Journal of Educational Supervision

Volume 5 | Issue 1 Article 2

2022

Reliability Evidence for the NC Teacher Evaluation Process Using a Variety of Indicators of Inter-Rater Agreement

T. Scott Holcomb
University of North Carolina at Charlotte, tholcom4@uncc.edu

Richard Lambert
University of North Carolina at Charlotte, rglamber@uncc.edu

Bryndle L. Bottoms
University of North Carolina at Charlotte, bbottom3@uncc.edu

Recommended Citation

Holcomb, T. S., Lambert, R., & Bottoms, B. L. (2022). Reliability Evidence for the NC Teacher Evaluation Process Using a Variety of Indicators of Inter-Rater Agreement. *Journal of Educational Supervision, 5* (1). https://doi.org/10.31045/jes.5.1.2

Empirical Research

Reliability Evidence for the NC Teacher Evaluation Process Using a Variety of Indicators of Inter-Rater Agreement

Journal of Educational Supervision
27 – 43
Volume 5, Issue 1, 2022
DOI: https://doi.org/10.31045/jes.5.1.2
https://digitalcommons.library.umaine.edu/jes/

T. Scott Holcomb[1], Richard Lambert[1], and Bryndle L. Bottoms[1]

Abstract

In this study, various statistical indexes of agreement were calculated using empirical data from a group of evaluators (n = 45) of early childhood teachers. The group of evaluators rated ten fictitious teacher profiles using the North Carolina Teacher Evaluation Process (NCTEP) rubric. The exact and adjacent agreement percentages were calculated for the group of evaluators. Kappa, weighted Kappa, Gwet's AC1, Gwet's AC2, and ICCs were used to interpret the level of agreement between the group of raters and a panel of expert raters. Similar to previous studies, Kappa statistics were low in the presence of high levels of agreement. Weighted Kappa and Gwet's AC1 were less conservative than Kappa values. Gwet's AC2 statistic was not defined for most evaluators, as there was an issue found with the statistic when raters do not use each category on the rating scale a minimum number of times. Overall, summary statistics for exact agreement were 68.7% and 87.6% for adjacent agreement across 2,250 ratings (45 evaluators ratings of ten profiles across five NCTEP Standards). Inter-rater agreement coefficients varied from .486 for Kappa, .563 for Gwet's AC1, .667 for weighted Kappa, and .706 for Gwet's AC2. While each statistic yielded different results for the same data, the inter-rater reliability of evaluators of early childhood teachers was acceptable or higher for the majority of this group of raters when described with summary statistics and using precise measures of inter-rater reliability.

Keywords

inter-rater reliability; teacher evaluation; agreement coefficients; in-service teachers

[1] University of North Carolina at Charlotte

Corresponding Author:
T. Scott Holcomb (Cato College of Education, University of North Carolina at Charlotte, 9021 University City Blvd, Charlotte, NC, USA, 28223)
email: tholcom4@uncc.edu

Introduction

Evidence from teacher observation systems has received a lot of focus and attention over the last decade, and has been examined in relation to student achievement, educational policy, and school funding (Hill et al., 2012; James & Wyckoff, 2020; MET Project, 2010; Ross & Walsh, 2019; Weisberg et al., 2009). In North Carolina alone, over 100,000 teachers are evaluated each year using the North Carolina Teacher Evaluation Process (NCTEP) (U.S. Department of Education, 2021). Evaluators of teachers in North Carolina are required to have some form of observation training, but there is not a required certification of inter-rater reliability (IRR) in place (National Council of Teacher Quality, 2019). However, this issue is not unique to North Carolina. With over three million public school teachers across the United States, and almost all states using some form of observational data to evaluate teachers, it is important that teachers are evaluated with valid, reliable, and fair measures of their professional practice (U.S. Department of Education, 2021). Despite the widespread use of observations to assess teacher performance, important empirical concepts related to the validity and reliability of evaluation scores from observations of teachers have been ignored (Cohen & Goldhaber, 2016; Herlihy et al., 2014; James & Wyckoff, 2020).

Information from teacher evaluations can be used to provide formative feedback to help teachers grow their practice and direct a teacher's professional development plan for improvement. These are two major factors in maintaining the best possible teacher workforce (Adnot et al., 2017; Herlihy et al., 2014; Hill & Herlihy, 2011). A required evaluator (or rater) certification and recertification process is among the suggestions offered as necessary pieces supporting the validity and reliability of teacher evaluation scores (Zepeda & Jimenez, 2019). An essential part of periodic training and recertification of raters of teachers involves investigation of IRR. This training and certification process looks differently across grade levels, subjects, and states. In some cases, raters co-rate lessons during observations with a certified rater and compare scores. However, in many states the decision regarding who observes and how the rater is credentialed to conduct observations is left up to a local school district. Regardless of these decisions most states and local districts do not attend to multiple issues related to teacher evaluation systems. Almost all states omit calculating IRR rates and statistics as a measure of reliability of scores produced by teacher evaluators (Herlihy et al., 2014).

The analysis and interpretation of rating quality beyond agreement percentages are necessary parts of teacher evaluation systems. The rating decisions made by teacher evaluators play a critical role in determining the effectiveness of a teacher. The use of chance-corrected agreement coefficients supplements efforts to provide valid and reliable scores of teacher performance. The ratings a teacher receives should not be dependent on which evaluator conducts the observations. Teacher performance evaluation ratings should be invariant to rater effects. In this study, a real-world training and IRR certification process of 45 external evaluators of early childhood educators provided evidence supporting the use of chance-corrected agreement coefficients as part of the reliability process of teacher observation scores.

Literature Review

Validity addresses the extent to which an interpretation of a test score is supported by the proposed test. The process of validation involves building an accumulation of evidence to support the basis for proposed score evidence. Interpretations or applications of scores "in a high-stakes environment is vulnerable to many validity threats, such as inadequate construct definition, construct underrepresentation, illogical reasoning…, negative consequences of test score use, and low reliability of test scores" (Haladyna & Downing, 2004, p. 25). Reliability is a pre-condition for validity, making it important to address inter-rater reliability. In order for classroom observation scores to be valid, raters must understand the instrument in use and the instrument in use must measure the construct it claims to be measuring. In addition, raters of teachers must be able to provide scores that are accurate and consistent (White, 2018). In order for teacher evaluation systems to meet this criteria, raters of teachers must be carefully trained and monitored (Bell et al., 2012; White, 2018).

While other studies related to teacher evaluation utilize generalizability theory, the primary interest of the current study is the application and use of IRR coefficients. Teacher evaluation studies reporting IRR statistics frequently stop at reporting percentage agreement between raters (Casabianca et al., 2015; Hill et al., 2012; Sartain et al., 2010). The *Standards for Educational and Psychological Testing* do not provide suggestions for a specific agreement level or reliability measure, but recommend appropriate measures are reported and calculated while an assessment is in use (American Educational Research Association [AERA] et al., 2014; Graham et al., 2012). As Hill et al. (2012) suggested, percentage agreement figures could be overstated through simply having less rating points on the observation instrument.

Construct Irrelevant Variance

As stated in *Standards for Educational and Psychological Testing*, construct irrelevant variance refers to the amount "scores may be systematically influenced to some extent by processes that are not part of the construct" (AERA et al., 2014, p. 13). Construct irrelevant variance can negatively impact the quality of teacher evaluation scores. Among many factors related to the existence of variance in teacher evaluation systems, some primary factors contributing to construct irrelevant variance can include the lesson observed, the rater, and the observational instrument in use (Hill et al., 2012). Some studies have attempted to address this issue through the use of multiple raters. In the case of a teacher evaluation system, construct-irrelevant variance may be added to the situation according to the time of day the observation occurs, the subject the lesson the teacher is focused on, the group of students the teacher is working with during that particular lesson and/or subject, etc. The list of factors in the area of construct-irrelevance involved with teacher evaluation systems and classroom observation is countless.

Rater-Mediated Assessment

For teacher evaluation systems based on observational data, the item responses consist of placements on rating scales made by evaluators. These placements result from a series of decisions made by each evaluator and inconsistencies between raters can be common. Wilson (2004) listed several reasons inconsistencies may occur: raters may never apply the scoring guide

in a correct way due to training differences, there may be differences in rater severity, raters often have natural tendencies to use rating categories more or less frequently, "halo effects", rater drift, and raters demonstrating inconsistencies themselves for a variety of reasons. Rater-mediated assessments are often presented with complex situations, such as in teacher evaluation processes. Raters of teacher performance must carefully and skillfully provide scores for an observation, no matter if the use of the score is summative to make a high-stakes decision or formative to provide feedback to the teacher.

Rater Accuracy, Agreement, and Consistency

In teacher evaluation systems, rater accuracy refers to the ability of a rater to provide accurate scores from an observation against a set of ratings provided by an expert panel or master rater. It is typical in calibration training of raters that passing a rating certification training process involves raters being able to provide accurate scores on a given teacher evaluation instrument (Cash et al., 2012; Hill et al. 2012). Whether accuracy is of the most interest in a given rating of a teacher observation is dependent upon the purpose of the observation. In situations where the observations can have high stakes for teachers, a rater's ability to provide accurate ratings of a teacher's performance is critical.

Research on rater agreement has utilized various methods for examining the consistency of ratings in performance assessment. One way to ascertain the consistency of a rater is to produce an index of the proportion of ratings of exact agreement. There are methods that can be developed related to adjacent agreement that can be adjusted based on the rating scale used for the construct of interest. Using adjacent agreement calculations can produce overly positive results, especially in the case where there are a small number of rating categories available. One step beyond these exact or adjacent percentage agreement measures involve using IRR coefficients that adjust scores based on chance-agreement.

To address issues of rater inconsistency it is important to have a process for training raters and a monitoring system that tracks the consistency of raters over time (Wilson, 2004). Wilson (2004) recommended five components to include in rater-mediated assessment training programs. Raters should have:

1. Understanding of the assessment or construct.
2. Opportunity to examine a large, representative sample of responses from the construct of interest.
3. Opportunity to have cognitive discussions with other raters on overlapping work.
4. Feedback provided to raters centering on how well they rate responses.
5. A system of rater calibration steps that result in raters passing training or having a need for further support.

Additionally, Wilson (2004) recommended a pre-developed monitoring system that could involve co-observations or re-ratings done by experts over a sample of a caseload. Using reference ratings of some sort allows for raters to see how consistent a sample of their ratings have been over time.

Teacher Evaluation

In educational research and evaluation, there is a strong emphasis placed on the reliability and validity of student achievement outcomes. These outcomes are even used as part of the teacher evaluation process in some states. However, there is not a similar focus on the validity, reliability, and fairness of teacher evaluation data. It is common for teacher evaluation systems to have unclear, or even absent, requirements related to the validity and reliability of scores from these observational instruments altogether (Herlihy et al., 2014). Across studies it was demonstrated that what makes scores from a teacher evaluation system valid and reliable is heavily context-specific (Cohen & Goldhaber, 2016; Herlihy et al., 2014).

Evidence exists to support the claim that raters are the largest source of error in evaluation systems (Casabianca et al., 2013; Cohen & Goldhaber, 2016; Hill et al., 2012). However, in all of these studies the rater is almost exclusively a school-level administrator. Some school districts and states require the use of an external evaluator or master rater during at least one observation (Adnot et al., 2017; Herlihy et al., 2014). Logically, this makes sense given the myriad responsibilities managed by administrators. Evaluation systems can be content and context specific, making the process of using an observation instrument complex and requiring a level of expertise to minimize subjectivity and enhance reliability of scores. These are measures suggested to enhance the overall quality of placements made across components of a given observation instrument. Steinberg & Sartain (2015) suggested the use of highly trained raters to conduct observations led to greater improvement in overall teacher quality. Other reports and studies debated who should be conducting observations and the best route to effectively train raters (Dee & Wyckoff, 2015; Hill & Herlihy, 2011; Sartain et al., 2009). While this is not the aim and scope of the current study, addressing who conducts teacher observations is something that needs further exploration and is supported by this study as it relates to a specific group of evaluators of early childhood educators.

Study Design

The Early Educator Support Office supports and evaluates preschool teachers holding birth through kindergarten licensure in North Carolina. The office uses the North Carolina Teacher Evaluation Process (NCTEP) rubric to perform formative and summative evaluations with designated teachers. This is the same rubric that is used to evaluate all licensed teachers in pre-kindergarten through grade 12 across North Carolina. The NCTEP rubric consists of five standards: Standard I – Teachers Demonstrate Leadership, Standard II – Teachers Establish a Respectful Environment for a Diverse Population of Students, Standard III – Teachers Know the Content they Teach, Standard IV – Teachers Facilitate Learning for their Students, and Standard V – Teachers Reflect on their Practice. The standards include a total of 25 elements that apply to specific components of teaching and allow for more targeted feedback. The performance rating scale has four levels and a not demonstrated option (see Table 1) applicable to all five standards and 25 elements.

Table 1
NCTEP Performance Rating Scale

Rating	Description
Developing	Teacher demonstrated adequate growth toward achieving standard(s) during the period of performance but did not demonstrate competence on standard(s) of performance.
Proficient	Teacher demonstrated basic competence on standard(s) of performance.
Accomplished	Teacher exceeded basic competence on standard(s) of performance most of the time.
Distinguished	Teacher consistently and significantly exceeded basic competence on standard(s) of performance.
Not Demonstrated	Teacher did not demonstrate competence on or adequate growth toward achieving standard(s) of performance. (If selected the evaluator must comment why it was used).

While the state of North Carolina currently does not have any required IRR protocols among evaluators who use the NCTEP, the Early Educator Support Office has implemented a systematic approach to evaluate the IRR of their evaluators. The "Phased Quality Improvement Model" spans eight phases to prepare and support the group of teacher evaluators responsible for evaluating licensed preschool teachers (Lambert, et al., 2021). The current study is situated within Phase VIII of the model: the IRR certification process. Early Educator Support Office evaluators with at least one year of experience rated ten fictitious teacher profiles containing videos, narratives, and classroom artifacts from real classrooms to represent a typical caseload of teachers served by the evaluators (Lambert, et al., 2021). This study is one piece of the IRR certification process and part of developing an IRR system for the Early Educator Support Office.

IRR is a measure of internal consistency that evaluates the level of agreement among raters. IRR can be utilized to include an agreement with a "correct" rating or agreement between raters. In some applications exact and adjacent agreement methods are used to identify rating consistency among raters (Graham et al., 2012; Johnson et al., 2009; Wind & Engelhard, 2012). Rater agreement is important in teacher evaluation, Kappa is a widely used method for assessing IRR. However, there are well documented statistical problems associated with this measure (Uebersax, 2002; Xie, 2013). Namely, a low value of the Kappa coefficient occurs in the presence of high-agreement and the Kappa value is heavily dependent upon the marginal distribution of ratings (Blood & Spratt, 2007; Xie, 2013). In order to assess its utility in our example, we evaluated it against a weighted Kappa, Gwet's AC1, and Gwet's AC2 for exact ratings and compared the results. Some of these widely used measures of IRR tend to be overly conservative (Porter & Jelinek, 2011; Rui & Feldman, 2012; Walsh et al., 2014). This is true of other measures of inter-rater agreement relying on the marginal distribution of categories, such as Intra-Class Correlations (ICC). In many cases even though actual rater agreement is high, the IRR statistic may be low. This could lead to useful information about teacher evaluators being mislabeled as unreliable. There is little empirical evidence directly addressing the validity,

consistency, and accuracy of evaluator ratings of early childhood teachers (Lambert,et al., 2021). Likewise, studies concerning the process of teacher evaluations and the use of evaluation data to develop early childhood teacher performance is rare.

Objectives and Research Questions

The aim of this study was to measure inter-rater agreement of standards ratings of a group of early childhood education evaluators and to compare methods for doing so. There are long-standing and well documented problems with the use of the Kappa coefficient as a measure of chance-corrected agreement between raters. In the area of teacher evaluation, it is important to make sure information from raters of teachers is reliable and valid. This made it necessary to further investigate other methods for calculating agreement coefficients. The following research questions guided the study:

1. How robust is Kappa when handling professional teacher evaluator data with low prevalence categories and high percentage agreement?
2. How well do alternative chance corrected agreement coefficients handle professional teacher evaluator data with low prevalence categories and high percentage agreement?
3. What are the reliability levels among evaluators across fictitious teacher profiles using ICC methods?

Methods

The overall exact percentage agreement was calculated for each of the 45 raters. Additionally, an adjacent percentage agreement for each rater was calculated (Lambert, Holcomb, & Bottoms, 2021). In this study, adjacent ratings were counted if the correct answer and rating were in the middle two categories, proficient and accomplished. Exact or adjacent percentage agreement of ratings may be an acceptable measure of agreement in some purposes; however, this does not account for agreement from chance alone. Methods vary for approaches to account for agreement by chance for continuous, ordinal, categorical, and nominal data. To investigate if general conclusions from percentage agreement statistics hold up, further analysis using chance-corrected agreement coefficients were calculated. The simple Kappa assumes two unique raters. This study applied methods for many raters. Overall agreement, adjacent agreement, Kappa, weighted Kappa, Gwet's AC1, and Gwet's AC2 were assessed in terms of overall standard ratings by 45 evaluators of ten fictitious teacher profiles. There is not a unified explanation of how to interpret agreement with IRR coefficients. Landis and Koch (1977), Fleiss (1981), and

Table 2

Benchmark Agreement Levels for Chance Corrected Agreement Indices

	< 0.00	0.00 – 0.20	0.21 – 0.40	0.41 – 0.60	0.61 – 0.80	0.81 – 1.00
Landis & Koch (1977)	No agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
Fleiss (1981)	Poor	Poor	Poor	Fair	Good	Excellent
Altman (1991)	Poor	Poor	Fair	Moderate	Good	Very Good

Altman (1991) provided suggestions for how to interpret agreement with the Kappa statistic (see Table 2).

The formulas and further explanation of how each of these methods were calculated are presented below. Cohen's (1960) Kappa was calculated as seen in Equation 1:

$$\hat{k} = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o is the observed proportion of agreement and p_e is the proportion of agreement by chance (Fleiss, 1981). This statistic estimates chance agreement through assuming ratings from different raters are completely random.

The weighted Kappa coefficient was defined as:

$$\hat{k}_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}} \tag{2}$$

Weighted Kappa considers predefined weights measuring the degree of disagreement. which distinguishes it from Kappa. In this study, weights on the main diagonal of the symmetrical matrix were equal to zero with all values off the main diagonal equal to one.

Gwet's AC1 and AC2 statistics are useful in situations with high levels of exact agreement (Gwet, 2008). Gwet's AC1 is "the conditional probability that two randomly selected raters might agree given that there is no agreement by chance" (Gwet, 2010). Gwet's AC1 was designed to use with any number of raters using categorical rating systems. Formulas used to calculate AC1 are shown in Equations 3 – 6 below:

$$\pi_q = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{iq}}{r} \tag{3}$$

$$p_{e\gamma} = \frac{1}{Q-1} \sum_{q=1}^{Q} \pi_q (1 - \pi_q) \tag{4}$$

$$p_a = \frac{1}{n} \sum_{i=1}^{n} [\sum_{q=1}^{Q} \frac{r_{iq}(r_{iq}-1)}{r(r-1)}] \tag{5}$$

$$AC1 = \frac{p_a - p_{e\gamma}}{1 - p_{e\gamma}} \tag{6}$$

Both overall agreement probability and chance agreement probability were estimated for AC1. The initial equation (solving for π_q) is used to calculate the probability that a rater classifies an object into a specific category, where p_a is the overall agreement probability and p_{ey} is the proportion of agreement by chance considering a random rating. Gwet's AC2 is a step beyond the Kappa statistic and Gwet's AC1. Gwet's AC2 can be adjusted for chance agreement and

misclassification errors. This is a method involving weighting disagreements in ratings differently. The meaning of the letters and symbols are the same across AC1 and AC2 formulas. However, the calculation of the values differs in some instances. Equations 7 – 12 were used to calculate AC2:

$$\pi_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{r_{il}}{r} \tag{7}$$

$$\alpha_{a|ql} = \sum_{k=1}^{Q} \beta_{k|q}\beta_{k|l} \tag{8}$$

$$p'_{e\gamma} = \frac{1}{Q-1} \sum_{q=1}^{0} \pi'_q\left(1 - \pi'_q\right) \tag{9}$$

$$\pi'_q = \sum_{l=1}^{Q} \beta_{q|l}\pi_l \tag{10}$$

$$p'_a = \frac{1}{n} \sum_{i=1}^{n} [\sum_{q=1}^{Q} \alpha_{a|qq} \frac{r_{iq}(r_{iq}-1)}{r(r-1)} + \sum_{q\neq 1}^{Q} \sum_{q\neq 1}^{Q} \alpha_{a|ql} \frac{r_{iq}r_{il}}{r(r-1)}] \tag{11}$$

$$AC2 = \frac{p'_a - p'_{e\gamma}}{1 - p'_{e\gamma}} \tag{12}$$

Data Sources

This study involved 45 evaluators of North Carolina early childhood educators. The group of evaluators rated ten fictitious teacher profiles using the NCTEP rubric. The profiles were developed by a panel of experts to represent a broad cross section of possible teachers the evaluators could evaluate in the field (Bottoms et al., 2021). All teachers were rated by all evaluators. The raters were not randomly selected, they were the group of evaluators employed by two specific programs within North Carolina during the study year. These raters can be conceptualized as the sample of all possible raters within the state in this context. Evaluators scored teachers across five overall standards and 25 elements. This study utilizes the five overall standard ratings to calculate the IRR methods. A total of 2,250 ratings using the teacher performance rubric were included in the study.

Results

Overall, the 45 evaluators had an exact agreement of 68.7% (*SD* = 10.2). Exact agreement percentages for the individual evaluators ranged from 42% to 88%. There was an overall agreement of 87.6% (*SD* = 7.0) using the adjacent agreement method. The evaluators ranged from 64% to 100% agreement using this method. Kappa values ranged from 0.121 to .790, with the overall sample Kappa at 0.486 (*SD* = .160). Ten raters had agreement levels within the

"good" or "substantial" .61 to .80 agreement range. There were 19 raters with an agreement level between .41 and .60, described as moderate agreement by Altman (1991) and Landis & Koch (1977).

Table 3
Exact and Adjacent Percent Agreements

	Agree	Lenient	Strict	Adjacent	AdjLen	AdjStr
Mean	68.7%	7.7%	23.6%	87.6%	2.4%	10.0%
Std. Deviation	10.2%	6.5%	12.0%	7.0%	2.8%	7.4%
Minimum	42.0%	0.0%	4.0%	64.0%	0.0%	0.0%
25th percentile	62.0%	3.0%	13.0%	84.0%	0.0%	4.0%
50th percentile	68.0%	6.0%	24.0%	88.0%	2.0%	10.0%
75th percentile	76.0%	12.0%	31.0%	93.0%	5.0%	14.0%
Maximum	88.0%	26.0%	58.0%	100.0%	10.0%	36.0%

Similar to previous studies applying the use of Kappa and weighted Kappa, the weighted Kappa produced a higher coefficient (Rui & Feldman, 2012). The weighted Kappa was .667 (*SD* = .143) for the full sample, individual evaluator values ranged from .315 to 1.000. The total amount of raters in the .61 to .80 agreement range was 23 for weighted Kappa. There were eight evaluators with weighted Kappa values greater than .80. Gwet's AC1 statistic was .563 (*SD* = .143) for the overall sample, with a range of .167 to .832. Tables 3 and 4 show results according to each agreement coefficient. Gwet's AC2 could only be calculated using all four rating scale categories for six out of 45 evaluators, for the other raters the statistic was calculated using three categories, reducing AC2 to the adjacent AC1 statistic. Gwet's AC2 statistic requires a rating category to have been used as a rating at a minimum of two times by the evaluator to be included in the chance corrected agreement calculation. Evaluators correctly did not use this rating category, as none of the overall standard ratings for the ten profiles did not have a rating of "distinguished" for any profile.

Table 4
Overall Inter-rater Reliability Between Raters

	Kappa	Weighted Kappa	AC1	AC2
Mean	0.486	0.667	0.563	0.706
Std. Deviation	0.160	0.143	0.143	0.109
Minimum	0.121	0.315	0.167	0.372
Maximum	0.790	1.000	0.832	0.886

ICCs by Standard

This study had 10 subjects to be rated. In this context, these were the 10 teacher profiles. While the profiles were not strictly randomly selected, they were developed to be representative of a broader population of all possible teachers an evaluator could encounter in the field. In this context, raters were the 45 evaluators who participated in the study. While these raters were not strictly randomly selected, they were simply the evaluators that worked for two specific programs within the state of North Carolina during the study year and could be conceptualized as a sample of all possible raters in the state. All subjects were rated by all raters. This study had five items or rating scales. In this context, these were the ratings on the rating scale progressions for each standard, and all five items had the same scaling. We chose to analyze each standard separately. This way of conceptualizing the reliability problem indicates a two-way random effects model with a single score, and absolute agreement between raters as the outcome of interest. For this model, the ICC can be defined this way:

$$ICC = \sigma^2_r / (\sigma^2_r + \sigma^2_p + \sigma^2_e) \quad (13)$$

where:

σ^2_p = Between-profile variance

σ^2_r = Between-rater variance

σ^2_e = Within-rater residual variance.

In HLM notation:

$$ICC = \sigma^2_{u0} / (\sigma^2_{\gamma} + \sigma^2_{u0} + \sigma^2_e) \quad (14)$$

where:

σ^2_{γ} = Between-profile variance

σ^2_{u0} = Between-rater variance

σ^2_e = Within-rater residual variance.

$$\text{Rating}_{mj} = \gamma_{00} + \gamma_{10}*P2_{mj} + \gamma_{20}*P3_{mj} + \gamma_{30}*P4_{mj} + \gamma_{40}*P5_{mj} + \gamma_{50}*P6_{mj} + \gamma_{60}*P7_{mj} + \gamma_{70}*P8_{mj} + \gamma_{80}*P9_{mj} + \gamma_{90}*P10_{mj} + u_{0j} + e_{mj} \quad (15)$$

where:

$_j$ = Rater number

$_m$ = Profile number

P = Profile

In our application, the within-rater residual variance was equivalent to the rater by profile interaction term. In cases where each rater made more than one rating for each profile, then this term would be the sum of the rater by profile interaction and the residual variance term. This approach is equivalent to equation 15 in Fleiss and Cohen (1973), and to ICC (2,1) in Shrout and Fleiss (1979). We estimated the ICCs using the ordinary least squares method through SPSS and using the restricted maximum likelihood (RML) method using the HLM software. Table 5 shows that the RML method resulted in similar but slightly more conservative estimates of the ICCs. The ICC values ranged from .521 to .729 and represent moderate to substantial levels of agreement (Landis & Koch, 1977). The ICCs differed greatly from Kappa values in the exact agreement method. Compared to the results from Kappa coefficients, the ICCs are very similar following the modified within one scoring method.

Table 5
ICC values

	Exact		Within One	
Standard	ICC (HLM)	ICC (SPSS)	ICC (HLM)	ICC (SPSS)
1 - Leadership	0.523	0.549	0.550	0.576
2 - Respectful Environment	0.555	0.581	0.507	0.533
3 - Content	0.669	0.692	0.639	0.662
4 - Facilitate Learning	0.643	0.667	0.708	0.729
5 - Reflective Practice	0.521	0.547	0.584	0.609

Discussion

This study was intended to be one step in the phases of developing a systematic approach to IRR for the NC Early Educator Support Office and contributed to the process by ensuring evaluators are providing consistent, accurate, and reliable ratings of teachers. This study contributed to the process of knowing which evaluators are receiving adequate training and which evaluators may need more focused support. Through examining IRR coefficients this group of evaluators of early childhood teachers are mostly providing reliable, valid, and fair ratings of teachers using the NCTEP rubric. In turn, this leads to teachers of young children being provided with quality and accurate support so that teachers can provide learning experiences to students that contribute to their growth and development. This protocol helps to achieve the goal of having an organized and communicable process in place to address best practices for this group of evaluators and the teachers they serve.

When using teacher evaluation instruments, it is difficult to determine where to attribute difference in ratings, whether it is teacher practices, the rater, or the classroom conditions. With studies demonstrating the majority of variance between ratings of teachers being attributed to the rater, the process of how raters are trained, certified, and monitored is essential to building a

valid and reliable system of teacher evaluation (Herlihy et al., 2014). While evaluation scores can be expected to vary across different applications of a given observational instrument, in this investigation of IRR through a training exercise most Early Educator Support Office evaluators provided reliable ratings of classroom teachers. This leads to higher quality teachers which positively influences young children's learning and development. In this study, there were reasonable levels of exact agreement for the majority of raters. Most rating disagreements occurred where the high-stakes decision would not have been present for the evaluated teachers. For the most part rater disagreement occurred when deciding between the middle rating categories of "proficient" and "accomplished". The difference between these two ratings does not impact high-stakes teacher licensure decisions made by evaluators, nor does it affect the type of professional development or mentorship provided to the teacher. The current study confirmed previous findings on the shortcomings of Kappa and the robustness of Gwet's AC1 at addressing the overly strict measures of chance agreement (Zepeda & Jimenez, 2019). These results also demonstrated Gwet's AC2 has shortcomings in a real-world example using a group of highly trained raters. Trained teacher evaluators do not appear to make random guesses, rather they follow a more complex process through interpreting evidences during an observation. Further development of a coefficient of rater-mediated agreement using highly trained raters with complex response processes is needed.

While the results of the group of evaluators were not in perfect agreement with the expert panel's ratings, they were acceptable for the majority of individual evaluators according to any of the IRR measures. As a result of findings from this study, and other parts of the Phased Quality Improvement Model, evaluators received targeted coaching, support, and professional development. Evaluators were involved in individual conferences going over their performance and were given a chance to discuss their decisions behind ratings as needed. For some evaluators this involved discussing ratings across specific standards, elements, and/or profiles. One area explored to target the development of evaluators demonstrating specific needs through the results of this study included conducting joint observations with a lead evaluator. Some evaluators will receive additional training on specific components of the NCTEP rubric and go through another IRR recertification process.

This study provides much needed evidence regarding reliability coefficients from one sample of evaluators of early childhood teachers who hold birth through kindergarten licensure. Most research on teacher evaluation focuses on evaluations of preservice teachers or evaluations conducted by school administrators. Research on observation scores of preservice teachers frequently occurs due to the need to effectively prepare preservice teachers and researchers have access to preservice teachers and their observation results in teacher education programs. Likewise, the majority of states confine the task of conducting teacher evaluations to classroom observations by school administrators. Administrators have numerous duties and may not be able to exclusively focus on teacher evaluation in the same manner as an external evaluator. The current study focused exclusively on the reliability of teacher evaluation scores of licensed teachers by a group of external professional evaluators.

Limitations and Future Directions

All raters included in this study had at least one year of experience in the field prior to participating in the IRR activity. This criterion emphasized the need for well-trained and experienced evaluators of early childhood teachers. The ten fictitious profiles were developed using video-recorded lessons, photographs, and artifacts from actual teachers served by the Early Educator Support Office. The profiles included written descriptions of classroom scenarios and lessons. While these profiles were designed with authentic artifacts from early childhood classrooms, they were still fictitious. However, the profiles allowed for all evaluators to participate in the same classroom scenario at the same time, this would be too cumbersome to complete authentically in overlapping classrooms.

As mentioned, these data are part of an ongoing process of developing a systematic approach to evaluating early childhood educators (Bottoms et al., 2021). Analyzing IRR agreement coefficients addresses just one part of the problem. Next steps include ensuring the items on the NCTEP rubric come together as a single measure of global teacher quality. This would provide further evidence the measure being used provides valid and reliable scores. A study utilizing a Many Facets Rasch Model should be conducted using the full set of fictitious teacher profile ratings, actual classroom ratings, and scores from evaluators currently undergoing IRR training and certification. The facets study would involve calibrating rater strictness and leniency, addressing racial bias, and administrative data (Bottoms, 2022). The problems related to traditional coefficients of IRR are well identified in the literature and emerged in this study using real-world data from evaluators of early childhood teachers. This underscores the need for further development of new methods more suitable for handling the complexities of rater-mediated assessment data.

Conclusion

As states continue to alter, monitor, and improve current teacher evaluation systems they should focus on empirical findings and establish rater-mediated assessment processes that yield valid and reliable data to strengthen their evaluation systems. Ensuring teacher evaluations are based on valid and reliable data is essential in providing all students access to effective teachers. This study provided an initial step in determining whether this group of evaluators scores teachers accurately and consistently, however it is important to continue to score raters over time as part of a monitoring and rater certification process. The high-stakes decisions behind measures of teacher performance require careful examination of reliability of ratings provided by teacher evaluators. Results from this study demonstrated a group of early childhood teacher evaluators provide acceptable levels of reliability. However, when more precise measurements were applied there was greater variability in scores produced by raters.

References

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis, 39*(1), 54-76. https://doi.org/10.3102/0162373716663646

Altman, D. (1991). *Practical statistics for medical research*. CRC Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2-3), 62-87. https://doi.org/10.1080/10627197.2012.715014

Blood, E., & Spratt, K. F. (2007). Disagreement and agreement: Two alternative agreement coefficients. SAS Global Forum: paper 186-2007, 1-12.

Bottoms, B. L. (2022). *A systematic approach to interrater reliability in teacher performance evaluations* [Doctoral dissertation, University of North Carolina at Charlotte]. ProQuest Dissertations and Theses Global.

Bottoms, B. L., Holcomb, T. S., Lambert, R. G., & Vestal, A. R. (2021). *The development of a systematic approach to evaluating early childhood teachers using the North Carolina teacher evaluation process* (CEMETR-2021-12). University of North Carolina at Charlotte, Center for Educational Measurement and Evaluation.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337. https://doi.org/10.1177/0013164414539163

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542. https://doi.org/10.1016/j.ecresq.2011.12.006

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.117/001316446002000104

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

Dee, T.S, & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management 34*(2), 267-297. https://doi.org/10.1002/pam.21818

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). John Wiley.

Fleiss, J.L., & Cohen, J. (1973). The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613-619.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Washington D.C.: Center for Educator Compensation Reform, U.S. Department of Education.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61,* 29-48. https://doi.org/10.1348/000711006X126600

Gwet, K. L. (2010). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (2nd ed.). Advanced Analytics LLC.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record, 116*(1), 1-28.

Hill, H. C., Charalambous, 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. https://doi.org/10.3102/0013189X12437203

Hill, H. C., & Herlihy, C. (2011). Prioritizing teaching quality in a new system of teacher evaluation. *Education Outlook*. http://www.aei.org/outlook/101089

James, J. & Wyckoff, J.H. (2020) Teacher evaluation and teacher turnover in equilibrium: Evidence from DC Public Schools. *AERA Open 6*(2), 1-21. https://doi.org/10.1177.2332858420932235

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance.* The Guilford Press.

Lambert, R., Holcomb, S., & Bottoms, B. (2021). Examining the inter-rater reliability of evaluators judging teacher performance: Proposing an alternative to Cohen's Kappa [Paper presentation]. National Council on Measurement in Education (NCME) Conference 2021, Virtual.

Lambert, R. G., Moore, C. M, Bottoms, B. L, Vestal, A., & Taylor, H. (2021). Use of Rasch modeling and focus group interviewing to inform the training of teacher evaluators. *International Journal of Multiple Research Approaches, 13*(2), 1-15.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174. https://doi.org/10.2307/2529310

MET Project. (2010). *Working with teachers to develop fair and reliable measures of effective teaching*. Bill and Melinda Gates Foundation.

National Council on Teacher Quality. (2019). Measures of Professional Practice: North Carolina results. *State Teacher Policy Database*. [Data set]. Retrieved from: https://www.nctq.org/yearbook/state/NC-Measures-of-Professional-Practice-95

Porter, J. M., & Jelinek, D. (2011). Evaluating inter-rater reliability of a national assessment model for teacher performance. *International Journal of Educational Policies*, *5*(2), 74-87.

Ross, E., & Walsh, K. (2019). *State of the states 2019: Teacher and principal evaluation policy*. National Council on Teacher Quality. https://www.nctq.org/pages/State-of-the-States-2019:-Teacher-and-Principal-Evaluation-Policy

Rui, N., & Feldman, J. M. (2012). Inter-rater reliability of a classroom observation protocol – a critical appraisal. *US-China Education Review, 3,* 305-315.

Sartain, L., Stoelinga, S. R., & Brown, E. (2009). *Evaluation of the excellent in teaching pilot: A report to the Joyce Foundation.* Consortium on Chicago School Research.

Steinberg, M.P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance & Policy 10*(4), 535-572. https://doi.org/10.1162/EDFP_a_00173

Uebersax, J. (2002). *Kappa coefficients: A critical appraisal*. https://www.john-uebersax.com/stat/kappa.htm

U.S. Department of Education, National Center for Education Statistics, Common Core of Data. (2021). *State Nonfiscal Public Elementary/Secondary Education Survey, 2018-19 v.1a.* https://nces.ed.gov/ccd/elsi/expressTables.aspx

Walsh, P., Thornton, J., Asato, J., Walker, N., McCory, G., Baal, Jo., Baal, Je., Mendoza, N., & Banimahd, F. (2014). Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the emergency department. *PeerJ, 2,* 1-19. https://doi.org/10.7717/peerj.651

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New Teacher Project.

White, M. C. (2018). Rater performance standards for classroom observation instruments. *Educational Researcher, 47*(8), 492-501. https://doi.org/10.3102/0013189X18785623

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Taylor & Francis Group.

Wind, S. A., & Engelhard, G. (2012). Examining rating quality in writing assessment: Rater agreement, error, and accuracy. *Journal of Applied Measurement, 13*(4), 1-15.

Xie, Q. (2013). *Agree or disagree? A demonstration of an alternative statistic to Cohen's Kappa for measuring the extent and reliability of agreement between raters*. In proceedings of the Federal Committee on Statistical Methodology Research Conference, The Council of Professional Associations on Federal Statistics, Washington, D.C.

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), 11-26. https://doi.org/10.31405/jes.2.2.2

Author Biographies

T. Scott Holcomb is a Ph.D. candidate in the Educational Research, Measurement, and Evaluation program at the University of North Carolina at Charlotte. Currently, Scott is a Graduate Research Assistant in the *Center for Educational Measurement and Evaluation*. His research interests include educational leadership, rater-mediated assessment, inter-rater reliability, educator effectiveness, and teacher evaluation.

Richard Lambert is a Professor in the Department of Educational Leadership in the Cato College of Education at the University of North Carolina at Charlotte and the Director of the *Center for Educational Measurement and Evaluation*. His research interests include formative assessment for young children, applied statistics, and teacher stress and coping.

Bryndle L. Bottoms is a Ph.D. candidate in the Educational Research, Measurement, and Evaluation program at the University of North Carolina at Charlotte. She works as a Graduate Assistant for the *Early Educator Support Office* and the *Center for Educational Measurement and Evaluation*. Also, she teaches Elementary Math Methods at Winthrop University in Rock Hill, SC. Her research interests include Rasch measurement, interrater reliability, teacher evaluation, and math education.