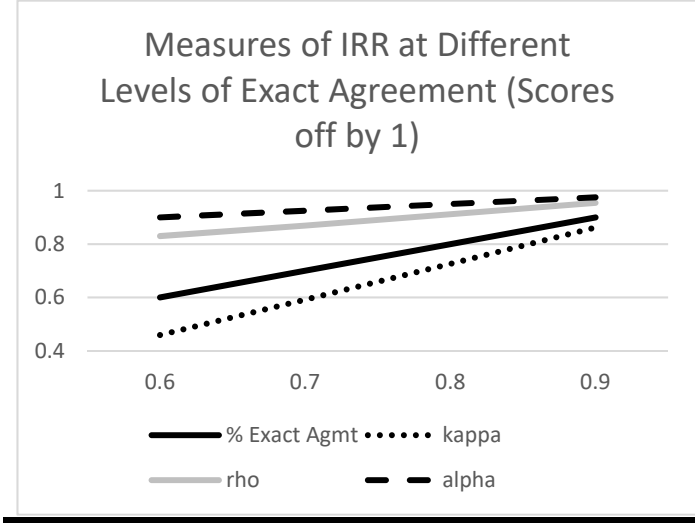

Figure 1. Measure of IRR at Different Levels of Exact Agreement (Score off by 1)

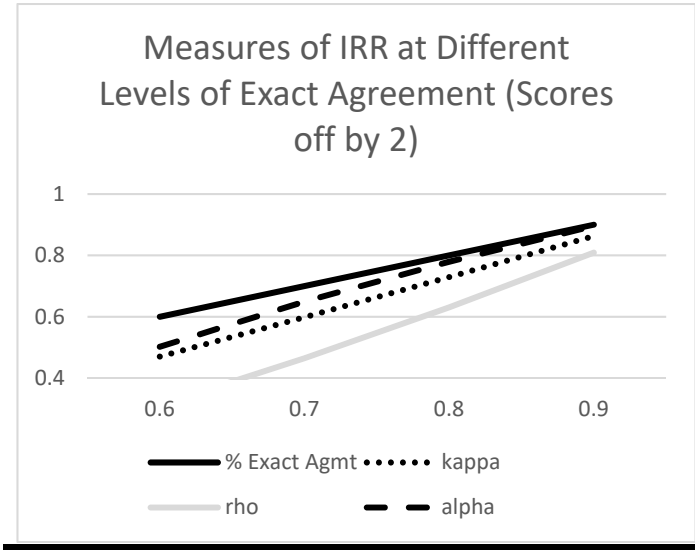

Figure 2. Measures of IRR at Different Levels of Exact Agreement (Scores off by 2)

Research Question 3: Understanding the Impact of Classroom Observation Scoring and Reliability Decisions

As described above, we generated classroom student achievement averages, expected to have an average correlation of .4 with the simulated original classroom-level scores. In our simulated data, the average correlation between classroom student achievement and original classroom scores was .402, with a minimum value of .190 and a maximum value of .573. We were interested in how correlations of those same student achievement averages and different classroom scores varied based on differences in IRR. Across the four different percentages of exact agreement, mean correlations ranged from .317 to .398 and correlations themselves ranged from .041 to .594 (See Table 3). When comparing correlations at different percentages of exact agreement, mean correlations decreased as percentages of exact agreement decreased. This means that imprecision from ratings of classroom instruction resulted in a reduction in effect size, on average. Further, this trend was exacerbated in the case where scores differed by 2 points. For example, the mean correlation at 60% exact agreement when scores were within 1 point (r = .381) was nearly the same as the mean correlation at 90% exact agreement when scores were within 2 points (r = .383). In other words, inter-rater reliability of 60% and 90% exact agreement resulted in the same average reductions to effect size. We discuss this finding in greater detail in the discussion section.

To better understand the average effects on correlations under the two simulated magnitudes of disagreement, we graphed differences between expected and average correlations at different percentages of exact agreement (see Figures 3 and 4). On the horizontal axis of each graph is the expected correlation, the correlation between the original classroom observation scores and the student achievement averages. Recall that these scores were simulated such that the average correlation would be .4 but that they ranged from .190 to .573. On the vertical axis is the difference between the expected correlation and the correlations for the data with different IRR characteristics. For both magnitudes of disagreement, correlations tended to be smaller than the expected correlation, represented by mostly positive y-values in the scatterplots. Comparing Figures 3 and 4 further demonstrates that on average, the correlation differences were greater when scores were off by 2 rather than off by 1. In fact, some of those correlations were

considerably smaller than the expected correlation. For example, if you take the data point marked with an X in the graph for 60% exact agreement in Figure 4, this represents a data set with an expected correlation of .400 and an actual correlation of .148. With respect to interpretation, a correlation of .1 is considered a small effect size, whereas a correlation of .3 is considered moderate, and a correlation of .5 is considered large (Cohen, 1988). Therefore, what was a moderate correlation became a small correlation because of error introduced by the rating process. Examining the graphs for 90% exact agreement in the bottom right corners of Figures 3 and 4 reveals that even in this scenario, correlations can be reduced by as much as .05 or .1 when the scores differ by 1 or 2, respectively. Therefore, these simulations suggest that the IRR of the scores—simulated in this analysis by percentage of exact agreement and the magnitude of the disagreement—has a measurable impact on correlations with an outcome. Specifically, the imprecision introduced by the rating resulted in a reduction in effect size, on average.

Table 3. Correlations between IRR Decision-Based Classroom Scores and Simulated Student Achievement Averages

Percent Exact Agreement		*M*	*SD*	Min	Max
60	Off by 1	0.381	0.074	0.141	0.531
	Off by 2	0.317	0.087	0.041	0.509
70	Off by 1	0.385	0.073	0.157	0.553
	Off by 2	0.340	0.086	0.065	0.594
80	Off by 1	0.392	0.073	0.169	0.538
	Off by 2	0.358	0.078	0.113	0.538
90	Off by 1	0.398	0.074	0.201	0.553
	Off by 2	0.383	0.078	0.161	0.537

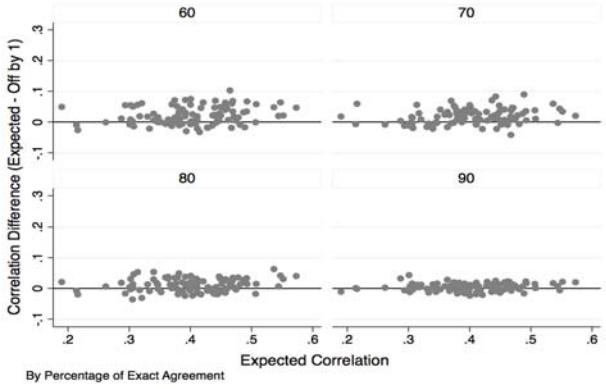

Figure 3. Comparing Expected and Actual Correlations when Magnitude of Disagreement is 1

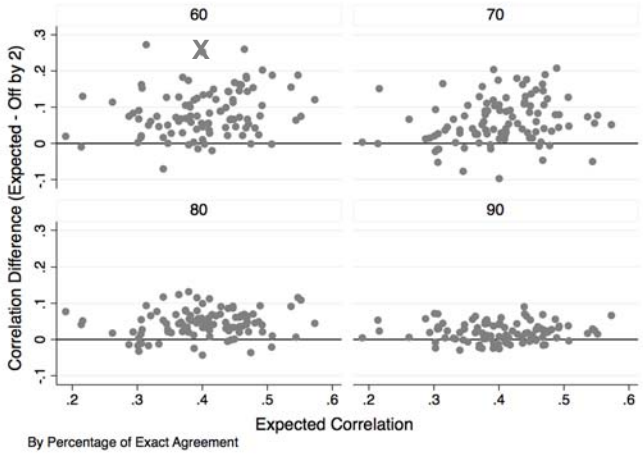

Figure 4. Comparing Expected and Actual Correlations when Magnitude of Disagreement is 2

Discussion

Classroom observation IRR matters because it is about trust. In particular, "we need to measure the extent of agreement among independent replications in order to estimate whether we can trust the generated data in subsequent analyses" (Krippendorff, 2016, p.139). We entered into this analysis as producers of quantitative classroom observation research, wanting more empirical evidence for our IRR decisions. Yet, we also view our findings as critical for consumers of quantitative classroom observation research. Whether, as a reviewer, needing to determine if a study's evidence is sufficient, or, as a reader, simply trying to determine the extent to which there is credible evidence for a study's claims, consumers need to understand the extent to which they should trust data as documenting (and supporting) what they claim to document. We view reporting information about IRR as one important piece of this building of trust.

As we looked across two top journals in two different fields of education, over a recent 10-year span, we learned quite a bit about trends in quantitative classroom observation research. First, we were struck by the relatively small percentage of studies (3.4%) that utilized data from quantitative classroom observation tools as a variable in quantitative analyses. It is possible that historically these two journals have been qualitatively oriented in their research traditions and this is one reason for this small percentage of studies. Alternatively, it could be that this work is challenging to carry out in a rigorous fashion (for example, we have had push back from reviewers when we have included

complex measures with 70% exact rater agreement), so few such studies are accepted to these top-tier journals. Despite the low occurrence of such studies in the two journals we sampled from, we believe that classroom observation research is here to stay and likely to increase in prevalence given the growing emphasis of classroom observation in schools and in policy circles (e.g., Swanson et al., 2012; Van der Lans et al., 2016).

Second, despite the recommendation in our field to move toward the use of observational systems with many scoring and reliability decisions specified by the developer (Hill et al., 2012), few of the studies we reviewed utilized classroom observation tools within a specified observational system. Therefore, it is likely that many of the researchers were confronted with decisions similar to those we modeled in this study. And, until well-specified observational systems become commonplace, scoring and reliability decisions will continue to be important. Even in the case where a researcher is utilizing a well-specified observational system, he or she still must attend to and report IRR information to ensure that the data obtained in using the system is to be trusted in representing the phenomena of interest.

Third, we were struck by the imprecision around the terms inter-rater agreement and reliability, with most studies reporting measures of consensus. Prior to this analysis, we, ourselves, used the terms interchangeably and hence were not surprised by the imprecision across the field, but our simulation studies suggested that precision around the particular reliability indices and knowing what each actually measures have important implications for study outcomes. In particular, we found through simulation that different types of IRR indices better account for different sources of variation within the data. By only providing measures of consensus (e.g., percent agreement or Cohen's kappa), researchers are omitting important information about the accuracy of the rating process (beyond the alignment between raters). For example, we found that correlation-based measures of consistency better account for the magnitude of disagreement between raters, whereas the measures of consensus better account for the agreement between raters (Hintze & Matthews, 2004). In addition, none of the studies used any of the newer indices representing methodological developments in assessing consensus as a dimension of IRR. Below, in the implications section, we offer suggestions for the reporting of information about classroom observation IRR.

Fourth, in reviewing the literature, we found that there was inconsistent reporting about the different phases of the IRR process (e.g., training reliability). While we attribute some of this variation to a lack of guidance about what to report, we attribute most of the variation and generally minimal reporting to a lack of space available in research studies. When faced with journal page limits, there is typically little room for researchers to report all of the necessary information about definition, training, development, reliability, validity, and limitations for classroom observation instruments (Vaughn & Briggs, 2003). All but one of the studies we reviewed provided some information about consensus or consistency dimensions of IRR. It seems that the trends in reporting varied by field with more studies reporting data collection reliability (over training reliability) in *JRME*, and more studies reporting training reliability (over data collection reliability) in *RRQ*. Overall, only 7 of 17 studies included information about both training reliability and data collection reliability. It is important to know that raters understood both how to use the observational instrument reliably and that raters continued to use the observational instrument reliably over the course of study data collection. Without this information, there is no way to know that the classroom observation ratings that have been collected accurately and represent the phenomena that they were intended to represent. While we have not focused on *intra*-rater reliability (i.e., internal consistency of a rater, Flemenbaum & Zimmermann, 1973) within this analysis, the attention to data collection reliability and rater drift is one way to account for intra-rater reliability over time (Kazdin, 1977).

Fifth, when studies reported IRR they often just reported reliability above a particular threshold (often .80). As discussed above, while the .80 threshold is convention in the field, we found no empirical basis for the threshold (Lance et al., 2006), especially for its use with respect to percentage of exact agreement. Combined with the findings from the simulation analyses, which suggest that mean correlations with an outcome are slightly lower at lower levels of percentage of exact agreement but that the magnitude of score differences has a more dramatic effect on correlations, it seems that adhering to strict thresholds for percentages of exact agreement is less helpful than reporting exact

point estimates and examining measures of both consensus and consistency.

In sum, findings from our simulation studies suggest that IRR matters when it comes to outcomes. In particular, error introduced by the rating process tended to decrease correlations, and in some cases, fairly significantly. Further, one surprising finding discussed briefly above was the significance of the magnitude of disagreement between scores. Our decision about modeling this phenomenon arose as we were simulating the data. We decided to model scores that were off by 1 and scores that were off by 2, with the assumption that an instrument with a range of 5, like our hypothetical instrument, could have more dramatic discrepancies, but that likely most disagreements would be either 1 or 2 off. Hence, simulating data that was consistently off by 2 allowed us to consider a possible, but likely worst-case, scenario for IRR. Through the simulations we found that the differences in magnitude of disagreement had a relatively large impact on the relation with the outcome. The impact of this decision and related reporting within studies became clearer as we realized that the most commonly reported measures of consensus do little, if anything, to reveal whether there are differences in the magnitude of disagreement. Therefore, there are clear implications for the measuring and reporting of IRR within studies employing classroom observation instruments and we describe those implications following our discussion of a few study limitations.

Limitations

While we feel that our study offers important empirically-based guidance with respect to quantitative classroom observation instruments and the reporting of results, there are a few study limitations. First, with respect to research question one, examining how researchers are reporting classroom observation IRR, we only examined 2 journals for 10 years to document recent approaches to the reporting and measuring of IRR. It is likely that other educational research journals differ from *JRME* and *RRQ* both with respect to the frequency of use of classroom observation instruments and the approaches to reporting information about reliability. Our intent was to get a feel for current practice in two different fields within education rather than to conduct a comprehensive literature review. Future research might systematically examine these same things in other educational research journals to better understand differences in the prevalence of quantitative classroom observational instruments and the reporting of IRR for such instruments between disciplines within education.

Second, our hypothetical instrument was generated to resemble several well-known classroom observational instruments but to not be exactly like any one of them. We acknowledge that there are a number of other approaches to classroom observation that are not captured by our hypothetical instrument, including time sampling approaches or frequency counts. We posit that some of the same issues or decisions apply even with slightly different approaches to classroom observation. To specifically focus on decisions that arise with those other approaches, future studies might replicate these simulations with instruments that model those other approaches. Given that another important decision that we have not modeled in this analysis is the decision about which instrument to use, it will be important to understand how these decisions matter based on other approaches to the quantitative measurement of teacher and student behavior.

Implications

In this section, we offer some implications for measuring and reporting IRR. First, we recommend that the following things be reported in each study utilizing a quantitative classroom observational instrument: 1) a validation argument for the choice of the instrument including information about any prior generalizability studies that have been conducted with the instrument; 2) information about the training process IRR; and 3) information about the data collection process, including key decisions as well as data collection IRR. When relying on commonly used indexes of IRR, we recommend using and reporting both an index of consensus (e.g., Cohen's kappa) as well as an index of consistency (i.e., a correlational measure).

Future research should investigate the use of newer consensus indices such as Krippendorf's (2013) α and alternative approaches to describing IRR in classroom observational research. We also recommend that when information about reliability is reported, it should include as much precision with respect to the index and the point estimate as possible. For example, reporting inter-rater agreement above .8 is not as precise as reporting the percentage of exact inter-rater agreement as .82. Given that the thresholds are relatively arbitrary, it is important for the consumer of the research to look across all of the information presented to make an

informed decision about the trustworthiness of the data and related results. While this would take up a bit more space in a journal article, it could be concisely presented in 1-2 paragraphs that lend considerable credence to the rigor of the study. Another approach that is common in other fields (e.g., chemistry, medical research) but has not yet been adopted in education is the use of a Bland-Altman plot (Bland & Altman, 1986) or some other graphical representation to describe trends in IRR and demonstrate the absence of systematic bias in measurement. Future research should examine the utility of such representations in examining and describing IRR in educational research.

Our unexpected finding about the importance of accounting for the magnitude of disagreement between raters also has implications with respect to how percentage of agreement is calculated. Although not common in the literature we surveyed, some users of classroom observational instruments choose to report percent of agreement within 1 score category rather than percent of exact agreement (and often use the same .80 threshold to justify results). This is a more relaxed approach to describing agreement (Stemler, 2004). For example, in the context of our hypothetical instrument, calculating agreement within 1 score category would mean that if one rater assigned a score of 3 and another rater assigned a score of 2 (or 4) then the two raters would be considered in agreement. If we had considered agreement that way, then any of the scores produced with 60%-90% exact agreement but off by 1 would be considered 100% in agreement. While this practice does attend to the magnitude of score differences, it ignores any disagreements that are off by 1 and collapses score categories so that any disagreements that "count" are relatively major. Given our findings, it seems that rater consensus reported with percentage of agreement within 1 should not be considered as an alternative to percentage of exact agreement, at least not without a much stricter set of standards.

Another key implication is the application of these findings to settings other than research studies employing quantitative classroom observational instruments. First, given the prevalence of classroom observations for teacher evaluation purposes and the high-stakes nature of those decisions, issues of validity and reliability need to be considered in those settings as well (Cohen & Goldhaber, 2016). For example, how are the raters (e.g., principals) being trained and how reliable are their ratings in assessing teacher quality? Second, our findings also apply to other approaches within education research including qualitative coding of teacher and/or student actions in classrooms as well as measures of fidelity of implementation of classroom-based interventions. In both of these cases, it is important to ensure that the data was reliably produced and to be able to convince research consumers of that reliability.

While we are arguing for slightly greater attention to IRR, our primary hope is to offer some guidance so that producers and consumers of research utilizing quantitative classroom observational instruments can make or evaluate decisions pertaining to IRR within study design and in the reporting of data and results. With a set of guidelines for reporting IRR and empirically-based information about the impact of different decisions, we hope that both producers and consumers of such information are better equipped for their role in the process of knowledge production.

References

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *British Medical Journal*, *314*(7080), 572.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307-310.

Blood, E. & Spratt, K. F. (2007). Disagreement on agreement: Two alternate agreement coefficients. SAS Global forum 2007. Retrieved from: http://www2.sas.com/proceedings/forum2007/186-2007.pdf

Boston, M., Bostic, J., Lesseig, K., & Sherman, M. (2015). A comparison of mathematics classroom observation protocols. *Mathematics Teacher Educator, 3*(2), 154-175.

Boston, M., & Smith, M. S. (2009). Transforming secondary mathematics teaching: Increasing the cognitive demands of instructional tasks used in teachers' classrooms. *Journal for Research in Mathematics* Education, *40*(2), 119-156.

Brown, S. A., Pitvorec, K., Ditto, C., & Kelso, C. R. (2009). Reconceiving fidelity of implementation: An investigation of elementary whole-number lessons. *Journal for Research in Mathematics Education*, *40*(4), 363-395.

Cash, A.H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scare: Degree of calibration and characteristics associated with calibration. *Early Childhood Research Quarterly*, *27*(3), 529-542.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284-290.

Clements, D. H. Sarama, J., Spitler, M. E., Lange, Alissa A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*(2), 127-166.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. doi:10.1177/001316446002000104

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., ... & Schatschneider, C. (2011). Testing the impact of child characteristics× instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly*, *46*(3), 189-221.

Copur-Gencturk, Y. (2015). The effects of changes in mathematical knowledge on teaching: A longitudinal study of teachers' knowledge and instruction. *Journal for Research in Mathematics Education, 46*(3), 280-330.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. doi:10.1007/bf02310555.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Fleiss, J. (1981). Statistical methods for rates and proportions. New York: Wiley.

Flemenbaum, A., & Zimmermann, R. L. (1973). Inter-and intra-rater reliability of the Brief Psychiatric Rating Scale. *Psychological Reports*, *33*(3), 783-792.

Grouws, D. A., Tarr J. E., Chavez, O., Sears, R., Soria, V. M., & Taylan, R. D. (2013). Curriculum and implementation effects on high school students' mathematical learning from curricula representing subject-specific and integrated content organizations. *Journal for Research in Mathematics Education*, *44*(2), 416-463.

Guthrie, J. T., Klauda, S. L., & Ho, A. N. (2013). Modeling the relationships among reading instruction, motivation, engagement, and achievement for adolescents. *Reading Research Quarterly*, *48*(1), 9-26.

Gwet, K. L. (2002). Kappa statistics is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability*, *1*(6), 1–5.

Gwet, K. L. (2012). Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters. Third Edition. Gaithersburg, MD: Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data. *Tutor* Quantitative *Methods Psychology*, *8*(1), 23-34.

Hartmann, D. P., Barrios, B. A., & Wood, D. D. (2004). Principles of behavioral observation. In M. Hersen (Ed.), *Comprehensive Handbook of Psychological Assessment* (Vol. 3, pp. 108-137). Hoboken, N. J.: John Wiley & Sons.

Hill, H. C., Charalambos, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64. doi:10.3102/0013189X12437203

Hill, H. C., & Shih, J. C. (2009). Examining the quality of statistical mathematics education research. *Journal for Research in Mathematics Education*, *40*(3), 241-250.

Hintze, J. M., & Matthews, W. J., (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, *33*(2), 258-270.

Ho, A. D., & Kane, T. J., (2013). The reliability of classroom observations by school personnel. Retrieved from http://www.metproject.org/downloads/met_reliability of classroom observations_researchpaper.pdf

Jackson, K. J., Garrison, A. L., Wilson, J., Gibbons, L., & Shahan, E. (2013). Exploring relationships between setting up complex tasks and opportunities to learn in concluding whole-class discussions in middle-grades mathematics instruction. *Journal for Research in Mathematics Education*, *40*(4), 646-682.

Jonnson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review 2*(2), 130-144.

Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, *1*(4), 427-452.

Kazdin, A. E. (1982). Single case research designs: Methods for clinical and applied settings. New York: Oxford Press.

Kelcey, B., & Carlisle, J. F. (2013). Learning about teachers' literacy instruction from classroom observations. *Reading Research Quarterly*, *48*(3), 301-317.

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis*, *21*(4), 345-363.

Kelsey, B. & Carlisle, J. F., (2013). Learning about teacher' literacy instruction from classroom observations. *Reading Research Quarterly*, *48*(3), 301-317.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hrobjartsson, A., . . . Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, *64*, 96-106.

Krippendorff, K. (2013). Content analysis. An introduction to its Methodology (3rd ed.). Thousand Oaks, CA: Sage.

Krippendorff, K. (2016). Misunderstanding reliability. *Methodology*, *12*(4), 139-144.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria what did they really say? *Organizational Research Methods*, *9*(2), 202-220.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159-174. doi:10.2307/2529310

Liao, S. C., Hunt, E. A., & Chen, W. (2010). Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Annals Academy of Medicine*, *39*(8), 613-618.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276-282.

MET Project. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET projects three-year study. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved April 25 2017, from metproject.org/downloads/MET_Ensuring_Fair_and_ Reliable_Measures_Practitioner_ Brief.pdf

Pianta, R. C. & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109, doi:10.3102/0013189X09332374.

Robert, C. P., & Casella, G. (2004). Monte Carlo Statistical Methods. Second edition, New York: Springer.

Rodgers, E., D'Agostino, J. V., Harmey, S. J., Kelly, R. H., & Brownfield, K. (2016). Examining the nature of scaffolding in an early literacy intervention. *Reading Research Quarterly*, *51*(3), 345-360.

Sailors, M., Hoffman, J. V., David Pearson, P., McClung, N., Shin, J., Phiri, L. M., & Saka, T. (2014). Supporting change in literacy instruction in Malawi. *Reading Research Quarterly*, *49*(2), 209-231.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321–325.

Semmelroth, C. L., & Johnson, E., (2014). Measuring inter rater reliability on a special education observation tool. *Assessment for Effective Intervention*, *39*(9), 131-145.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428.

Silverman, R., & Crandell, J. D. (2010). Vocabulary practices in prekindergarten and kindergarten classrooms. *Reading Research Quarterly*, *45*(3), 318-340.

Silverman, R. D., Proctor, C. P., Harring, J. R., Doyle, B., Mitchell, M. A., & Meyer, A. G. (2014). Teachers' instruction and students' vocabulary and comprehension: An exploratory study with English monolingual and Spanish–English bilingual students in Grades 3–5. *Reading Research Quarterly*, *49*(1), 31-60.

Spearman, C (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101. doi:10.2307/1412159.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approached to estimating integrated reliability. *Practical Assessment, Research & Evaluation*, *9*(4), 1-11.

Stolarova, M., Wolf, C., Rinker, T., & Bielmann, A. (2004). How to assess and compare inter-rater reliability, agreement and correlation of ratings: An exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in Psychology*, *5*, 1-13. doi:10.3389/fpsyg.2014.00509

Strong, M. (2011). The highly qualified teacher: what is teacher quality and how do we measure it? NY: New York: Teachers College Press.

Stuhlman, M. W., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2014). How to select the right classroom observation tool. Retrieved from http://curry.virginia.edu/uploads/resourceLibrary/CASTL_practioner_Part3_single.pdf

Swanson, E., Solis, M. Ciulio, M., & McKenna, J. W. (2012). Special Education Teachers' Perceptions and Instructional Practices in Response to Intervention

Implementation. *Learning Disabilities Quarterly*, *35*(2), 115-126.

Tarr, J. E., Grouws, D. A., Chavez, O. & Soria, V. M. (2013). The effects of content organization and curriculum implementation on students' mathematics learning in second-year high school courses. *Journal for Research in Mathematics Education*, *44*(4), 683-729.

Van de Lans, R. M., van de Grift, W. J. C. M., van Veen, K., & Marjon, F. B., (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Students in Educational Evaluation*, *50*, 88-95.

Vaughn, S., & Briggs, K. L., (Eds.) (2003). Reading in the classroom: Systems for the observation of teaching and learning. Baltimore: Brookes Publishing.

Vaughn, S., Swanson, E. A., Roberts, G., Wanzek, J., Stillman-Spisak, S. J., Solis, M., & Simmons, D. (2013). Improving reading comprehension and social studies knowledge in middle school. *Reading Research Quarterly*, *48*(1), 77-93.

von Eye, A., & Mun, E. Y. (2005). Analyzing rater agreement: Manifest variable methods. Mahwah, New Jersey: Erlbaum.

White, T. G., Kim, J. S., Kingston, H. C., & Foster, L. (2014). Replicating the effects of a teacher-scaffolded voluntary summer reading program: The role of poverty. *Reading Research Quarterly*, *49*(1), 5-30.

Wilhelm, A. G. (2014). Mathematics teachers' enactment of cognitively demanding tasks: Investigating links to teachers' knowledge and conceptions. *Journal for Research in Mathematics Education*, *45*(5), 637-675.

Citation:

Wilhelm, Anne Garrison, Gillespie Rouse, Amy, & Jones, Francesca. (2018). Exploring Differences in Measurement and Reporting of Classroom Observation Inter-Rater Reliability. *Practical Assessment, Research & Evaluation*, *23*(4). Available online: http://pareonline.net/getvn.asp?v=23&n=4

Corresponding Author

Anne Garrison Wilhelm
Assistant Professor
Southern Methodist University

email: awilhelm [at] mail.smu.edu