Journal of Educational Supervision

Volume 2 | Issue 2 Article 1

2019

Teacher Evaluation and Reliability: Additional Insights Gathered from Inter-rater Reliability Analyses

Sally J. Zepeda
University of Georgia, szepeda@uga.edu

Albert M. Jimenez
Kennesaw State University, Ajimen17@kennesaw.edu

Recommended Citation

Zepeda, S. J., & Jimenez, A. M. (2019). Teacher Evaluation and Reliability: Additional Insights Gathered from Inter-rater Reliability Analyses. *Journal of Educational Supervision, 2* (2). https://doi.org/https://doi.org/10.31045/jes.2.2.2

Empirical Research

Teacher Observation and Reliability: Additional Insights Gathered from Inter-rater Reliability Analyses

Journal of Educational Supervision
11 – 26
Volume 2, Issue 2, 2019
DOI: https://doi.org/10.31045/jes.2.2.2
https://digitalcommons.library.umaine.edu/jes/

Sally J. Zepeda[1] & Albert M. Jimenez[2]

Abstract

Using a newly created teacher evaluation instrument, Inter-rater Reliability (IRR) analyses were conducted on four teacher videos as a means to establish instrument reliability. Raters included 42 principals and assistant principals in a southern US school district. The videos used spanned the teacher quality spectrum and the IRR findings across these levels varied. Key findings suggest that while the overall IRR coefficient may be adequate to assess the validity of a classroom observation instrument, the overall coefficient may be unstable across the various teacher performance levels. Findings also strongly suggest that raters are much more likely to agree when they see high-quality teaching when compared to levels of agreement regarding low-quality teaching.

Keywords

teacher evaluation; classroom observation; inter-rater reliability; observation instrument construction; Kappa; Gwet's AC1

[1] University of Georgia, USA
[2] Kennesaw State University, USA

Corresponding Author:
Sally J. Zepeda (Educational Administration and Policy, University of Georgia, 815 College Station, River's Crossing, Room 312. Athens, GA. 30602, USA)
Email: szepeda@uga.edu

Introduction

Broadly, this study was designed to contribute to the research and discussion surrounding the establishment of inter-rater reliability for classroom observers in the context of teacher evaluation. Specifically, this study aimed to address a gap in the literature by assessing if inter-rater reliability was consistent in a classroom observation instrument developed for a new teacher evaluation system across the teacher quality spectrum. This topic is timely given that teacher evaluation systems around the country were revamped in response to conditions set forth in the 2009 federal grants program, Race to the Top (RTTT). RTTT, a segment of the American Recovery and Reinvestment Act of 2009, was designed, in part, to increase the effectiveness of educators—both teachers and educational leaders (Clifford & Ross, 2011; Lohman, 2010). After the No Child Left Behind Act of 2001 laid dormant waiting to be amended and driven by the incentives associated with RTTT, policy, research, and practice focused on teacher and leader effectiveness addressing primarily how to assess it and what should or should not be used to assess primarily teacher effectiveness (Donaldson & Papay, 2015; Firestone, 2014; Hallgren, James-Burdumy, & Perez-Johnson, 2014; Steinberg & Donaldson, 2016; Steinberg & Quinn, 2017).

Including in the mix, new classroom observation instruments, which are undeniably centerpieces of new teacher evaluation systems, albeit fraught with diverging points-of-view, must be based on multiple measures and assessed for reliability (AERA, 2015; Darling-Hammond, 2015; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Garver, 2019; Jimenez & Zepeda, 2016). Given that 48 states require classroom observations as part of their teacher evaluation systems (Doherty & Jacobs, 2015) and classroom observation scores are factored with other value-added processes to "sum total" a teacher's performance, it is imperative to examine issues of reliability and the inter-reliability of classroom observations made by school leaders.

Literature Review

With the increased focus on teacher quality, issues relating to measuring teacher effectiveness is of increased importance. One of these issues is rater agreement for classroom observations. As the evaluation of teachers becomes tied to issues such as merit pay and continued employment, consistency across raters, typically principals and assistant principals, is essential to fair and accurate teacher performance assessment. A key aspect of measuring teacher performance is the teacher during classroom observations. We examine issues measuring teacher effectiveness, classroom observations, and the design of this quantitative study of inter-rater reliability using the percent of overall agreement, Fleiss' Kappa (1971), and Gwet's AC1 (2002) statistics. These statistics provide IRR coefficients used to assess the reliability of the observation instrument.

Issues Measuring Teacher Effectiveness

Moving beyond the press for highly-qualified teachers as found in the language of the No Child Left Behind Act of 2001, RTTT shifted the conversation and controversies in its language to include teacher effectiveness as measured by bundled algorithms within value-added measures (VAMs), inextricably linking student test scores to individual teachers (Mette et al., 2017). In brief, the provisions of RTTT stipulated that "teacher effectiveness is evaluated, in *significant*

part, by student growth" including "multiple observation-based assessments of teacher performance" as part of the evaluation system (U.S. Department of Education, 2009, p. 12, emphasis added).

The classroom observation has become embedded now formally in teacher evaluation systems. This shift focused attention on a critical need to create new evaluation instruments designed to measure teacher effectiveness, including rubric-centric classroom observation instruments. Historically, the data gathered during teacher evaluation have been problematic, offering little to improve the quality of teaching and learning (Darling-Hammond, 2016; Kraft & Gilmour, 2017; Peterson, 2000; Weisberg, Sexton, Mulhern, & Keeling, 2009).

Through examining research and the literature, the problems that are most impactful to teacher evaluation data, center around three issues. First, teacher observations, which are a large part of any evaluation system, are completed too infrequently and for too short in duration (Zepeda, 2017); teacher and leader observations do not culminate with conversations focused on the improvement of instruction (Weisberg et al., 2009; Zepeda, Lanoue, Price, & Jimenez, 2014).

Second, many teacher evaluation instruments cannot pass a validity challenge because they fail to completely assess all areas of what it means to be a teacher (Darling-Hammond, 2013; Haefele, 1993; Kraft & Gilmore, 2017; Weisberg et al., 2009). Third, the research has shown that the data gathered from teacher observations, completed by principals, are unreliable (Stodolsky, 1984) primarily because they lack sufficient knowledge about teacher evaluation, the observation process, and the accurate rating of teachers (Kauchak, Peterson, & Driscoll 1985; Medley & Coker, 1987; Peterson, 2000; Stodolsky, 1984; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984). Moreover, evaluator bias may dictate scores and feedback based on what is measured by the individual conducting the observation (Steinberg & Donaldson, 2016). There are other thorny issues associated with classroom observations and leaders; namely they have difficulty discerning teacher performance levels, and the observation instruments are segmented cueing them to measure only small portions of the broader construct of teaching (Cohen & Goldhaber, 2016).

Weisberg et al. (2009) published a particularly troubling study examining both the ratings of teachers and uses of teacher evaluation data. In this study, they examined school districts, both large and small, and found that nearly all teacher evaluation scores were found to be good or great, excellence in teaching was not rewarded by districts, professional development was rarely tied to the results of the evaluation, new teachers were generally rated above being satisfactory, and negative results of a teacher evaluation rarely led to dismissal. These findings are striking examples of the many problems associated with teacher evaluation, and they provide further examples of why research and professional associations fail to support using such systems for personnel decisions (AERA, 2015; National Council of Teachers of English, 2012; National Council of Teachers of Mathematics, 2011). Essentially, the findings of the Weisberg et al. (2009) study bring to light the need to have valid and reliable observation instruments prior to using these types of data to make personnel decisions. Moreover, reliable data are needed to support school-wide efforts in the identification of professional learning needs of teachers (Zepeda, 2019).

Classroom Observations

Stemming from the 1800s when classroom observations were conducted by "visitors" external to the schoolhouse and despite the lack of attention in the research, classroom observations have been considered to be the heart of teacher evaluation systems (Ponticell et al., 2019; Zepeda, 2017). Emerging in the late 1950s, the clinical model of instructional supervision included a cyclical process—the pre-observation conference, the classroom observation, and the post-observation conference (Cogan, 1973; Goldhammer, 1969; Ponticell et al., 2019). The intent was for teachers to gain insights about their teaching and the relationship on improving it within the context of the classroom environment. Through this cyclical process, teachers were engaged in discussions about classroom practices and the improvement of teaching. Fast forward to the 1980s, these components of instructional supervision became the "mainstay of teacher evaluation systems" (Zepeda, 2013, p. 65).

The shift of embedding "parts of the" clinical model of instructional supervision, namely the classroom observation, in the process of teacher evaluation, will more than likely add fuel to the argument that supervision and evaluation are irreconcilable processes within the field of instructional supervision (Glanz & Neville, 1997; McCarty, Kaufman, & Stafford, 1986). Time will tell. However, there are studies that illustrate that teacher evaluation, supervision, and professional development can co-exist and enhance the end result for teachers (Derrington & Campbell, 2018; Mette et al., 2017) supporting coherence across systems (Zepeda, 2017). For the purposes of this article, we focus on examining reliability and validity of what school leaders in a single school district see, report, and rate about teacher performance during classroom observations embedded in a teacher evaluation system. This study focuses broadly on the consistency and inconsistencies with ratings, especially classroom observations, and the need to have reliable observation instruments.

The lack of consistency in ratings has, at least in part, been associated with performance assessments in an educational context that are often designed and implemented before methodological issues are examined and addressed (Linn & Baker, 1996). As such, outlining clear methods for validating these instruments, including establishing inter-rater reliability, are vital because "observation ratings inherently rely on evaluators' professional judgment" and "there is always a question of how much the ratings depend on the particular evaluator rather than the educator's actual performance" (Graham, Milanowski, & Miller, 2012, p. 4). This concern can be addressed, in part, through ensuring that ratings are consistent across raters.

Research Design and Methodology

For the present study, the methods employed are quantitative. The inter-rater reliability coefficients reported, namely the percent of exact agreement, Fleiss' kappa (1971), and Gwet's AC1 statistic (2002), were all calculated using AgreeStat, version 2011.3. The data for this study came from one primary source. Inter-rater reliability, in this study, refers to "a measurement of the consistency of the *absolute value* of evaluators' ratings" (Graham et al., 2012, p. 5, emphasis in the original), as the classroom observation instruments examined in this study required either a "yes" or "no" response across performance standards. This "yes" or "no" response is also known as an exact match.

Inter-rater reliability is a technique for determining the consistency of raters when tasked with accurately assessing what they have seen, and also that trained raters are able to be consistent across ratings. Assessing inter-rater reliability in a teacher evaluation context is vital as the teacher evaluation process becomes tied to outcomes such as continued employment and merit pay. Using this instrument and assessing the inter-rater reliability of the raters in this study provides insight into the procedures needed and the importance of such procedures.

There are a number of statistical means of evaluating inter-rater reliability, and some of the more common methods used include percentage of exact agreement, Cohen's kappa (1968) and its variations, and the intra-class correlation coefficient (Cook & Beckman, 2006; Graham et al., 2012). While these methods are common, additional methods, such as Gwet's AC1 statistic (2002) have also been created as a means to assess inter-rater reliability. Typically used in medical research, Gwet's AC1 statistic has been empirically shown to be a more stable measure of chance-corrected inter-rater reliability, combatting the paradox that sometimes appears in variations of kappa, primarily high values of overall agreement that can produce low values of kappa (Feinstein & Cicchetti, 1990; Gwet, 2002, 2012; Jimenez & Zepeda, in press). These are the measures of inter-rater reliability calculated and reported in the current study.

Context of the Research Site

Located in a southeastern state, the Developmentally Appropriate School District (DASD, a pseudonym) serves just over 13,000 students. Of these students, 51% are African American, 23% are Hispanic, 20% are White, 2% are Asian, and the racial breakdown of students has remained relatively consistent. Nearly 12% of the students have English as their second language and approximately 9% of the district's students are served through the English as a Second Language (ESOL) program. Approximately 11% of students in the district are served through gifted education programs and about 11% are special needs students. Students are served through the work of 2691 employees—1038 of which are teachers. Over 70% of the district's teachers have advanced degrees, 250 are certified in gifted education, 16 are National Board-Certified teachers, and 7 are designated as state-level Master Teachers.

At the time of this study, the 21 schools in the Developmentally Appropriate School District (DASD) were led by a remarkably stable cadre of principals whose tenure in the district was marked by longevity measured by years in the position at their schools. A total of

- 14 principals had served as a principal in the district for the past 6 years;
- 11 schools had the same principal for the past 6 years.

One school had experienced a principal change in the past six years, and three have had two principal changes. It is important to note that three of the principal changes were the result of current principals taking a position at another school within the district. Only one of the district's 21 schools has experienced a principal change in the past 3 years.

The Context of Teacher Evaluation in the Developmentally Appropriate School District

The philosophy of evaluation for the DASD was grounded in the fundamental belief that the purpose of evaluation is to develop master teachers. Master teachers improve their effectiveness by embracing research-based instructional practices through on-going professional development that impacts the practice of colleagues in their school and district; and most importantly, provides classroom experiences where all students achieve at the highest level. Supporting teacher growth, the DASD invested time and resources in developing its own teacher evaluation system that was predicated on classroom observations to reflect its focus on instructional improvement. The DASD worked with a local university professor with expertise in instructional supervision to lead its efforts in creating classroom observation protocols and instruments, including rubrics associated performance standards and elements that describe each. Great care was taken to develop a teacher evaluation system that was reflective of its emphasis on instruction.

The DASD made a purposeful decision to develop their own evaluation system to align support to teachers while implementing standards-based instruction in their classrooms. The DASD spent two years developing the teacher evaluation system before beginning its use. At the onset of this undertaking, the empirical research and best practices were culled from the literature on standards-based instruction and classroom practices. System-wide district and building-level leaders as well as teachers were involved in the process of creating the language used to describe standards-based teaching. Second, a rubric was developed identifying performance-based standards (e.g., standards-based instruction, assessment of student learning, etc.) with accompanying elements that amplified what the standard would look, and sound liken in practice. Third, from this rubric, the classroom observation instrument was developed.

The process of developing the teacher evaluation system was an iterative process in that teachers and administrators reviewed, edited, and gave feedback to the rubric and then the classroom observation instrument. Changes were made based on numerous rounds of feedback and input. Once there was agreement on these documents, professional learning for teachers and school leaders began. For principals and assistant principals, professional development focused on the uses of the classroom observation instruments, and this professional development was ongoing for two years prior to the current study. Principals and assistant principals met monthly spending approximately four hours at each meeting focusing on using the observation protocols, applying ways to collect data, to present the data from classroom observations, and to engage in conversations with teachers. Numerous simulations included watching videos of teachers "teaching to the standards." After viewing videos, school leaders engaged in large-and small debriefing on how teachers would be rated, what types of instructional behaviors were noteworthy around the standards, and then, if applicable, what was missing from the lesson observed. In between the monthly professional learning, system and building-level leaders conducted walk-throughs at each principal's school using the classroom observation instruments. Through these associated learning opportunities, principals and assistant principals actively engaged in refining the Tier I and Tier II rubrics and classroom observation instruments.

Classroom Observation Instrument

There are two classroom observation instruments—Tier I and Tier II—used in the DASD teacher evaluation system, and both are aligned primarily with years in teaching. In Tier I, new teachers in years one, two, and three receive assistance designed to promote their successful transition into the profession. The evaluation process at this tier focuses on enhancing strengths and improving weaknesses related to performance standards. Tier I is also where teachers with experience but new to the system begin their growth and development within the DASD. However, any teacher regardless of years in the system, needing greater support and on a Plan of Improvement, as determined by the supervisor or principal is placed on Tier I. In Tier II, teachers in years four, five, and six or a veteran teacher new to the district are observed using the Tier II classroom observation instrument.

The observation instrument under examination included either 6 (Tier I) or 7 (Tier II) performance standards agreed on by the system and a content expert as covering the requirements of being a highly effective, professional educator in a standards-based classroom. The standards for the Tier I classroom observation instrument included *Curriculum and Planning (CP), Standards-Based Instruction (SBI), Assessment of Student Learning (ASL), Instruction Environment (IE), Building Positive Student Relationships (BPSR) and Artifacts and Evidence (AE).*

Tier II teachers are assessed on the same 6 standards as Tier I teachers, though there is an additional standard, *Teacher Leader (TL),* which asserts that as teachers becomes more seasoned veteran educators, they have a responsibility to assume additional responsibilities benefitting the school community and/or their peers. For this study, the *Artifacts and Evidence* and *Teacher Leader* standards were not examined, as they do not lend themselves to video observation analysis.

Within each of the performance standards, there are itemized elements (numbering from 1 to 9 amplifying the components of the standard). At the time of the study, the state-approved teacher evaluation system for each classroom observation, teachers received either a "yes" or a "no" on each individual element. The "yes" or "no" was given based on whether or not the rater observed the behavior outlined in the individual element. This information was gathered, and the observation forms were retained and used as a main source of data in calculating the final evaluation rating score each teacher receives at the conclusion of the school year. Examples of elements include "The teacher demonstrates high expectations for all students," "The teacher intentionally solicits feedback from students on their understanding of the standard," and "The teacher fosters a sense of community and belonging by acknowledging diversity, achievements, and/or accomplishments of learners in the classroom." This study ultimately gathered data on 25 such elements.

Teacher Observation Process in the Developmentally Appropriate School District

The observation process in the school district Teacher Evaluation System is similar for both Tier I and Tier II teachers. Namely, each teacher will have a pre-observation conference (though the pre-observation conference was strongly suggested for Tier II teachers), a classroom observation,

and a post-observation conference. The major difference between the tiers is the number of times each teacher goes through this process. In Tier I, the teacher was to be observed a minimum of three times, with as many as two being unannounced. In Tier II, there was no required pre-observation conference, but the evaluation system and its policies mandated post-observation conferences after each observation had concluded.

For Tier II teachers, the observation cycle occurs twice per year, with one observation being announced and one being unannounced. For all teachers, additional observation cycles may be performed at the discretion of the evaluator and may be either announced or unannounced. The need to establish inter-rater reliability for the school district teacher evaluation system observation instrument provided the ideal situation to examine inter-rater reliability not only as a total measure in a teacher observation context, but also how these coefficients may or may not vary along the teacher quality spectrum.

Data Sources

The data for this research come from a larger study designed to establish both the validity and reliability of a newly created teacher evaluation system. This study relies on data gathered from ratings of four videos of teachers teaching am actual lesson to students. Each of the four teachers in the videos and each administrator charged with rating the videos is an employee of the same school district. Inter-rater reliability calculation requires that "two or more observers have rated the same set of observable evidence" (Graham et al., 2012, p. 13), and this study is strengthened by having 42 raters.

Criteria were established for the selection of the videos from a total sample of 15 videos. The videos were chosen to cover all levels of schooling (e.g., elementary, middle, and high school), cover a range of academic subjects (e.g., mathematics, language arts, reading), be of an appropriate number of minutes of instruction (each video was at least 30 minutes so as to align with the system requirement that all classroom observations had to be a minimum of 30 minutes), and cover the teacher quality spectrum (i.e., poor quality teaching to high quality teaching). Each of the videos was rated by 42 principals and assistant principals in the school district.

Similarly, the principals and assistant principals rating the videos also covered the range of schooling as they came from all elementary, middle, and high schools within the district. It is important to note that the raters received training at monthly Principal Learning Communities (PLCs) on the instrument, the overall teacher evaluation system, and the accompanying rubrics and job aids to assist in the rating process for a two-year period prior to the study. The ratings were completed to adhere to a strict protocol that was explained to the participants in the study. The inter-rater reliability analyses included calculations of coefficients for all of the participating teachers combined, as well as analyses for each teacher, individually. All participants rated each video, and there were no missing data.

Data Analysis and Findings

This study, designed primarily to establish the inter-rater reliability of the observation instrument, resulted in two primary findings. The first finding from the study is that the inter-rater reliability coefficients for the instrument overall, may not be indicative of the effectiveness of the instrument at various levels of the teacher quality spectrum. Overall, the instrument was found to be adequately reliable for the purposes of the district. Each of the inter-rater reliability coefficients presented in this study ranges in value from 0 to 1 (Graham et al., 2012; Gwet, 2002). The interpretation of the statistics, however, can be difficult. In fact, the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) do not suggest a criterion for the interpretation of these measures, only that they should be calculated and reported. What is considered adequate varies and is ultimately a decision for the school district.

While the overall inter-rater reliability of the instrument was determined to be adequate by district administrators to continue the use of the instrument, further examination into these coefficients suggests a potential issue, one with broad impact for all districts and states evaluating teachers through observation instruments. Table 1 highlights the performance of the instrument using all teachers combined, which was used to establish the inter-rater reliability of the instrument.

Table 1: Inter-Rater Reliability Coefficients for all Items Combined for all Teachers Combined and the Highest and Lowest Rated Teacher

	% Agreement	Fleiss' Kappa	Gwet's AC1
All Teachers Combined	0.69	0.34	0.6
Highest Rated	0.84	0.08	0.83
Lowest Rated	0.52	0.16	0.33

Additionally, Table 1 also shows the performance of these coefficients calculated for the highest and lowest rated teachers in the study. While it can be argued that 0.69 as the overall reliability for the instrument is borderline, what is clear is that the value for the percent agreement for the lowest rated teacher of 0.52 is well-below what can be argued as acceptable. School districts endeavoring to create teacher evaluation instruments should be mindful that the adequacy of the overall inter-rater reliability coefficient may not be indicative of the adequacy of the coefficients associated with the instrument at various levels of the teacher quality spectrum; these coefficients may vary across the levels of teacher quality.

The second finding is the percent of overall agreement, values of kappa, and values for Gwet's AC1 statistic, are, by far, lowest for the lowest rated teacher, suggesting that there is likely a deficiency in the ability of trained raters to agree what poorer teacher quality looks like. As seen in Table 2, the value for the percent of overall agreement, when examining all the teachers

combined, is 0.69 for all standards combined, and ranges from 0.51-0.82 when examined by standard.

Table 2: Percent Agreement for All Four Videos Combined and the Highest and Lowest Rated Teachers

	All Teachers		Highest Rated		Lowest Rated	
	% Agree	S.E.	% Agree	S.E.	%Agree	S.E.
All Items	0.69	0.02	0.84	0.03	0.52	0.03
Standard 1	0.51	0.04	0.66	0.07	0.41	0.04
Standard 2	0.74	0.03	0.89	0.05	0.61	0.05
Standard 3	0.67	0.05	0.83	0.08	0.47	0.05
Standard 4	0.82	0.05	0.94	0.04	0.60	0.11
Standard 5	0.75	0.09	0.89	0.12	0.39	0.06

N=168

When examining the highest and lowest rated teacher individually, the value for the highest rated teacher for all standards combined is 0.84, with a standard range of 0.66-0.94, while the lowest rated teacher has a value of 0.52 for all standards combined and a range of 0.39-0.61 for individual standards. These results are similar for the values of kappa and Gwet's AC1 coefficients, as seen in Tables 3 and 4, respectively.

Table 3: Fleiss' Kappa for All Four Videos Combined and the Highest and Lowest Rated Teachers

	All Teachers		Highest Rated		Lowest Rated	
	Kappa	S.E.	Kappa	S.E.	Kappa	S.E.
All Items	0.34	0.03	0.08	0.02	0.16	0.05
Standard 1	0.18	0.03	0.01	0.01	0.01	0.02
Standard 2	0.40	0.04	0.09	0.04	0.22	0.15
Standard 3	0.24	0.06	0.06	0.03	0.13	0.06
Standard 4	0.51	0.10	0.02	0.02	0.08	0.06
Standard 5	0.23	0.09	0.03	0.05	0.05	0.01

N=42

Table 4: Gwet's AC1 for All Four Videos Combined and the Highest and Lowest Rated Teachers

	All Teachers		Highest Rated		Lowest Rated	
	AC1	S.E.	AC1	S.E.	AC1	S.E.
All Items	0.60	0.04	0.83	0.04	0.33	0.05
Standard 1	0.29	0.07	0.58	0.10	0.16	0.07
Standard 2	0.66	0.05	0.89	0.05	0.48	0.07
Standard 3	0.57	0.08	0.82	0.09	0.25	0.07
Standard 4	0.78	0.07	0.94	0.04	0.49	0.18
Standard 5	0.70	0.12	0.88	0.13	0.09	0.13

N=42

These findings suggest that raters may have a much keener sense of what represents high quality teaching, but a much less developed idea of what represents lower quality teaching. Seemingly, the raters agree when observing a high-quality teacher but have ratings with much greater variation when observing a poorer quality teacher.

Significance

The first finding that there is great variation in the inter-rater reliability coefficients along the teacher quality spectrum, suggests districts and states should examine inter-rater reliability not only in total, but also at the various levels of teacher quality and that the values at each of these levels should be considered when determining both the overall performance of the raters using the instrument and whether or not additional rater training or alteration to the observation instrument may be warranted. With the push to link teacher evaluation to income, creating instruments and providing adequate training to assure accurate assessment of teacher performance and continuity of ratings is vital. Those being rated have a right to know that the instrument's effectiveness is not just for those performing at the top end of the teacher quality spectrum.

The finding that there is greater variation in the ratings of the lowest rated teacher compared to the highest rated teacher is suggestive that policies and practices may need to be developed which assist in training raters to better identify poorer teacher quality. While rater training is likely a key area where the variation in ratings identifying poorer performing teachers can likely be minimized, there are other potential areas that can impact this variation. The standards and elements on the observation instrument, though validated for content, need further examination to determine if item wording makes it difficult to apply to teachers across the teacher quality spectrum. Finally, factors outside of rater training and instrument wording and construction can also impact the results of an evaluation. Some raters tend to "rate up" based on previous knowledge or out of a sense of niceness (Antonioni & Park, 2001). These are other areas of interest that need to be examined to help better accurately identify poorer quality teaching. Each of these factors potentially impacts the ability for raters to accurately identify poorer quality teaching, and in a policy and practice context, must be reconciled.

Limitations

This study was limited in two ways. First, the study population was limited to one school district in the southeastern US. Though there are several characteristics of the district that makes it interesting for research purposes, expanding the research population geographically, and in other ways, could improve the generalizability of the finding of the study. Another limitation is there were just four teachers assessed, purposefully chosen, due to the time available for assessing the videos. While four teachers provided plenty of data to establish inter-rater reliability coefficients, future research could benefit from more examples of teachers across the quality spectrum which could strengthen the findings, particularly at the lower end.

Conclusions

The findings from this study are important for both practice and policy. The need to have valid and reliable observation instruments is clear, especially in the age of accountability where the push to link salary, retention, and promotion is palpable. In the area of policy, these findings support the notion that policies should be developed to highlight the need to assess the reliability of an instrument at various levels of the teacher quality spectrum. If personnel and salary decisions are to be tied to teacher evaluations, all teachers should know that the observation instrument is reliable for teachers of all ability levels. In practice, findings highlight the need for high-quality, ongoing professional development for principals to better prepare them to be consistent and fair in the evaluation of teachers. Without fair and consistent teacher evaluation, there is no justification to tie teacher evaluation scores to such high-stakes decisions as continuous employment or merit pay.

Another noteworthy finding indicates that leaders can easily and accurately rate higher-performing teachers but the data suggests that there is less confidence in leaders rating teachers whose performance is less than "average," and to this, more research is needed as well as professional development that must prepare principals to not only be consistent and fair, but also to evaluate teachers in a similar way to other raters in the district or state. The raters participating in this research, as previously outlined, received extensive, ongoing professional learning on the classroom observation instrument, evaluation procedures, and rubrics. Given the amount of training and ongoing support, coupled with the longevity and stability of the school leaders, we are puzzled that the results yielded such low levels in the identification and amplification of poorly performing teachers. More research into the accurate identification of poorer performing teachers is needed to highlight the issues of accurate assessment of these teachers (Jimenez, & Zepeda, in press; Zepeda, 2017; Zepeda, 2016; Zepeda et al., 2014). Finally, this research suggests the overall IRR coefficient relating to an observation instrument may not be enough to determine if the instrument is in fact reliable. To determine the actual reliability of an instrument, assessment of teachers across the teacher quality performance spectrum might be needed and targeted training to assess teachers at the lower end of the teacher quality spectrum is likely warranted.

References

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448-452. https://doi.org/10.3102/0013189X15618385

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Recovery and Reinvestment Act (ARRA) of 2009, Pub. L. No. 111–5, 123 Stat. 115, 516 (Feb. 19, 2009).

Antonioni, D., & Park, H. (2001). The relationship between rater affect and three sources of 360-degree feedback ratings. *Journal of Management*, 27, 479-495.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. doi: 10.1037/h0026256

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378–387. https://doi.org/10.3102/0013189X16659442

Cogan, M. (1973). *Clinical supervision*. Boston, MA: Houghton Mifflin.

Cook, D.A., & Beckman, T.J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119 (2), 166.e7-16. doi: 10.1016/j.amjmed.2005.10.036

Clifford, M., & Ross, S. (2011). Designing principal evaluation: Research to guide decision-making. Washington, DC: National Association of Elementary School Principals.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement.* New York, NY: Teachers College Press.

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation?. *Educational Researcher, 44*(2), 132-137. https://doi.org/10.3102/0013189X15575346

Darling-Hammond, L. (2016). Research on teaching and teacher education and its influences on policy and practice. *Educational Researcher, 45*(2), 83–91. https://doi.org/10.3102/0013189X16639597

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. H., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8–15. doi:10.1177/003172171209300603

Derrington, M. L., & Campbell, J. W. (2018). High-stakes teacher evaluation policy: US principals' perspectives and variations in practice. *Teachers and Teaching: Theory and Practice, 24*(3), 246–263. https://doi.org/10.1080/13540602.2017.1421164.

Doherty, K. M., & Jacobs, S. (2015). *State of the States 2015: Evaluating teaching, leading and learning.* National Council on Teacher Quality. Retrieved from https://files.eric.ed.gov/fulltext/ED581451.pdf

Donaldson, M. L., & Papay, J. P. (2015). Teacher evaluation for accountability and development. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy* (pp. 174–193). New York, NY: Routledge

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*, 551-558. doi:10.1016/0895-4356(90)90159-M

Firestone, W. A. (2014). Teacher Evaluation Policy and Conflicting Theories of Motivation. *Educational Researcher, 43*(2), 100–107. https://doi.org/10.3102/0013189X14521864

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378. doi: 10.1037/h0031619

Garver, R. (2019). Evaluative relationships: teacher accountability and professional culture. *Journal of Education Policy*, 1-25. https://doi.org/10.1080/02680939.2019.1566972

Glanz, J., & Neville, R.F. (Eds.) (1997). *Educational supervision: Perspectives, issues, and controversies.* Norwood, MA: Christopher Gordon Publishers.

Goldhammer, R. (1969). *Clinical supervision: Special methods for the supervision of teachers*. New York, NY: Holt, Rinehart, & Winston.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Madison, WI: Center for Educator Compensation Reform. Retrieved from http://cecr.ed.gov/pdfs/Inter_Rater.pdf

Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-rater Reliability Assessment*, *1*(6), 1-6. Retrieved from http://www.agreestat.com/research_papers/kappa_statistic_is_not_satisfactory.pdf

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters*. Gaithersburg, MD: Advanced Analytics, LLC. Retrieved from http://www.agreestat.com/book3/bookexcerpts/first13pages.pdf

Haefele, D. L. (1993). Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education*. *7*(1), 21-31. doi: 10.1007/BF00972346

Hallgren, K., James-Burdumy, S., & Perez-Johnson, I. (2014). *State requirements for teacher evaluation policies promoted by Race to the Top. NCEE Evaluation Brief.* National Center for Education Evaluation and Regional Assistance. https://files.eric.ed.gov/fulltext/ED544794.pdf

Jimenez, A., & Zepeda, S. J. (2017). Building the plane in flight: Establishing post-hoc inter-rater reliability coefficients in an educational context. *Sage Research Methods Cases.* doi: http://dx.doi.org/10.4135/9781473958050

Jimenez, A., & Zepeda, S. J. (In Press). A Comparison of Gwet's AC1 and kappa when calculating inter-rater reliability coefficients in a teacher evaluation context. *Journal of Education Human Resources.*

Kauchak, D., Peterson, K., & Driscoll, A. (1985). An interview study of teachers attitudes toward teacher evaluation practices. *Journal of Research and Development in Education*, *19*(1), 32-37. Retrieved from http://www.researchgate.net/publication/232424773_An_interview_study_of_teachers'_attitudes_toward_teacher_evaluation_practices

Kraft M.A., & Gilmour A.F. (2017) Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234-249. https://doi.org/10.3102/0013189X17718797

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron and D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities, Ninety-fifth Yearbook of the National Society for the Study of Education.* Chicago, IL: University of Chicago Press.

Lohman, J. (2010). *Comparing no child left behind and race to the top* (2010-R0235). OLR Research Report. Retrieved from http://www.cga.ct.gov/2010/rpt/2010-R-0235.htm

McCarty, D., Kaufman, J., & Stafford, J. (1986). Supervision and evaluation: Two irreconcilable processes? *The Clearing House, 59*(8), 351-353. Retrieved from http://www.jstor.org/stable/30186568

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research, 80*(4), 242-247. doi: 10.2307/40539630

Mette, I. M., Range, B. G., Anderson, J., Hvidston, D. J., Nieuwenhuizen, L, & Doty, J. (2017). The wicked problem of the intersection between supervision and evaluation. *International Electronic Journal of Elementary Education, 9*(3), 709-724. Retrieved from https://www.iejee.com/index.php/IEJEE/article/view/185

National Council of Teachers of English. (2012). *NCTE position statement on teacher evaluation.* Retrieved from http://www.ncte.org/positions/statements/teacherevaluation

National Council of Teachers of Mathematics. (2011). *Teacher evaluation: A position of the National Council of Teachers of Mathematics*. Retrieved from http://www.nctm.org/uploadedFiles/About_NCTM/Position_Statements/Teacher%20Evaluation%20(with%20references,%202011).pdf

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Ponticell, J. A., Zepeda, S. J., Lanoue, P. D., Haines, J. G., Jimenez, A. M., & Ata, A. (2019). Observation, feedback, and reflection. In S. J. Zepeda & J. A. Ponticell (Eds.), *The Wiley Handbook of Educational Supervision.* (249-279*)* John Wiley & Sons, Inc.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340-359. https://doi.org/10.1162/EDFP_a_00186

Steinberg, M. P., & Quinn, R. (2017). Education reform in the post-NCLB era: Lessons learned for transforming urban public education. *Cityscape, 19*(1), 191-216. https://www.jstor.org/stable/26328306

Stodolsky, S. S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher, 13*(9), 11-18. doi: 10.3102/0013189X013009011

United States Congress. (2002). *No child left behind act of 2001*. US Congress. Retrieved from https://eric.ed.gov

U.S. Department of Education. (2009). *Race to the top program executive summary.* Retrieved from http://www.ed.gov/programs/racetothetop/executive-summary.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on teacher differences. Brooklyn, NY: The New Teacher Project. Retrieved from http://gcpstv.org/gcps-mainweb01.nsf/092DF14366B4598F8525788C0067CF48/$file/TNTPTheWidgetEffect.pdf

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: RAND. Retrieved from: http://www.rand.org/pubs/reports/2006/R3139.pdf

Zepeda, S. J. (2013). *The principal as instructional leader: A handbook for supervisors* (3rd ed.). New York, NY: Routledge.

Zepeda, S. J., Lanoue, P. D., Price, N. F., & Jimenez, A. M. (2014). Principal evaluation–Linking individual and building-level progress: Making the connections and embracing the tensions. *School Leadership & Management, 34*(4), 324-351. doi:10.1080/13632434.2014.928681

Zepeda, S.J. (2016). *The leaders guide to working with underperforming teachers: Overcoming marginal teaching and getting results*. New York, NY: Routledge.
Zepeda, S. J. (2017). *Instructional supervision: Applying tools and concepts* (4th ed.). New York, NY: Routledge.
Zepeda, S. J. (2019). *Professional development: What works* (3rd ed.). New York, NY: Routledge.

Author Biographies

Sally J. Zepeda is a professor of educational administration and policy at the University of Georgia (Athens). Her research centers on instructional supervision, teacher evaluation, and professional learning in the context of schools. Her book, *Instructional Supervision: Applying Tools and Concepts*, is in its fourth edition and was simultaneously translated into Turkish. She was the lead co-editor for the *Wiley Handbook on Educational Supervision* (2019).

Albert M. Jimenez is associate professor of educational leadership at Kennesaw State University. His research focuses primarily on teacher evaluation, the improvement of evaluation techniques, and improving school climate through leadership. He currently is on the advisory board of Bridging Theory and Practice: The Rowman and Littlefield Leadership Series and has an upcoming publication in the inaugural issue of the *Journal of Education Human Resources*.